Reflecting the Male Gaze: Quantifying Female Objectification in 19th and 20th Century Novels

Kexin Luo, Yue Mao, Bei Zhang, and Sophie Hao
Center for Data Science, New York University

Abstract

Inspired by the concept of the *male gaze* (Mulvey, 1975) in literature and media studies, this paper proposes a framework for analyzing gender bias in terms of *female objectification*—the extent to which a text portrays female individuals as objects of visual pleasure. Our framework measures female objectification along two axes. First, we compute an *agency bias* score that indicates whether male entities are more likely to appear in the text as grammatical agents than female entities. Next, by analyzing the word embedding space induced by a text (Caliskan et al., 2017), we compute an *appearance bias* score that indicates whether female entities are more closely associated with appearance-related words than male entities. Applying our framework to 19th and 20th century novels reveals evidence of female objectification in literature: we find that novels written from a male perspective systematically objectify female characters, while novels written from a female perspective do not exhibit statistically significant objectification of any gender.

Keywords: bias, gender, stereotypes, sexism, representational harms, literature, humanities

1. Introduction

In literature and media studies, the *male gaze* (Mulvey, 1975) refers to a phenomenon in which women are depicted in film, literature, and the visual arts as objects of aesthetic pleasure, to be consumed and enjoyed by a heterosexual male viewer. The male gaze, and the practice of *female objectification* more generally, perpetuates an understanding of women as tools meant to serve the interests of others, with little attention paid to women's agency, individuality, or subjectivity (Nussbaum, 1995). Beyond these representational harms (Barocas et al., 2017; Crawford, 2017), there is evidence that internalization of female objectification can result in averse mental health impacts on girls and women (Fredrickson and Roberts, 1997; Szymanski et al., 2011). Despite its harms, however, female objectification seems to be ubiquitous in Western culture. Within the computational linguistics literature, male-gaze-like depictions of women have been documented in film and television (Agarwal et al., 2015; Singh et al., 2023), Tweets (Anzovino et al., 2018), internet memes (Fersini et al., 2022; Singh et al., 2023), text corpora (da Cunha and Abeillé, 2022), text generated by language models (Lucy and Bamman, 2021; Wolfe et al., 2023), and even theoretically "gender-neutral" media such as linguistics textbooks (Macaulay and Brice, 1997) and journal articles (Kotek et al., 2020).

This paper proposes a quantitative framework for studying female objectification in text, leveraging techniques developed for the analysis of gender bias in natural language processing (NLP). We operationalize the concept of female objectification by factoring it into two biases that a text might exhibit: an *agency bias* that favors treating male entities as grammatical agents, and an *appearance bias* that favors mentioning female entities in collocation with words related to appearance. To measure these biases within a collection of texts, we develop a pipeline that combines several NLP tools: we measure agency bias by using a semantic role labeler (Shi and Lin, 2019) to analyze the argument structure of male and female entities in a text, and we measure appearance bias by using the Word Embedding Association Test (Caliskan et al., 2017) to analyze stereotypical associations in the word embedding space induced by the text.

We apply our framework by presenting a quantitative study of female objectification in English-language novels and translations from the late modern era. We show that commonly-downloaded open-source novels exhibit positive, statistically significant levels of agency bias and appearance bias, revealing the existence of systematic female objectification within this repertoire. When controlling for the gender of authors and narrators, we find that novels with a female author or a first-person female narrator do not exhibit statistically significant levels of either agency bias or appearance bias—some are strongly biased, but most are weakly biased in the opposite direction (i.e., they exhibit mild levels of *male objectification*). On the other hand, novels with a male author or a first-person male narrator consistently exhibit both agency bias and appearance bias.

To summarize, our contributions are as follows. In Section 2, we define two bias metrics for texts, which formalize the concept of female objectification. In Section 3, we develop a methodology for studying female objectification in a collection of texts. Finally, in Section 4, we present empirical evidence of a male gaze phenomenon in 19th and 20th century English-language novels.[1]

[1] Our code is available at https://github.com/Bellaaazzzzz/Female_Objectification_

LREC-COLING 2024, pages 13803–13812
20-25 May, 2024.

