ROHAN CHARUDATT SALVI, Department of Computer Science, University of Illinois Chicago, USA

NIGEL BOSCH, School of Information Sciences and Department of Educational Psychology, University of Illinois Urbana-Champaign, USA

Artificial Intelligence has expanded its influence far beyond traditional boundaries in our society. One prominent application of artificial intelligence is the use of large language models, which have transcended their initial roles in high-tech industries and academic research and are now actively utilized by individual users. These models have continually improved over the years in their generative capabilities and performance across numerous tasks. However, they still pose a persistent risk of reproducing biases and stereotypes. Previous research has predominantly focused on quantitatively measuring biases in these large language models. In this study, we seek to assess not just the presence of bias itself, but the perception of stereotypes by these models via in-depth exploration of their responses. We demonstrate how the computational grounded theory framework, which integrates qualitative and quantitative approaches, can be applied in this context to assess the conceptualization of stereotypes. Furthermore, we contrast language model results with a survey of 400 human participants who also completed similar prompts as the model in order to understand people's perception of gender stereotypes. The results indicate substantial similarities between language model and human perceptions of stereotypes, highlighting that a model's perception stems from societal perception of stereotypes.

CCS Concepts: • Computing methodologies \rightarrow Natural language processing; • Human-centered computing \rightarrow Collaborative and social computing.

Additional Key Words and Phrases: Large language models, mixed methods, grounded theory, computational social science, LLM bias

ACM Reference Format:

Rohan Charudatt Salvi and Nigel Bosch. 2025. Investigating Perception of Gender Stereotypes in Large Language Models: A Computational Grounded Theory Approach. ACM J. Responsib. Comput. X, X, Article XX (May 2025), 28 pages. https://doi.org/10.1145/3737882

1 Introduction

Large language models (LLMs) are artificial intelligence models designed to process and generate a wide range of text content, such as human-like language [19] and computer code [79]. They use deep learning models, such as transformer architectures [83], to process and generate text. These models are trained on massive datasets, containing billions of sentences, to learn grammar, context, and semantic relationships. The size of these models is characterized by the number of trainable parameters they possess, often ranging from hundreds of millions to tens of billions [7, 25, 78, 83]. LLMs have various applications, including machine translation [41], text summarization [14], and chatbots [45].

Authors' Contact Information: Rohan Charudatt Salvi, rcsalvi2@uic.edu, Department of Computer Science, University of Illinois Chicago, Chicago, Illinois, USA; Nigel Bosch, pnb@illinois.edu, School of Information Sciences and Department of Educational Psychology, University of Illinois Urbana-Champaign, 43 Champaign, Illinois, USA.

50 Manuscript submitted to ACM

51

1 2

3

10

11

12

13 14

15

16

17

18

19

20

21 22

23

24 25

26 27

28

29

30 31 32

33

34

35 36

37

38

39

40 41

42

44

⁴⁵ Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not 46 made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components 47 of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on 48 servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

⁴⁹ © 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

53 While useful, however, large language models can develop biases across different categories due to their training data 54 reflecting societal prejudices [6, 10, 13, 17]. These biases encompass socio-cultural factors like race, gender, religion, 55 and nationality, potentially perpetuating stereotypes [13, 15, 17, 48, 58]. Gender bias has been shown quite evident, for 56 example, as these models tend to reinforce traditional role associations with gender [59, 62]. Additionally, these models 57 58 may produce ethically questionable outputs, including promoting violence or hate speech [35]. The utilization of large 59 language models imbued with biases can yield various societal consequences, warranting careful consideration [22]. 60 The consequences include, for example, biased outputs perpetuating false narratives and deepening the dissemination 61 of misleading content [38]. The unintended generation of discriminatory outcomes in domains such as hiring [22], 62 63 medicine [64], content moderation [81], and customer service, may reinforce systemic prejudices and yield unfair 64 treatment. Such instances will eventually erode public faith and trust in AI and machine learning systems, potentially 65 hampering their adoption in various spheres. Efforts to tackle these biases involve employing diverse, carefully curated 66 data, incorporating fairness measures during training, and applying bias mitigation techniques [52, 62]. However, fully 67 68 eradicating biases remains challenging, representing an ongoing focus of research in natural language processing.

69 In this study, we investigate how LLMs perceive stereotypes, focusing specifically on gender-based stereotypes as a 70 common example. We examine it through two key research questions. First, how do LLMs perceive gender stereotypes? 71 Second, do these perceptions align with human perceptions? We utilized open-ended text generation, followed by text 72 73 analysis as an approach to analyze how models perceive these gender-based stereotypes. This approach is novel with 74 respect to prior research using related methodologies (e.g., [17, 26, 55, 73]) because (1) we focus specifically on the 75 perceptions of stereotypes as opposed to their propagation, yielding different insights, and because (2) we demonstrate a 76 means to use computational grounded theory [63], a methodology leveraging both large-scale quantitative analysis 77 78 and fine-grained qualitative analysis. In this study, we highlight the significance of this methodology, as it effectively 79 balances human interpretation through qualitative assessments with the support of computational capabilities. We 80 believe this balance is essential given the diverse nature of stereotypes across different societies and perspectives, which 81 may be more suitably identified by a somewhat open-ended approach. Finally, we conducted a survey to gauge people's 82 83 perceptions of gender stereotypes and compared them to the model's perceptions as a means of validating findings and 84 generalizing them from one LLM (where it is possible they are idiosyncratic) to a broader context (where many LLMs 85 share the perceptions of people who wrote the text on which they were trained). As the use of LLMs across various 86 domains and applications grow researchers will need to develop approaches beyond quantitative analysis that also 87 involves having humans qualitatively assessing the outputs. Thus in future such CGT-based approach can be employed 88 89 as a comprehensive mixed approach for LLM generated text. 90

91 92

2 Quantitative Approaches to Assessing LLM Bias

Language models have always suffered from biases, like essentially any other machine learning model [56]. [13] showed, 93 94 for example, that word embeddings-a representation technique used to represent text as numeric vectors that encode 95 semantic relationships between words-can capture gender biases if they are developed using a corpus containing 96 gender biases. With the introduction of transformer-based architectures [83], the language generation capabilities of 97 98 these models have improved significantly [19]. Transformers are advanced neural network architectures that utilize 99 self-attention mechanisms to process sequences of data, such as text, in parallel. This allows them to exploit the context 100 and relationships within the data, leading to more accurate and coherent text generation. However, they require huge 101 datasets for training [74]. The training data is typically collected from text on the web, which is prone to containing 102 103 stereotypical text [23]; thus, during training, a model can encode stereotypes-as well as, perhaps, perceptions about 104 Manuscript submitted to ACM

stereotypes themselves, as studied in this paper. Researchers have studied what different types of biases exist in these models. [1] highlighted bias against Muslims in LLMs especially GPT-3. [73] highlighted gender stereotypes in LLMs. [57] highlighted racial bias in models, bias due to language dialect [12], disability [43] and sexual orientation [29].

Various approaches have been suggested to identify and measure bias in language models [31]. As stated by [59] "In order to assess the adverse effects of these models, it is important to quantify the bias captured in them." The methods to quantify biases broadly fall into two categories [55].

2.1 Predefined Stereotype-associated Tasks using Datasets

115 In this approach, researchers curate several sentences or prompts associated with the bias they want to measure. The 116 datasets have a task associated with sentences (e.g. selecting pronouns for text prompts around a given occupation), 117 and the model is judged based on its performance on the task. "Performance" in this case is usually a measure of 118 119 probability for the model's inclination toward stereotypical and a less stereotypical response. [59] developed Stereoset, 120 a large-scale natural dataset in English to measure stereotypical biases of four types: gender, profession, race, and 121 religion. WinoBias and WinoGender [87], [69] are perhaps the most widely used datasets developed to capture gender 122 bias in LLMs. [60] proposed another dataset, CrownS-Pairs, to capture bias across nine dimensions, such as race, religion, 123 124 and age. CrownS-Pairs data focus on stereotypes about historically disadvantaged groups and contrasts them with 125 advantaged groups. Recently, new datasets have been released following a similar methodology but extending to 126 different languages such as Japanese and Russian [47] and measuring biases against other minority groups, such as 127 LGBTQ+ individuals through the WinoQueer dataset [29]. 128

2.2 Language Generation and Analysis

Another popular approach to measure bias is allowing the model to freely generate text given some prompts and then 132 analyze the generated text [73]. Similar to the previous approach, numerous prompts are designed by the researchers 133 134 related to the demographic group for which they wish to assess the bias. Following generation, researchers subsequently 135 perform various automatic analyzes to investigate the properties of generated samples. Properties of interest could be 136 sentiment or toxicity, for instance; analyzing sentiment reveals if certain groups are consistently discussed in more 137 negative or positive tones [73], while assessing toxicity can show if some groups are more frequently associated with 138 139 harmful or offensive content [1]. By examining these factors across various groups, it becomes possible to identify 140 and quantify biases. Both of these properties can be analyzed using off-the-shelf rule-based tools, individually trained 141 transformer-based classification models, or publicly available inference APIs [26, 35, 73]. Beyond this, [48] uses named 142 entity recognition pipelines to detect the mention of specific occupations. 143

144 Recent studies have provided highly relevant criticism for both approaches. [11] highlighted that benchmark datasets 145 lack a clear definition of what is being measured and suffer from several conceptual and operational pitfalls such as 146 inconsistencies in the anti-stereotype being a negation, contrastive factual, or an irrelevant statement, and names 147 being employed in place of the group name in sentence pairs respectively. [2] highlight how the measurement of bias 148 149 through text completion using prompts is prone to yielding contradictory results under different experiment settings. 150 [65] investigate how gender bias is measured using various extrinsic metrics and also demonstrate how a dataset such 151 as Winobias [87] can be coupled with different metrics. Finally, they examine how the selection of the dataset and its 152 makeup, along with the choice of the metric, impact the measurement of bias, revealing notable differences within 153 154 each aspect. These findings have spurred increasing interest in discovering innovative and more effective methods to 155 assess bias without encountering the limitations currently observed; for example, [55] proposed to quantify stereotype 156 Manuscript submitted to ACM

105

106

111

112 113

114

129 130

biases by measuring the probability of demography based on the stereotype, such as measuring the probability of pronouns given the profession of the person is nurse, as opposed to what the Steroset dataset does [59], and measure the probability associated with only one word representing a demographic such as "he", "she" and "they", with respect to above example to reduce the effect of the noise. [24] propose two new measures for gender biases to facilitate a more refined understanding of gender bias, by identifying both the unequal preference for male or female pronouns and the reinforcement of gender stereotypes in different professions.

In sum, quantitative measures have been well studied and do provide some insights into biases present in LLMs. However, there are limitations to purely qualitative approaches, some of which may be addressed by qualitative methods, as discussed next.

167 168 169

171

165

166

¹⁷⁰ 3 Qualitative Approaches for In-depth Examination of Text

Qualitative methodologies provide many approaches to gain insights from data via human interpretation of lowlevel data such as video, text, and observation [75]. Qualitative methods may thus address some of the limitations of quantitative measures stemming from the higher-level, aggregate nature of most quantitative analyzes. Quantitative approaches for studying LLMs in particular often focus on predefined constructs such as sentiment [73] or toxicity [35]. Conversely, certain qualitative approaches, such as open coding [72], content analysis [49], and grounded theory [37], are designed to discover what constructs are apparent in a bottom-up fashion informed by both data and researcher knowledge.

180 Grounded theory, in particular, is one of the most widely used qualitative methodologies [37]. Grounded theory 181 focuses on developing theories through a systematic and iterative process of data collection, coding, and analysis. 182 Researchers using grounded theory engage in constant comparative analysis, where they continually contrast new data 183 184 with previously collected data and emerging concepts, allowing theories to evolve and emerge naturally. This approach 185 is particularly useful for exploring complex and dynamic social phenomena [20], since it emphasizes understanding the 186 perspectives and experiences of individuals within their social context. Grounded theory has been widely adopted across 187 various disciplines, providing a robust framework for generating novel insights and contributing to the development 188 189 of new theoretical perspectives. However, [63] highlighted challenges with grounded theory such as that it involves 190 researchers making subjective decisions as they code and analyze data, hence infusing their own personal inclinations 191 and traits into the analysis [70]. These biases infused in turn make the process of validating and replicating the studies 192 challenging [8]. Moreover, grounded theory faces limitations in dealing with large-scale data sets, particularly in the 193 194 context of unstructured social data [4].

195 Content analysis is one common approach to employ grounded theory with textual data [49]. Content analysis is a 196 pivotal method within qualitative research, frequently employed with versatile data (e.g., text, visual, audio) in fields 197 such as media and political sociology [28, 30]. There are three main approaches to content analysis: empirical, emergent 198 199 coding/grounded theory, and theoretical [77]. The grounded theory approach allows "an analysis without a particular 200 theory in the first place, but then use the data under investigation to develop a theory. This theory is then applied to the 201 subsequent data" [77]. The process involves meticulously reading through the data, systematically coding it to assign 202 labels or categories to specific elements [49, 70], and subsequently delving into the analysis of the coded text to unveil 203 204 patterns, connections, and disparities that reside within texts [28, 30]. By deciphering these patterns, researchers can 205 uncover trends, and recurring themes to provide insights such as understanding the impact of educational policies on 206 the mission statements of schools [5]. 207

209 The drawback, however, is the time-intensive nature of such analyzes, particularly in problems such as the analysis 210 of LLMs where the amount of data that could be analyzed is essentially unlimited. An alternative approach to purely 211 qualitative research is to apply qualitative methods in only a subset of cases identified via quantitative methods. These 212 213 techniques range from basic calculations of word or phrase frequency to more complex strategies such as supervised and 214 unsupervised machine learning algorithms [44, 71]. Additionally, the quantitative techniques may incorporate natural 215 language processing mechanisms that consider the structure of language and relationships between words during 216 computations [27]. Similar to any scientific approach, the selection of text analysis method should align with the research 217 218 question and the available dataset [42]. Fortunately, the wide range of technical options allows for careful application 219 across various research questions, effectively utilizing different data types to offer insightful solutions to a variety of 220 inquiries [27, 44, 71]. For instance, [84] uses a qualitative approach to examine gender biases in LLM-generated texts, 221 focusing on stylistic and lexical differences informed by social science findings on gender communication. Similarly, 222 223 [61] looks at nationality biases through sensitivity analysis to see how factors like economic status and internet usage 224 affect sentiment in generated content. While much of the existing work in LLM research has focused on identifying 225 hallucinations or employing LLM primarily for qualitative analyses, the qualitative analysis of LLM-generated text for 226 bias and stereotypes is still an emerging field that will likely expand as LLM usage grows across different domains. 227 228 These studies underscore the necessity of a mixed-methods strategy, blending qualitative insights with quantitative 229 accuracy to thoroughly explore the complex dynamics of language model outputs. 230

In this paper, we focus in particular on computational grounded theory (CGT), discussed next.

3.1 Computational Grounded Theory

231 232

233 234

235

236

237 238

239

240

To address the inherent challenges of subjective judgment and limited scalability associated with grounded theory, [63] introduced an innovative three-step methodology known as CGT. This approach effectively combines human expertise with computational techniques, especially machine learning, to enhance the content analysis process. The CGT framework strikes a balance between interpretive and computational elements, thereby mitigating the constraints associated with each, facilitating both meaningful interpretation and large-scale analysis.

The three-step CGT methodology entails the initial step of pattern detection and refinement, where computational 241 techniques are employed to extract and explore text patterns through the list of frequent words or topics, word networks, 242 243 or other quantitative measures. Subsequently, CGT involves a stage of deep reading and interpretation, allowing for 244 the identification of meaningful patterns. In the final step, computational methods are utilized to validate the patterns 245 that have been identified across a large dataset. Researchers can delineate the data analysis procedures, computational 246 tools, and models employed at each step and any criteria for data manipulation. Through this, researchers not only add 247 248 transparency to the research but enable others to replicate studies with identical datasets and computational strategies. 249 Moreover, CGT leverages automation to make it scalable and thus well-suited for exploring "big data" applications. 250 Hence, this systematic approach "brings inductive content analysis closer to the validity, reliability, reproducibility, and 251 scalability necessary for scientific research"[63]. 252

Like traditional grounded theory, CGT focuses on the importance of staying grounded in the data, where analysis can be directly linked to the textual evidence. Both methodologies advocate for an iterative process where the findings are continually refined and hypotheses are developed based on new data they uncover [37, 63]. This iterative process ensures that both methodologies adapt and evolve as new insights are gained. Additionally, both methods prioritize a deep understanding of the data, with CGT employing computational tools to process data and complement grounded theory's thorough, deep approach. However, traditional grounded theory does not scale well to large datasets, which Manuscript submitted to ACM limits its applicability in contexts where researchers have access to vast amounts of unstructured social data. Therefore,
 an advantage of CGT is its ability to facilitate an in-depth examination of content, effectively considering both the
 quality and quantity of data. Moreover, unlike grounded theory, CGT does not use fixed coding schemes. Instead, it
 uses flexible, high-level coding that starts with computational methods for initial categorization and is later refined
 manually to reduce subjectivity and improve reliability.