Alice saw *Bob* at the park. **She** waved to *him* and said, “Hello!” **Bob** smiled and walked over.

Male Agentivity:	$1/3$	$=$	$.33$
Female Agentivity:	$2/2$	$=$	1.00
Agency Bias:	$\frac{1/3}{2/2} - 1$	$=$	$-.67$

Figure 1: Calculating agency bias for a short story by counting occurrences of **female agents**, **male agents**, and *male patients*.

2. Measuring Female Objectification

Our treatment of female objectification is inspired by the following elements of the male gaze: (1) that women are depicted as *passive* objects, (2) to be consumed for *visual* pleasure. We operationalize these two concepts by defining the following bias metrics.

1. **Agency Bias:** A text exhibits *agency bias* if male entities are more likely than female entities to appear in the text as grammatical agents.
2. **Appearance Bias:** A text exhibits *appearance bias* if “female” words are distributionally closer to “appearance” words than “male” words.

For both bias metrics, a value of 0 represents lack of bias, a positive value represents presence of female objectification, and a negative value represents presence of male objectification.

2.1. Agency Bias

The *grammatical agent* of a clause is the entity that initiates the event denoted by the main predicate of that clause. For example, in the sentence *Bob was seen by Alice*, *Alice* is the agent, since Alice initiates the act of seeing. *Bob*, the person who receives the act of seeing, is the *patient*. Our agency bias metric is based on the intuition that if a text portrays women as passive objects, then female entities are less likely than male entities to appear in the text as grammatical agents.

2.1.1. Definition of Agency Bias

We define the *female agentivity* (resp. *male agentivity*) of a text as the conditional probability that an entity appears in the text as a grammatical agent, given that it is female (resp. male). The agency bias of a text is defined as follows:

$$\text{agency bias} = \frac{\text{male agentivity}}{\text{female agentivity}} - 1.$$

Example. Figure 1 illustrates how agency bias is calculated for a short story. We estimate male agentivity to be $1/3$, since there are three occurrences of a male entity (Bob), and one of them is an agent. The female agentivity of this text is 1, since the sole female entity in this story (Alice) acts as an agent in both occurrences. The final agency bias is $-2/3$, meaning that female entities are 67% more likely than male entites to appear in this story as agents.

2.1.2. Calculating Agency Bias

To calculate agency bias for a text at scale, we use a procedure consisting of the following steps.

- **Entity Extraction:** We extract entities from the text using spaCy’s named entity recognizer.
- **Gender Classification:** We classify these entities as *male* or *female* using a procedure similar to the one used in Toro Isaza et al. (2023).
- **Agency Classification:** We classify the entities as *agents* or *non-agents* using Shi and Lin’s (2019) semantic role labeler.

The full details of our implementation are given in Appendix A.

2.2. Appearance Bias

The Word Embedding Association Test (WEAT, Caliskan et al., 2017) quantifies the extent to which a word embedding space conveys stereotypical associations between demographic groups. Our appearance bias metric uses WEAT to compare “male” and “female” words in terms of their similarity to “appearance” words. If a text depicts women as bearers of visual pleasure, then we expect the female words to be closer than male words to the appearance words when an embedding space is trained on that text.

2.2.1. Definition of Appearance Bias

Let M, F, and A be sets of *male words*, *female words*, and *appearance words*, respectively.[2] Let $\mathbb{W}$ be a word embedding space, where the embedding of a word w is denoted by $\vec{w}$. The *WEAT score for* $\mathbb{W}$ is defined as the quantity

$$\operatorname{weat}(\mathbb{W}) = \frac{\operatorname{mean}_{a \in A} s(a)}{\operatorname{stdev}_{a \in A} s(a)},$$

where

$$s(a) = \operatorname*{mean}_{f \in F} \cos(\vec{f}, \vec{a}) - \operatorname*{mean}_{m \in M} \cos(\vec{m}, \vec{a}).$$

Quantifying_in_Novels.