267 CGT also benefits from a rigorous yet interpretive method that allows for both close and distant reading techniques, 268 enabling researchers to measure meaning more effectively [63]. However, the approach is not without its limitations. 269 One significant challenge, as discussed in [18] is that algorithms like the latent Dirichlet allocation topic model often 270 271 struggle with unbalanced classes, leading to the generation of duplicate and conglomerate clusters. This indicates 272 that the ability to locate planted topics or ensure representative document selection is compromised, questioning the 273 reliability of the output. Furthermore, CGT's reliance on computational models raises concerns about the interpretation 274 of how meaning corresponds to word patterns, suggesting that existing validation strategies such as face validity 275 276 and indirect validity may not adequately protect against substantial measurement errors [3]. Therefore, while CGT 277 leverages the strengths of both qualitative insights and quantitative rigour, it requires a careful and skeptical application 278 to overcome these inherent challenges. 279

Recent studies such as [82] have employed CGT to analyze student physics problem-solving approaches and gain 280 insight into physics education research. Furthermore, [39] employed key principles from CGT to investigate and 281 282 compare conspiracy theories generated by humans and bots on social networks. A few studies, such as those by [34] 283 and [51], have employed only the first step of CGT to look for patterns through methods such as topic modeling. Others 284 like [54] draw parallels from CGT to employ mixed methods. [36] uses it to analyze interpretive topic modeling for 285 286 content analysis. These instances show that CGT can be used to explore and understand patterns across diverse areas, 287 thus helping researchers improve their studies. 288

The three steps of CGT can be summarized as follows:

290 3.1.1 Pattern Detection through Human-Centered Computational Exploratory Analysis. The initial computational step 291 in pattern detection serves a dual purpose. First, by simplifying text through computational methods, we can uncover 292 patterns that may not be immediately apparent to human readers. Patterns emerging from the exploration analysis 293 294 can prompt researchers to adopt fresh and possibly unexpected perspectives on their data, leading them to discover 295 novel avenues for analysis. Essentially, computers bring to light what may escape human perception. Second, the 296 computational methods perform rapid analysis of the text while ensuring full reproducibility since another researcher 297 can get the same results by following the exact computational steps on the same dataset. Thus, computational exploration 298 299 methods excel at simplifying and revealing patterns within data at scale, with minimal extra effort required from the 300 researcher. 301

302 3.1.2 Hypothesis Refinement through Human-Centered Interpretation. In traditional grounded theory, researchers 303 alternate between examining their data and interpreting the analysis of the data. The second step of CGT, influenced by 304 grounded theory, aims to replicate this process but incorporates computational elements. By initially identifying patterns 305 306 through words and forming groups of those words in step 1, researchers can pinpoint texts that represent specific 307 themes or categories. These selected texts can also be used to calculate the relative prevalence of each category. This 308 computer-guided reading approach allows researchers to validate their interpretations of word or sentence groupings 309 generated in the first step. It also helps researchers better understand how the most frequent words are expressed in 310 311 complete sentences or topics in the documents respectively. Hence, the researcher does not need to go through the 312 Manuscript submitted to ACM

7

entire text. In the example provided in [63], based on 12 different topics identified in step 1, Nelson read 10 documents for each topic. Additionally, the second step of CGT can either confirm or amend the patterns identified in the initial step. Selecting representative text based on the patterns identified in the previous step, the second step not only enhances efficiency since the reader does not need to go through the entire document, but also ensures that important passages are not overlooked due to fatigue. Moreover, by involving human reading, the word count and the distribution of sentences into three sentiments (in the [63] example) gain meaningful context and are interpreted in a theory-informed manner.

The iterative initial two steps of CGT refine data analysis. The interpretive reading step 2 translates computational results into sociologically significant concepts, empowering researchers to generate hypotheses about the societal context that gave rise to the data. Remaining aligned with the grounded theory framework of examining and interpreting data, these two CGT steps might also guide researchers toward collecting supplementary or distinct data to arrive at accurate conclusions about their subject of interest. Once data-driven patterns are identified and refined through the first two steps, computational techniques support researchers in the crucial final step of pattern confirmation.

3.1.3 Pattern Confirmation. This third step of CGT tests the generalizability of patterns discerned in the initial two steps, serving as a conclusive and essential validation of the previous two inductive steps. Additionally, step three compels researchers to operationalize patterns identified through the first two steps in quantifiable terms, formalizing concepts and patterns in a manner not always undertaken in qualitative analysis settings.

4 Method

313

330

331

332

333 334

335 336

337 338

339

340

341

342 343

344

345

346 347

348

To explore and understand the perception of stereotypes by the LLM versus human perceptions, we conducted two studies. The first study employed generation by an LLM, followed by analysis of the text generated through CGT to identify and confirm stereotype perceptions in the LLM. Based on the results from the first study, we conducted a survey and examined the information collected to understand human's perception of gender stereotypes. In our study, it is essential to acknowledge that, for simplicity's sake, we have primarily considered two genders, namely men and women, though it will be crucial to continue this work (and other work on LLM biases) to understand stereotypes related to more gender identities.

4.1 Study 1

349 We employed a GPT-Neo model for gender-based stereotypes in large language models in this study. "GPT-Neo 2.7B" is 350 a transformer model designed by EleutherAI that builds upon the transformative capabilities of its predecessor, GPT-3, 351 to facilitate even more sophisticated natural language understanding and generation. It was trained on the Pile [32], 352 353 a large-scale curated dataset created by EleutherAI for the purpose of training this model. The model was trained for 354 420 billion tokens over 400,000 steps and has 2.7 billion parameters which is a billion more than GPT-3. To ensure 355 better reproducibility, we chose GPT-Neo because it is open-source and provides clear documentation of the training 356 357 data it uses. These factors are crucial for understanding and mitigating biases inherent in the training corpus and for 358 comprehending model outputs. 359

GPT-Neo can leverage its extensive training on a vast array of internet text, to engage in coherent conversations [50], answer questions, draft creative pieces [85], offer code snippets [86], and the previous research has demonstrated that the model exhibits social biases concerning gender, religion, and disability [21, 40, 53]. In this study, we employed prompt-based language generation and subsequent analysis to identify gender biases. However, we want to emphasize Manuscript submitted to ACM



Fig. 1. Method overview illustrating how we implement computational grounded theory for analyzing stereotypes in language models and align that process with humans' perceptions of stereotypes.

that While our approach focuses on GPT-Neo and gender stereotypes, our methodology is not limited to this model or stereotype. It can also be applied to other LLMs and different stereotypes.

We have adopted the computational grounded theory approach and have employed three similar steps.

4.1.1 Step 1: Pattern Discovery. For step 1, we required generated text from the LLM. Thus, we employed a prompt to generate textual data from GPT-Neo [32]. In prior works, researchers stated the subject in the prompt and let the model generate the rest. For instance, [61] used the prompt "The <Nationality>people" to generate short sentences with GPT-2 and understand nationality bias in it. Since we were examining the perception of gender stereotypes and wanted only one stereotype per sentence, we developed a prompt "One of the stereotypes for [GENDER]" to allow the model to generate varied gender-based stereotypes and highlight that there could be many stereotypes associated with a gender.

We began by generating 1800 sentences using the prompt with "women". Initially, we attempted to identify common themes in these sentences using topic modeling, specifically, latent Dirichlet allocation [9], but we did not observe clear themes. Hence, we adopted a simpler approach by examining the most frequently occurring words. The analysis of the top words and manual inspection of sentences associated with them led to the identification of four distinct sentence structures, which facilitated categorization into clusters. The first sentence structure, which emerged after prompt generation, was characterized by the use of the word "is" and often entailed stating a stereotype about women. The second structure typically had the word "in" after the prompt, indicating a stereotype about women in a specific context (typically an occupation or region). The third involved comparisons between female and male subjects. Finally, the fourth sentence structure encompasses all the remaining sentences. We recorded the top 30 most frequent words from each cluster to further analyze the prevalent sentence patterns. From the top 30 words, we could observe that the "is" cluster's sentences focused on nature and characteristics, while the "in" cluster highlighted occupations and regions Manuscript submitted to ACM

417

418

419

420 421

422

423

424

425 426

427

436 437

438

439

440

441 442

443

444

445

446 447

448

449

450

451

452 453

454 455

456

457

461

462

466 467

468

where gender stereotypes are prevalent. We also observed the top 30 words varied substantially for the "is" cluster whereas we found common terms for "in" cluster, especially the area of work or region such as "military", "tech", and, "america".

To gain further insights and account for variations in frequent words such as antonyms (e.g., "weak" and "strong") and negation in sentences (e.g., "not strong"), we incorporated sentiment analysis for the sentences using a pre-trained model¹. This addition allowed us to better understand the biases and patterns present in the data, since, despite differing keywords such as "not strong" and "weak", the sentences exhibited similar sentiments. We followed a similar procedure for analyzing "men", with the only difference being the gender word change in the prompt.

428 4.1.2 Step 2: Human Interpretation. In this step of our work, our primary task was to study text, spot certain patterns, 429 and form initial ideas. We began by manually examining the sentences related to the most frequently used words in 430 two different categories, namely "is" and "in". Afterward, we delved deeper into analyzing the sentiments of these 431 sentences to discern the stereotypes they conveyed. For instance, "strong" and "weak" were both observed as top words 432 433 for "women". The sentences containing these words were predominantly negative. Thus, by closely observing these 434 sentences, we hypothesize that one of the stereotype patterns perceived by the LLM is that women are weak. 435

For the most frequently observed words associated with a specific gender such as "women", we conducted a search and analysis of sentences containing those words in sentences associated with "men". For instance, "cook" was often associated with women, but we also examined sentences where the word "cook" was used for men. This helped us understand how the model generated text for each gender.

Throughout this step, we actively searched for common gender-related stereotypes. These stereotypes could pertain to women on a global scale or within specific fields of work or even in particular countries. For instance, we discovered stereotypes for women working in fields like business or science or pertaining to countries like India or the United States.

Our work during this step helped us identify the broader themes among the biases we encountered, including the global stereotypes commonly applied to all women or men, stereotypes in specific fields such as business, and those associated with a certain country such as the United States. We would like to highlight that if the model-generated text does not specify a region, we interpret the stereotypes as not specific to any geographic region. It is imperative to underscore that the outcomes observed in this step of our research are preliminary and not yet conclusive. There exists a requirement for thorough validation through statistical methods and taking into consideration the effect of prompt templates in the subsequent steps of our research to ensure the reliability of our findings.

4.1.3 Step 3: Pattern Confirmation. In the third step, we adopted a strategy akin to that of the initial step. However, this time, our goal was not to identify patterns but to confirm them. To ensure that the perceptions of stereotypes we observed in the second step would commonly emerge in the model and not solely with one specific prompt, we 458 459 created sentences using two additional prompts: "A common perception about [GENDER]" and "One of the common 460 perceptions about [GENDER]". For each of these prompts, we generated a set of 900 sentences, resulting in a total of 2,700 sentences for each gender category. To maintain consistency in our analytical process, we organized these sentences into the four distinct groups described in step 1, namely "is", "in", "both genders", and "rest", extracted key 463 464 terms, and then conducted sentiment analysis. 465

¹Sentiment analysis model: https://huggingface.co/j-hartmann/sentiment-roberta-large-english-3-classes

Manuscript submitted to ACM

[2] highlighted the importance of prompts in influencing the output of language models. Therefore, we performed a 469 470 correlational analysis to examine whether the choice of these two new prompts in step 3 had an impact on both the most 471 frequently occurring words and their associated sentiments. We first sorted the top 30 words alphabetically and, for 472 each word, recorded the total word count as well as how frequently it appeared in sentences with different sentiments: 473 474 positive, negative, and neutral. We then calculated the correlation between the top 30 words from the new prompts and 475 the top 30 words from step 1, based on their distribution across different sentiments. Following this, we repeated the 476 process of calculating the correlation matrix, once again based on the distribution across different sentiments. However, 477 this time around we sorted the top words by word count rather than alphabetically to examine the data through a 478 479 different lens, emphasizing influential words in terms of frequency, analyzing their sentiment associations, and offering 480 a comparative perspective to the initial alphabetical sorting. 481

The third step while maintaining the same methodology as step 1 verifies the presence of stereotypical patterns that were identified during steps 1 and 2. To ensure patterns were not influenced by prompts we employed two new prompts in this step and also conducted a correlation analysis to comprehend how these prompts affected the top words and sentiment.

4.2 Study 2

In the second study, we conducted a survey to gather insights on gender stereotypes. This survey was designed based 490 on the insights obtained by our first study. Our survey involved 221 male, 170 female, 7 trans or non-binary participants, 491 492 and 2 participants who did not disclose their gender. All of them were from the United States, and we also recorded 493 both their genders and age. To conduct this survey, we created a set of 6 prompts that closely mirrored the sentences 494 generated by GPT-Neo. We selected "tech industry," "business," "sports," "science," and "military" as the fields for the 495 "in" prompts because these terms were the most common in the "in" cluster for women during step 1 of our Study 496 497 1 (Table A3). Participants were tasked with constructing complete sentences using these prompts. The survey was 498 divided into two sections: "is" and "in." The survey structure is outlined below.

Survey Structure:

٠	"Is"	Prompts
---	------	---------

- One of the common stereotypes for [Gender] is
- "In" Prompts
 - One of the common stereotypes for [Gender] in the tech industry is
 - One of the common stereotypes for [Gender] in business is
 - One of the common stereotypes for [Gender] in sports is
 - One of the common stereotypes for [Gender] in science is
 - One of the common stereotypes for [Gender] in the military is

In the "is" section, participants were asked to create five sentences each for both women and men. Participants used a single prompt for this purpose, following the specified structure. In the "in" section, we used five distinct prompts, each focusing on various domains such as science, technology, business, sports, and the military. These domains were chosen based on the observations made in the second step of study 1, particularly due to their relevance to the identified stereotypes. These domains also appeared among the top 30 words. Participants were required to complete sentences Manuscript submitted to ACM

10

499 500 501

> 502 503

482

483 484

485

486 487 488

489

> 508 509

510

511

addressing gender-specific stereotypes within these areas. In total, each participant completed 20 sentences, resulting in a total of 8,000 sentences gathered for analysis.

We followed a similar analysis process as in study 1. Sentences were categorized into "is", "in", "men and female comparison", and "rest". We examined the top 30 words within each category and evaluated the sentiment of sentences containing these words.

Finally, to ensure that the stereotypical patterns we identified in step 3 of our study 1 were not merely LLMconfabulated stereotypes, but could also reflect societal perceptions of stereotypes, we conducted a Kolmogorov–Smirnov (K-S) test [66]. This non-parametric statistical test determines whether a sample adheres to a specific probability distribution. It is commonly used to compare datasets and ascertain if they share the same underlying distribution. In our study, we employed the K-S test to compare the distributions of top words and assess the similarities or differences between sentences in the survey and the sentences generated by the LLM.

Our first step involved finding the common top words between the survey and the two newly introduced prompts, i.e., "A common perception about [GENDER]" and "One of the common perceptions about [GENDER]". For each word we found, we calculated its probability by dividing the word count by the total number of words in the respective sentence collections. We then used these probabilities associated with common words in the K-S test.

5 Results

521

522

523

524 525

526

527

528

529 530

531

532

533

534 535

536

537

538

539 540 541

542

543 544

545

546

547

568

5.1 Study 1

5.1.1 Step 1. We identified our sentences broadly fell into four categories based on the sentence structure. First were sentences of the form "One of the common stereotypes for women is", the second was "One of the common stereotypes for women in", the third involved comparing them with "men", and finally, the fourth was everything else.

548 By analyzing the top words, we gained insights into prevalent stereotype perceptions. For women, as highlighted in 549 Table 1, perceived stereotypes included terms such as passive, weak, emotional, mother, and housewife, which align 550 with previous research on stereotypes (as opposed to perceptions of them) in natural language processing methods 551 [16, 33]. Nevertheless, these top words alone failed to fully capture the nuances of meaning. Therefore, we employed 552 553 sentiment analysis, revealing that 56% and 50% of sentences exhibited a negative sentiment in the "is" and "in" categories, 554 respectively. Within the "in" category, we identified five prominent fields (or types of fields) in the top words: technology, 555 science, business, military, and sports. Some of these stereotypes were region-specific, applying mainly to women in 556 the United States or India. This suggests that the data used for training may have been biased towards these countries. 557 558 Furthermore, although the majority of sentences remained negative, this category contained slightly more positive 559 sentiments compared to the previous one, with words like weak, lazy, and good characterizing it. 560

Similarly, we examined the top words associated with men, revealing perceived stereotypes that portrayed men as aggressive, lazy, hardworking, and highly interested in sex and money (Table 1). We observed regional variations for men, primarily in the United States and India, and also identified fields such as technology and the military in the "in" category. The top words in the "in" category emphasized characteristics like laziness, sexual activity, and hard work. In terms of sentence sentiment, the "is" category predominantly featured negative sentiments, whereas the "in" category displayed a majority of positive sentiments, in contrast to the women's category.

5.1.2 Step 2. Upon reviewing sentences containing the most frequently used words, we made four key observations.
 We examined sentences that contained top words associated with the gender mentioned in the prompt, as well as how
 these words were used in sentences relating to other gender categories included in our study.

Method	Gender	Word	Total	Neg	Pos	Neu
		weak	24	23	0	1
Stop 1	Women	passive	18	5	1	2
Step 1	women	cook	21	12	5	4
		mother	18	5	11	2
		strong	19	2	12	5
	Mon	aggressive	25	14	2	9
	Men	lazy	47	47	0	0
		sex	25	14	9	2
		weak	20	20	0	0
Star 2	Women	passive	14	13	0	1
Step 5		emotional	7	5	1	1
		sex	13	11	2	0
		aggressive	31	19	4	8
	Mon	strong	11	2	7	2
	Men	sex	43	29	9	5
		lazy	10	10	0	0
		emotional	167	58	0	109
Chi da o	Wannam	weak	91	91	0	0
Study 2	women	mothers	23	1	3	19
		cook	33	0	0	33
		strong	137	1	0	136
	Mon	aggressive	64	3	0	61
	wien	sex	52	2	2	48
		emotion	46	23	0	23

Table 1. Selected top words from "is" category

• We observed that stereotypical and non-stereotypical sentences were generated across genders. Example: "cook" was a word associated with "women" but we found it in sentences for "men"

- One of the stereotypes for women is that they love to cook.
- One of the stereotypes for men is that they love food, they love to cook, do it all the time.
- One of the stereotypes for men is that they do not like to cook.
- One of the stereotypes for men is to go into the kitchen to cook an entire meal.
- We found a mix of sentences that portrayed a stereotype associated with gender in varying sentiments, although the majority of them portrayed the stereotype in a negative light.
 - Example: "Mother" was a stereotype associated with women in different sentiments
 - One of the stereotypes for women is that they're just bad mothers.
 - One of the stereotypes for women (some would say the only ones) is the role of the housewife or mother.
 - One of the stereotypes for women is to be a good mother or wife first, and then have lots of time for herself.
- The emergence of sentences with genders "men" and "women" revealed the model's perception that men are biased against women.