[2] Unlike Caliskan et al. (2017), we only use one set of target words.

Male Words (M)	boy, brother, father, he, him, himself, husband, male, man, mr, sir, uncle, male named entities
Female Words (F)	aunt, female, girl, her, herself, lady, miss, mother, she, sister, wife, woman, female named entities
Appearance Words (A)	belt, complexion, dress, eye, lip, outfit, plain, pore, purse, ravishing, ugly, voluptuous, 992 others

Table 1: Words used to calculate WEAT scores for the appearance bias metric.

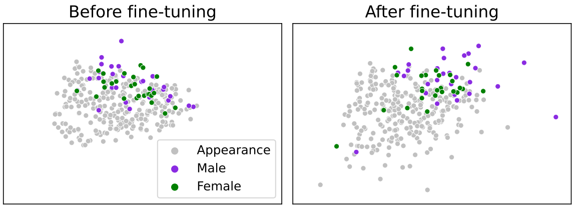

Figure 2: GloVe embeddings (Pennington et al., 2014) for words in Table 1, plotted along two principal components. Fine-tuning the embeddings on the novel *Lady Audley's Secret* by Mary Elizabeth Braddon causes "male" words, but not "female" words, to move away from the cluster of "appearance" words. The novel's appearance bias is 1.575.

To compute the appearance bias of a text, we take a pre-trained embedding space $\mathbb{W}$, and fine-tune it on the text. Letting $\mathbb{W}'$ denote the fine-tuned embedding space, the appearance bias of the text is defined as:

$$\text{appearance bias} = \mathrm{weat}(\mathbb{W}') - \mathrm{weat}(\mathbb{W}).$$

Word Sets. Table 1 shows the male, female, and appearance words used to calculate WEAT scores on our pre-trained and fine-tuned embedding spaces. The full set of appearance words is obtained from the Oxford Learner's Dictionaries' "Appearance" topic vocabulary.[3] The male and female words include those listed in Table 1, as well as named entities that can be assigned a gender.

Example. Figure 2 visualizes how a word embedding space changes after it has been fine-tuned on a novel. In this example, the three word clusters overlap in the pre-trained embedding space, but the male words drift away from the appearance words after fine-tuning. The appearance bias of this text is 1.575, indicating that this text associates females with appearance more than males.

2.2.2. Calculating Appearance Bias

Here we briefly describe elements of our procedure for calculating the appearance bias of a text. Full implementation details are given in Appendix A.

Gendered Entities. Similar to our agency bias pipeline, our calculation of appearance bias includes entity extraction and gender classification steps that produce the named entities included among the male and female words.

Embedding Spaces. We use `glove.6B` embeddings (Pennington et al., 2014), pre-trained on Wikipedia articles and the Gigaword corpus (Parker et al., 2011). We fine-tune our embeddings using the CBOW objective with negative sampling (Mikolov et al., 2013a,b). Since proper nouns may be out of vocabulary, we randomly (re-)initialize the embeddings for named entities before fine-tuning.

3. The Male Gaze in Literature

Our main experiment uses our framework to quantitatively study female objectification in English-language novels and translations from the late modern period. We compile a collection of open-source novels primarily from the 19th and 20th centuries, and measure the appearance bias and agency bias of these novels. We conclude that our collection of novels exhibits evidence of systematic female objectification if we are able to reject the null hypothesis that the average appearance bias and agency bias of the novels are both 0. Our goal is to answer the following questions.

Q1. Do novels exhibit systematic female objectification in general?

Q2. Is the use of female objectification influenced by the gender of a novel's author or narrator?

3.1. Texts

The texts used in our study consist of the 100 most downloaded books from Project Gutenberg as of August 25, 2023.[4] All texts are novels written in or translated into English. After filtering out novels published before 1800, our final dataset consists of 79 novels. As shown in Figure 3, most novels in our dataset were published between 1840 and 1940, roughly coinciding with the Victorian Era and the start of World War II.