Example:

- One of the stereotypes for men in the tech industry is that they don't get women into positions of power.
- We encountered sentences that lacked meaning and often coherence, consisting mainly of negative or toxic words.
- Example
 - One of the stereotypes for women is that of the a bitchy bitch (or the bitchy bitch you would kill for).

This step in our analysis provided us with a glimpse into the model's generation related to stereotypes and its perception of stereotypes in our society. We identified a range of representations, encompassing both stereotypical and non-stereotypical portrayals across genders, along with perceived biases in gender characterizations, and a handful of Manuscript submitted to ACM

⁶²⁵ instances featuring illogical or blatantly offensive language. These insights contributed to our understanding of the

626
 627
 628
 628
 629
 629
 629
 620
 620
 621
 622
 623
 624
 625
 625
 626
 627
 628
 628
 629
 629
 620
 620
 620
 620
 621
 622
 623
 624
 625
 625
 626
 627
 628
 628
 628
 629
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620
 620

Table 2. Correlation analysis on the top sorted by word count between the prompts used in step 1 and step 3

(a) Prompt from step 1 and prompt 1 from step 3

			Step 3	
		Negative	Neutral	Positive
	Negative	0.9525	0.7970	0.8523
Step 1	Neutral	0.9485	0.7871	0.8407
	Positive	0.9629	0.7957	0.8528
(b) P	rompt from s	step 1 and pro	ompt 2 from	step 3
			Step 5	
		Negative	Neutral	Positive
	Negative	0.9717	0.7949	0.8309
Step 1	Neutral	0.9543	0.7842	0.8158

0.9560

0.7936

0.8280

5.1.3 Step 3. We replicated the analytical procedure carried out during the initial step in the third step of our study. However, this time around, we introduced two novel prompts. We discovered that certain keywords resurfaced regardless of the specific prompts used. For instance, words like "weak", "care", "passive", and "emotional" were consistently generated in relation to women. In the "in" category, the term "weak" remained prominent across both prompts, and it appeared within the fields of business, technology, and the military. Notably, in terms of geographical distribution, our findings once again highlighted the prevalence of the United States and India in generated text.

Positive

On the other hand, for men, the top recurring words from step 1 included "sex", "hard work", "aggressive", and "strong". When focusing on the "in" category, words like "lazy" and "sex" reappeared, particularly within domains such as business, the military, and science.

We conducted a correlation analysis to compare the prompts used in step 3 with those in step 1. Our observations revealed a strong correlation between sentiments when sorted by word count. However, when sorted alphabetically, we noticed weak correlations, with only a few exceptions. Alphabetical sorting was adopted to provide an initial view of the words without the influence of other factors like frequency or sentiment, enabling us to establish a baseline before diving into more complex analyzes. Table 2 presents the findings of our correlational analysis between step 1 and step 3 for sentences generated with the prompt related to women and categorized as "is." Prompt 1 in step 3 was "A common perception about [GENDER]" and Prompt 2 was "One of the common perceptions about [GENDER]".

5.2 Study 2

629 630

631

644

645 646 647

648 649

650

651

652

653 654

655

656

657

658 659

660

661

662

663 664

665

666

667 668

669

When we examined human-generated sentences from the survey, we found the prominent perceived stereotypes about
 women were that women are emotional and not as strong as men. Women were often seen as mothers and homemakers,
 focused on caring and cooking. Some participants also highlighted that a common stereotype for women was that
 they are not considered as smart or capable as men in different fields such as technology, military, and business. On
 the other hand, people believed that the stereotypes for men were being strong, aggressive, and interested in sex and
 Manuscript submitted to ACM

Gender	Category	Study 1	Study 2	K-S Test Statistic	<i>p</i> -value	Data Points
		Study 2	Prompt 1	.500	.283	8
	Is	Study 2	Prompt 2	.500	.283	8
Men		Prompt 1	Prompt 2	.375	.660	8
		Study 2	Prompt 1	.500	.771	4
	In	Study 2	Prompt 2	.500	.771	4
		Prompt 1	Prompt 2	.500	.771	4
		Study 2	Prompt 1	.714	.053	7
	Is	Study 2	Prompt 2	.714	.053	7
Women		Prompt 1	Prompt 2	.420	.571	7
		Study 2	Prompt 1	.570	.212	7
	In	Study 2	Prompt 2	.420	.575	7
		Prompt 1	Prompt 2	.285	.960	7

Table 3. Kolmogorov-Smirnov test results for distribution differences for the top words ("Data Points") in LLM and survey studies.

money—similar to those perceived by the LLM. Participants believed a stereotype exists that men are "smart" and "good" at their jobs, even though they might be seen as "nerds". Interestingly, the sentiment of the sentences for both genders throughout the survey was predominantly either neutral or negative, and only a few responses had a positive sentiment.

The results of the K-S test, shown in Table 3, consistently indicate that we could not reject the null hypothesis for all the tests we conducted, whether based on gender or categories. This suggests that these studies may share the same underlying distribution. Therefore, it implies that the frequent stereotypical patterns observed in the model are accurate reflections of the stereotypes that are widely perceived by people as well.

6 Discussion

We identified keywords in two sentence categories namely "is" and "in" that remained consistent when analyzing patterns in sentences completed by the model and humans. Our research involved comparing these words at both the model level and within societal contexts. We also conducted qualitative tests to evaluate them. When we combined different observations, insights from various stages, and quantitative assessments, we found that the model's perception of gender-based stereotypes aligns with people's existing perceptions on numerous points. Additionally, the model produced statements that challenge stereotypes, such as men and cooking [68]. In fact, we observed mixed perceptions of stereotypes present in the model that may or may not exist in society, and vice versa. For example, in our survey, we observed that people believe a common stereotype for women is "women being bad drivers", with the word "driver" ranking in the top 10 words. However, in the case of LLM "driver" was never generated by any prompt.

We also discovered that sentiments related to a stereotype can vary and may lead to biases. We noticed a slight difference between stereotypes and biases. For instance, the model might depict women as either good or bad mothers. In our survey, participants usually saw women simply as mothers or caring mothers. This observation is crucial because it helps bridge the gap between people's understanding of stereotypes and models' understanding of stereotypes. A prominent gender stereotype often perceived about women is the role of a mother [67]. However, labeling women as "bad mothers" could indicate bias in the model. Approaches like sentiment analysis may result in highlighting sentences stating "a bad mother" and "physically weaker" as a negative sentiment but the context and comparison must also be considered while identifying stereotypes and biases.

728 Manuscript submitted to ACM

729 Finally, the biases in these models and their perceptions of stereotypes primarily arise from the training data [62]. 730 While we did find that frequent contexts, such as science, business, and sports, are associated with stereotypes for 731 women, as well as countries like the United States and India, we cannot determine whether the perceived stereotypes 732 733 generated for women, in general, were influenced by stereotypes prevalent in these countries and fields, and if so, to 734 what extent. Similarly, for women in different fields, it is unclear whether the model extensively generated stereotypes 735 based on these regions. Therefore, such an open-ended generation approach can also help us understand the potential 736 contexts from where the model gains its notion of the stereotypes in our society. 737

738 An important theoretical implication of the results is that gender stereotypes related to personal characteristics are 739 carried over into job roles, particularly for negative stereotypes about women. We observed this pattern clearly in both 740 of our studies. In the "is" cluster for the LLM and in the survey, we found men commonly associated with words like 741 "strong", "hardworking", "aggressive", and "tough". For women, conversely, it was "weak", "passive", and "emotional". 742 743 Such stereotypes may lead to a perception of women as not being competent enough for certain roles, especially in 744 male-dominated areas, which we observed clearly in the "in" cluster where the stereotypes emerged for fields such 745 as the tech industry, science, business, military, and sports. Our findings showed stereotypes describing women with 746 limiting phrases like "not good enough" or "not strong enough" especially in tech and the military. On the contrary, 747 748 men were described with positive terms such as "hard work", "strong", and "smart" in both the LLM-generated text 749 and the survey. This highlights societal biases that favor a gender for roles in certain occupations as one of the most 750 prominent forms of stereotype. Our observed theory aligns with previous research by [80] on gender stereotypes, that 751 individual factors, sociocultural factors, and stereotyped thinking in human beings cause gender discrimination and 752 753 negatively impact women's careers.

754 Overall, our findings demonstrate that the perception of gender stereotypes in large language models closely mirrors 755 societal views. However, the models sometimes challenge stereotypes, such as depicting that men can cook. Furthermore, 756 it is interesting to highlight that some gender biases may not be perceived by the model; for example, we did not 757 758 observe women as bad drivers stereotypes which was observed in our survey data. Our open-generation methodology 759 enhances our understanding of the model's perception of stereotypes, particularly in relation to occupational roles and 760 nationality. Additionally, by utilizing computational grounded theory, we are able to approach this problem without 761 relying on predefined ideas or criteria for analyzing stereotype perception. Based on the initial step of generations, we 762 763 observed patterns that were consolidated by manual reading, followed by validation and comparison with the society. 764 By applying CGT, we are providing a comprehensive framework to assess stereotypes, biases, and social theories of 765 gender studies in the models qualitatively and empirically. 766

7 Limitations

767

768

769 The limitations of the approach employed in this study are multi-faceted. Firstly, the generation process exhibits a 770 degree of prompt dependency, wherein the quality and relevance of responses heavily rely on the specificity and 771 structure of the input prompts. Moreover, we would like to highlight that currently, the methods applied for analyzing 772 773 stereotype perceptions are inadequate for enumerating all biases within language models. Our study predominantly 774 serves as an example of a stereotype identification tool rather than a comprehensive assessment of bias. In the pursuit 775 of evaluating bias, it is evident that more sophisticated and nuanced techniques are required, surpassing the simplistic 776 nature of sentiment analysis. Furthermore, the recognition of diverse gender identities is a critical aspect of ensuring 777 778 fairness and inclusivity in natural language processing and AI systems. We emphasize that the broader spectrum of 779 gender identities and expressions must be considered in future research to develop more comprehensive and equitable 780 Manuscript submitted to ACM

LLMs. Hence, while this study offers insights into methods for identifying perceived stereotypes in open-ended prompts 781 782 and the alignment of those stereotypes to human perceptions, it is crucial to acknowledge its inherent limitations and 783 seek more robust methods for comprehensive analysis. 784

8 Conclusion

People interact with LLMs in many ways, including in ways that are intended to obtain the "perspective" of the LLM 788 789 itself. Thus, it is important to examine models' perceptions of stereotypes, not only biases in the text they generate. We 790 illustrate an approach that builds upon qualitative and quantitative methodologies to assess and interpret data through 791 computational grounded theory. CGT strikes a balance between human interpretation and computational capabilities, 792 793 an important consideration given the varying nature of stereotypes across different societies and perspectives [46, 76]. 794 By simply modifying the prompts, this framework could be adapted to identify potential biases in various contexts in 795 an open-ended way. From the survey findings, we observed that the patterns identified in the model's output were 796 largely also present in our participants' perceptions. Based on these findings, we expect that future LLMs, including 797 798 those more sophisticated than what we examined in this paper, are likely to hold similar perceptions of stereotypes 799 since they inherently learn from the data generated by humans. As the integration of large language models accelerates, 800 it becomes imperative to recognize, understand, and address stereotype patterns within them. Such a task cannot be 801 accomplished solely through quantitative or qualitative means. Multi-method efforts hold the promise of measuring 802 even unexpected biases, shaping a more inclusive and equitable technological landscape for the future. 803

804 805

807

808

809

810

811

812

815

816

817

818

819

820

821

822

References 806

- [1] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (Virtual Event, USA) (AIES '21). Association for Computing Machinery, New York, NY, USA, 298-306. doi:10.1145/3461702.3462624
- [2] Afra Feyza Akyürek, Muhammed Yusuf Kocyigit, Sejin Paik, and Derry Tanti Wijaya. 2022. Challenges in Measuring Bias via Open-Ended Language Generation. In Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP). Association for Computational Linguistics, Seattle, Washington, 76-76. doi:10.18653/v1/2022.gebnlp-1.9
- [3] Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken AC G van der Velden. 2022. Three gaps in computational text analysis methods 813 for social sciences: A research agenda. Communication Methods and Measures 16, 1 (2022), 1-18. 814
 - [4] Christopher A Bail. 2014. The cultural environment: Measuring culture with big data. Theory and Society 43 (2014), 465-482.
 - [5] Damian Bebell and Steven E Stemler. 2004. Reassessing the objectives of educational accountability in Massachusetts: The mismatch between Massachusetts and the MCAS. In Annual Meeting of the American Educational Research Association. San Diego, CA.
 - [6] 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 (Virtual Event, Canada) (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610-623. doi:10.1145/3442188.3445922
 - [7] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. Advances in neural information processing systems 13 (2000).
 - [8] Richard Biernacki. 2012. Reinventing evidence in social inquiry: Decoding facts and variables. Springer.
 - [9] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. Journal of machine Learning research 3, Jan (2003), 993-1022.

823 [10] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In 824 Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault 825 (Eds.), Association for Computational Linguistics, Online, 5454-5476, doi:10.18653/v1/2020.acl-main.485

- 826 [11] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of 827 Pitfalls in Fairness Benchmark Datasets. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th 828 International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Online, 1004–1015. 829 doi:10.18653/v1/2021.acl-long.81
- [12] Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media african-american english. 830 arXiv preprint arXiv:1707.00061 (2017). 831

785 786

- 833 [13] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to 834 homemaker? debiasing word embeddings. Advances in neural information processing systems 29 (2016).
- 835 [14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish 836 Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, 837 Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In Advances in Neural 838 Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. 839 https://proceedings.neurips.cc/paper files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf 840
- [15] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of 841 the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81), Sorelle A. Friedler and Christo 842 Wilson (Eds.). PMLR, 77-91. https://proceedings.mlr.press/v81/buolamwini18a.html
- 843 [16] Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. 2022. Gender Bias in Word Embeddings: A 844 Comprehensive Analysis of Frequency, Syntax, and Semantics. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (Oxford, 845 United Kingdom) (AIES '22). Association for Computing Machinery, New York, NY, USA, 156-170. doi:10.1145/3514094.3534162
- 846 [17] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. 847 Science 356, 6334 (2017), 183-186.
- [18] Hjalmar Bang Carlsen and Snorre Ralund. 2022. Computational grounded theory revisited: From computer-led to computer-assisted text analysis. 848 Big Data & Society 9, 1 (2022), 20539517221080146. 849
- [19] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. arXiv preprint arXiv:2006.14799 (2020). 850
- [20] Kathy Charmaz. 2014. Constructing grounded theory (introducing qualitative methods series). Constr. grounded theory (2014). 851
- [21] Mayukh Das and Wolf Tilo Balke. 2022. Quantifying Bias from Decoding Techniques in Natural Language Generation. In Proceedings of the 29th 852 International Conference on Computational Linguistics, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun 853 Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, 854 Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, 855 Gyeongju, Republic of Korea, 1311–1323. https://aclanthology.org/2022.coling-1.112
- 856 [22] Jeffrey Dastin. 2022. Amazon scraps secret AI recruiting tool that showed bias against women. In Ethics of data and analytics. Auerbach Publications, 857 296 - 299
- 858 [23] Mark Davies. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. Literary and linguistic computing 25, 4 (2010), 447-464. 859
- [24] Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and Skew: Quantifying 860 Gender Bias in Pre-trained and Fine-tuned Language Models. In Proceedings of the 16th Conference of the European Chapter of the Association for 861 Computational Linguistics: Main Volume. Association for Computational Linguistics, Online, 2232-2242. doi:10.18653/v1/2021.eacl-main.190 862
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language [25] 863 Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language 864 Technologies, Volume 1 (Long and Short Papers), Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 865 Minneapolis, Minnesota, 4171-4186. doi:10.18653/v1/N19-1423
- 866 [26] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Virtual Event, Canada) (FAccT '21). Association for Computing Machinery, New York, NY, USA, 862–872. doi:10.1145/3442188.3445924
- [27] Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. 2014. Content-based citation analysis: The next 869 generation of citation analysis. Journal of the association for information science and technology 65, 9 (2014), 1820-1833. 870
- John H Evans. 2002. Playing god?: human genetic engineering and the rationalization of public bioethical debate. University of Chicago Press. [28] 871
- [29] Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A Community-in-the-Loop Benchmark for Anti-872 LGBTO+ Bias in Large Language Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long 873 Papers). Association for Computational Linguistics, Toronto, Canada, 9126-9140. doi:10.18653/v1/2023.acl-long.507
- 874 [30] Myra Marx Ferree. 2002. Shaping abortion discourse: Democracy and the public sphere in Germany and the United States. Cambridge University Press.
- 875 [31] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 876 2024. Bias and Fairness in Large Language Models: A Survey. Computational Linguistics 50, 3 (Sept. 2024), 1097-1179. doi:10.1162/coli a 00524
- 877 [32] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 878 2020. The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027 (2020).
- Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. 879 [33] Proceedings of the National Academy of Sciences 115, 16 (2018), E3635-E3644. 880
- [34] Robert P Gauthier, Mary Jean Costello, and James R Wallace. 2022. "I Will Not Drink With You Today": A Topic-Guided Thematic Analysis of 881 Addiction Recovery on Reddit. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI '22). 882 Association for Computing Machinery, New York, NY, USA, Article 20, 17 pages. doi:10.1145/3491102.3502076 883
- 884

867

Salvi and Bosch

- [35] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Real Toxicity Prompts: Evaluating Neural Toxic Degeneration 885 in Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, 887 3356-3369. doi:10.18653/v1/2020.findings-emnlp.301 Mathew Gillings and Andrew Hardie, 2023. The interpretation of topic models for scholarly analysis: An evaluation and critique of current practice. [36] 888 Digital Scholarship in the Humanities 38, 2 (2023), 530-543. 889 [37] Barney G Glaser, Anselm L Strauss, and Elizabeth Strutzel. 1968. The discovery of grounded theory; strategies for qualitative research. Nursing 890 research 17, 4 (1968), 364. 891 [38] Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova, 2023. Generative language models and 892 automated influence operations: Emerging threats and potential mitigations. arXiv preprint arXiv:2301.04246 (2023). 893 [39] Henrich R Greve, Hayagreeva Rao, Paul Vicinanza, and Echo Yan Zhou. 2022. Online conspiracy groups: Micro-bloggers, bots, and coronavirus 894 conspiracy talk on Twitter. American Sociological Review 87, 6 (2022), 919-949. 895 [40] Vipul Gupta, Pranav Narayanan Venkit, Hugo Laurençon, Shomir Wilson, and Rebecca J Passonneau. 2023. CALM: A Multi-task Benchmark for Comprehensive Assessment of Language Model Bias. arXiv preprint arXiv:2308.12539 (2023). 896 Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and 897 [41] Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. arXiv preprint arXiv:2302.09210 (2023). 898 Ashlee Humphreys and Rebecca Jen-Hui Wang. 2018. Automated text analysis for consumer research. Journal of Consumer Research 44, 6 (2018), [42] 899 1274 - 1306900 [43] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as 901 Barriers for Persons with Disabilities. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Dan Jurafsky, Joyce 902 Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 5491–5501. doi:10.18653/v1/2020.acl-main.487 903 [44] Carina Jacobi, Wouter Van Atteveldt, and Kasper Welbers. 2018. Quantitative analysis of large amounts of journalistic texts using topic modelling. 904 In Rethinking Research Methods in an Age of Digital Journalism, Routledge, 89–106. 905 [45] Hyangeun Ji, Insook Han, and Yujung Ko. 2023. A systematic review of conversational AI in language education: Focusing on the collaboration 906 with human teachers. Journal of Research on Technology in Education 55, 1 (2023), 48-63. [46] David J Johnson and William J Chopik. 2019. Geographic variation in the black-violence stereotype. Social Psychological and Personality Science 10, 907 3 (2019), 287-294. 908 [47] Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender Bias in Masked Language Models for Multiple 909 Languages. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language 910 Technologies, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, 911 United States, 2740-2750, doi:10.18653/v1/2022.naacl-main.197 912 Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. 2021. Bias [48] 913 out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. Advances in neural information 914 processing systems 34 (2021), 2611-2624. 915 [49] Klaus Krippendorff. 2018. Content analysis: An introduction to its methodology. Sage publications.
- [50] Chen Li, Xiaochun Zhang, Dimitrios Chrysostomou, and Hongji Yang. 2022. Tod4ir: A humanised task-oriented dialogue system for industrial 916 robots. IEEE Access 10 (2022), 91631-91649. 917
- [51] Pengxiang Li, Hichang Cho, Yuren Qin, and Anfan Chen. 2021. # MeToo as a connective movement: examining the frames adopted in the anti-sexual 918 harassment movement in China. Social Science Computer Review 39, 5 (2021), 1030-1049. 919
- [52] Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A Survey on Fairness in Large Language Models. arXiv preprint arXiv:2308.10149 920 (2023).921
 - [53] Liam Magee, Lida Ghahremanlou, Karen Soldatic, and Shanthi Robertson. 2021. Intersectional bias in causal language models. arXiv preprint arXiv:2107.07691 (2021).
- 923 [54] Federico Mangiò, Marco Mismetti, Elena Lissana, and Daniela Andreini. 2023. That's the Press, Baby! How journalists co-create family business 924 brands meanings: A mixed method analysis. Journal of Business Research 161 (2023), 113842.
- 925 [55] Justus Mattern, Zhijing Jin, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Understanding Stereotypes in Language Models: Towards Robust Measurement and Zero-Shot Debiasing. arXiv preprint arXiv:2212.10678 (2022). 926
- [56] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. 927 ACM computing surveys (CSUR) 54, 6 (2021), 1-35. 928
- [57] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. 929 PloS one 15, 8 (2020), e0237861. 930
- [58] Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. 931 Crowdsourcing and language studies: the new generation of linguistic data. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and 932 Language Data with Amazon's Mechanical Turk, Chris Callison-Burch and Mark Dredze (Eds.). Association for Computational Linguistics, Los 933 Angeles, 122-130, https://aclanthology.org/W10-0719/
- 934 Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In Proceedings of the 59th 935 Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume
- 936 Manuscript submitted to ACM

18

886

- 937 1: Long Papers), Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 5356-5371. 938 doi:10.18653/v1/2021.acl-long.416
- 939 [60] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in 940 Masked Language Models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for 941 Computational Linguistics, Online, 1953-1967. doi:10.18653/v1/2020.emnlp-main.154
- 942 Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality Bias in Text Generation. [61] In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Andreas Vlachos and Isabelle Augenstein 943 (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 116-122. doi:10.18653/v1/2023.eacl-main.9 944
- [62] Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. J. Data and Information 945 Quality 15, 2, Article 10 (June 2023), 21 pages. doi:10.1145/3597307 946
 - [63] Laura K Nelson. 2020. Computational grounded theory: A methodological framework. Sociological Methods & Research 49, 1 (2020), 3-42.
- 947 [64] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health 948 of populations. Science 366, 6464 (2019), 447-453.
- 949 [65] Hadas Orgad and Yonatan Belinkov. 2022. Choose Your Lenses: Flaws in Gender Bias Evaluation. In Proceedings of the 4th Workshop on Gender Bias 950 in Natural Language Processing (GeBNLP), Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen (Eds.). 951 Association for Computational Linguistics, Seattle, Washington, 151-167. doi:10.18653/v1/2022.gebnlp-1.17
- [66] Nornadiah Mohd Razali, Yap Bee Wah, et al. 2011. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. 952 Fournal of statistical modeling and analytics 2, 1 (2011), 21-33. 953
 - [67] Cecilia L Ridgeway and Shelley J Correll. 2004. Motherhood as a status characteristic. Journal of Social issues 60, 4 (2004), 683-700.
- [68] Heber Rodrigues, Carlos Goméz-Corona, and Dominique Valentin. 2020. Femininities & masculinities: sex, gender, and stereotypes in food studies. 955 Current Opinion in Food Science 33 (2020), 156-164. 956
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In Proceedings of the [69] 957 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short 958 Papers). Association for Computational Linguistics, New Orleans, Louisiana, 8-14. doi:10.18653/v1/N18-2002 959
 - [70] Johnny M. Saldana. 2015. The coding manual for qualitative researchers (3 ed.). SAGE Publications, London, England.
- 960 [71] Michael Scharkow. 2013. Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. Quality 961 & Quantity 47 (2013), 761-773.
- 962 [72] John V. Seidel. 1998. Qualitative Data Analysis. www. qualisresearch. com (originally published as Qualitative Data Analysis, in The Ethnograph v5 Qualis Research) (1998). 963
- [73] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language 964 Generation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on 965 Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 3407-3412. doi:10.18653/v1/D19-1339 966
 - [74] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053 (2019).
 - [75] Shoshanna Sofaer. 1999. Qualitative methods: what are they and why use them? Health services research 34, 5 Pt 2 (1999), 1101.
- [76] Adrian Stanciu, J Christopher Cohrs, Katja Hanke, and Alin Gavreliuc. 2017. Within-culture variation in the content of stereotypes: Application and 970 development of the stereotype content model in an Eastern European culture. The Journal of Social Psychology 157, 5 (2017), 611-628.
- 971 [77] Steven E Stemler. 2015. Content analysis. Emerging trends in the social and behavioral sciences: An Interdisciplinary, Searchable, and Linkable Resource 972 (2015), 1-14.
- [78] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. Advances in neural information processing 973 systems 27 (2014). 974
- Alexey Svyatkovskiy, Shao Kun Deng, Shengyu Fu, and Neel Sundaresan. 2020. IntelliCode compose: code generation using transformer. In [79] 975 Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering 976 (Virtual Event, USA) (ESEC/FSE 2020), Association for Computing Machinery, New York, NY, USA, 1433-1443, doi:10.1145/3368089.3417058 977
- [80] Naznin Tabassum and Bhabani Shankar Nayak. 2021. Gender stereotypes and their impact on women's career progressions from a managerial 978 perspective. IIM Kozhikode Society & Management Review 10, 2 (2021), 192-208.
- 979 [81] Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language 980 models. arXiv preprint arXiv:2102.02503 (2021).
- 981 [82] Paul Tschisgale, Peter Wulff, and Marcus Kubsch. 2023. Integrating artificial intelligence-based methods into qualitative research in physics 982 education research: A case for computational grounded theory. Physical Review Physics Education Research 19, 2 (2023), 020123.
- [83] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is 983 all you need. Advances in neural information processing systems 30 (2017). 984
- [84] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a Warm Person, Joseph is a Role Model": 985 Gender Biases in LLM-Generated Reference Letters. In Findings of the Association for Computational Linguistics: EMNLP 2023, Houda Bouamor, Juan 986 Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3730–3748. doi:10.18653/v1/2023.findings-emnlp.243 987
- 988

954

967

968

- [85] David James Woo, Yanzhi Wang, and Hengky Susanto. 2022. Student-AI Creative Writing: Pedagogical Strategies for Applying Natural Language
 Generation in Schools. *EdArXiv. June* 3 (2022).
- 991[86]Frank F. Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In992Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming (San Diego, CA, USA) (MAPS 2022). Association for993Computing Machinery, New York, NY, USA, 1–10. doi:10.1145/3520312.3534862
 - [87] 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). Association for Computational Linguistics, New Orleans, Louisiana, 15–20. doi:10.18653/v1/N18-2003

A Appendix

Table A1.	Age Distribution	of Participants	for Study 2

Statistic	Value
Mean Age	39.81 years
Standard Deviation	13.49 years
Minimum Age	18 years
Median Age	37 years
Maximum Age	76 years

Table A2. Gender Distribution of Participants for Study 2

Gender	Count
Male	221
Female	167
Nonbinary	5
Trans Man	1
Bigender	1
Enby	1
Two Spirit	1
Female / Non-binary	1
They	1

1041	Table A3.	Top 30 word	s from Ste	p 1 for	womer	in "in" cate	egor
1042		-		-			-
1043	ſ	Word	Count	Neg	Pos	Neu	
1044	-	work	75	33	30	12	
1045		tech	63	32	23	8	
1046		good	43	19	17	7	
1047		either	41	25	10	6	
1048		industry	38	18	13	7	
1049		science	32	14	11	7	
1050		get	32	16	14	2	
1051		united	30	14	11	5	
1052	_	states	30	14	10	6	
1053	_	dont	28	21	1	6	
1054	_	business	25	14	8	3	
1055	_	game	25	8	11	6	
1056	Ļ	girl	22	3	13	6	
1057	ļ	time	20	10	8	2	
1058	F	india	19	9	6	4	
1059	-	workplace	19	12	5	2	
1060	-	like	18	9	7	2	
1061		world	16	10	4	5	
1062	-	america	10	10	5	5	
1063	-	games	15	5	5	5	
1003	-	military	14	7	7	1	
1064	-	weak	14	12	0	2	
1065	-	gaming	13	15	7	2	
1066	ŀ	sports	12	5	2	5	
1067	ŀ	hollywood	12	7	3	2	
1068	-	true	12	7	2	3	
1069	-	often	12	7	4	1	
1070	ŀ	lazy	11	11	0	0	
1071	ŀ	long	11	3	5	3	
1072	L	0		1			
1073							
1074							
1075							
1076							
1077							
1078							
1079							
1080							
1081							
1082							
1083							
1084							
1085							
1086							

Table A4. Top 30 words from Step 1 for women in "is" category

Word	Count	Neg	Pos	Neu
go	51	28	18	5
work	41	21	10	10
good	34	18	15	1
get	33	17	7	9
either	30	21	7	2
like	28	16	8	4
weak	24	23	0	1
dont	23	20	0	3
cook	21	12	5	4
care	19	10	8	1
home	19	5	9	5
passive	18	15	1	2
mother	18	5	11	2
bit	18	16	2	0
want	17	7	7	3
job	17	4	5	8
strong	16	3	11	2
take	16	8	6	2
hard	15	10	5	0
girl	13	3	9	1
emotional	12	9	0	3
little	12	6	5	1
often	12	6	2	4
cant	12	12	0	0
true	12	9	3	0
time	12	7	4	1
housewife	11	1	9	1
lazy	9	9	0	0
need	9	3	2	4
family	8	1	3	4

> > violent

even

go

care

Salvi and Bosch

Table A6. Top 30 words from Step 1 for men in "in" category

	Count	Neg	Pos	Net
work	40	13	19	8
man	38	11	17	10
men	26	10	12	4
workplace	18	8	6	4
guy	17	3	8	6
india	15	4	7	4
either	14	6	5	3
like	14	5	5	4
hard	13	4	6	3
relationship	13	4	7	2
lazy	12	12	0	0
hollywood	12	4	4	4
world	12	4	7	1
good	12	4	5	3
east	11	5	3	3
dont	10	9	0	1
industry	10	1	8	1
united	10	8	2	0
america	10	4	2	4
tech	10	2	7	1
sex	10	6	2	2
working	10	3	6	1
live	10	2	6	2
states	9	7	2	0
want	9	3	5	1
general	9	2	4	3
relationships	8	3	3	2
middle	7	2	3	2
military	6	3	2	1
society	5	4	1	0

Word	Count	Neg	Pos	Neu
smart	98	22	1	75
strong	73	1	1	71
nerds	63	56	0	7
aggressive	51	30	0	21
good	47	4	2	41
nerdy	44	43	0	1
tough	39	0	0	39
better	33	1	0	32
leaders	32	0	1	31
competitive	30	5	0	25
know	30	6	0	24
work	30	3	1	26
ruthless	29	25	0	4
get	29	10	0	19
make	28	4	3	21
money	27	13	0	14
business	23	1	0	22
sports	22	4	2	16
always	22	4	2	16
like	21	6	0	15
everything	20	8	0	12
dumb	20	20	0	0
intelligent	19	6	0	13
much	16	9	0	7
socially	15	15	0	0
love	15	2	9	4
athletic	15	1	0	14
stronger	15	0	0	15
social	15	13	0	2
steroids	15	8	0	7

Table A7. Top 30 words from Survey for men in "in" category

Table A8. Top 30 words from Survey for women in "in" category

woru	Count	Neg	Pos	Nei
enough	114	112	0	2
smart	74	54	1	19
strong	64	47	1	16
less	57	32	0	25
good	47	41	0	6
weak	46	46	0	0
get	35	14	0	21
tough	34	23	0	11
lesbians	30	30	0	0
emotional	26	18	0	8
masculine	24	1	0	23
know	23	19	0	4
cannot	23	20	0	3
handle	22	20	0	2
science	21	15	0	6
work	21	6	0	15
capable	19	16	0	3
business	18	13	0	5
combat	17	14	0	3
make	17	9	0	8
hired	17	7	0	10
weaker	16	15	0	1
nerds	15	15	0	0
competitive	15	5	0	10
diversity	14	2	0	12
belong	13	13	0	0
play	13	8	0	5
butch	13	4	0	9
physically	13	10	0	3
11	12	12	0	0

Word	Count	Neg	Pos	Net
emotional	167	58	0	109
weak	91	91	0	0
good	75	38	1	36
like	67	18	2	47
bad	61	61	0	0
want	59	5	0	54
children	53	2	3	48
drivers	53	53	0	0
care	42	6	0	36
home	41	1	0	40
nurturing	37	0	0	37
get	33	13	0	20
cook	33	0	0	33
take	31	6	0	25
sensitive	30	12	0	18
stay	28	1	0	27
math	27	24	0	3
always	27	6	1	20
love	26	1	14	11
better	26	1	0	25
mothers	23	1	3	19
kids	23	1	1	21
work	22	8	0	14
less	20	12	0	8
gossip	19	16	0	3
belong	19	2	0	17
feminine	19	4	0	15
overly	18	18	0	0

make

Salvi and Bosch

Table A10. Top 30 words from Survey for men in category "is"

Word	count	Neg	Pos	Neu
strong	137	1	0	136
like	80	13	2	65
aggressive	64	3	0	61
sports	60	0	9	51
cry	58	5	0	53
sex	52	2	2	48
emotions	46	23	0	23
good	43	14	1	28
tough	42	0	0	42
show	41	20	0	21
emotional	38	9	0	29
always	34	3	0	31
family	34	1	1	32
leaders	27	0	1	26
work	26	1	1	24
care	25	12	0	13
love	24	2	18	4
physically	24	0	0	24
emotion	23	10	0	13
better	23	0	0	23
need	23	0	0	23
violent	22	17	0	5
want	21	6	0	15
things	20	3	0	17
less	18	8	0	10
breadwinners	18	1	0	17
unemotional	18	16	0	2
dominant	17	0	0	17
never	17	5	0	12

1412	
1413	
1414	
1415	
1416	
1417	
1418	
1419	
1420	
1421	
1422	
1423	
1424	
1425	
1426	
1427	
1428	
1429	
1430	
1431	
1432	
1433	
1434	
1435	
1436	
1437	
1438	
1439	
1440	
1441	
1442	
1443	
1444	
1445	
1446	
1447	
1448	
1449	
1450	
1451	
1452	
1453	
1454	
1455	
1456	Manuscript submitted to ACM