[3] `http://www.oxfordlearnersdictionaries.com/topic/category/appearance_1`

[4] `https://www.gutenberg.org/ebooks/search/?sort_order=downloads`

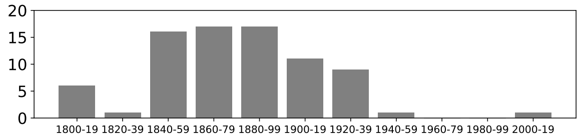

Figure 3: Distribution of publication dates for novels used in our main experiment. Most of our novels were published between the Victorian Era (1837–1901) and the start of World War II (1939–1945).

Narrator	Author			Total
	F	M	Unknown	
1p-F	7	2	0	9
1p-M	2	19	1	22
3p	13	31	1	45
Multiple	1	2	0	3
Total	23	54	2	79

Table 2: The distribution of author genders and narrative perspectives in our dataset. Narrators may be third-person (3p) or first-person (1p-F or 1p-M). "Multiple" refers to novels with more than one narrator.

3.2. Experimental Setup

Question Q2 introduces two independent variables: *author gender* and *narrator gender*. Intuitively, we expect that a novel is more likely to exhibit female objectification if it is written from a male perspective. To test this, we assume that a novel takes on a male perspective if it is written by a male author, or if it uses a male first-person narrator.

Table 2 shows the distribution of author and narrator features among our novels. We distinguish among three *narrative perspectives*: female first-person (1p-F), male first-person (1p-M), and third-person (3p). Since the gender of third-person narrators is often unspecified, our analysis of narrator gender is limited to the comparison between 1p-F and 1p-M narrators.

Hypothesis Testing. We use a one-sample t-test to determine whether systematic bias has been detected in our dataset. We test the null hypothesis that the mean agency or appearance bias of our novels is 0, and we conclude that systematic bias exists if (1) the observed mean values of both bias metrics are positive, and (2) the null hypothesis is rejected with $p < .05$ for both bias metrics.

4. Results

Our results are shown in Table 3. Both research questions are answered in the affirmative: our full set of novels shows evidence of systematic female objectification, but novels written from a female perspective—either by a female author, or using a 1p-F narrator—do not. As illustrated in Figure 5, higher agency bias is associated with higher appearance bias in general, though the correlation between the two metrics is weak.

	Agency Bias		**Appearance Bias**	
Overall	**.067**	($p < .001$)	**.176**	($p = .005$)
Authors				
F	.014	($p = .660$)	.185	($p = .267$)
M	**.090**	($p < .001$)	**.164**	($p = .004$)
Narrators				
1p-F	.095	($p = .135$)	.069	($p = .764$)
1p-M	**.144**	($p < .001$)	**.186**	($p = .015$)

Table 3: Mean agency and appearance bias scores measured in our novels dataset. Bolded results indicate evidence of systematic female objectification (mean > 0, $p < .05$).

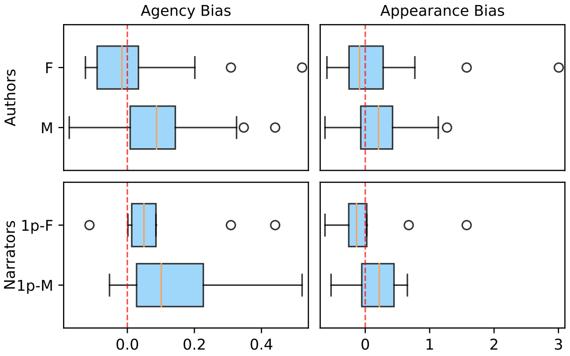

Figure 4: The distribution of agency bias and appearance bias scores for our novels, conditioned on author gender (F vs. M) and narrator gender (1p-F vs. 1p-M).

The overall presence of female objectification across the entire dataset is attributable to the fact that most novels in our dataset are written from a male perspective. Although female-perspective novels have positive mean bias scores, Figure 4 shows that this is driven by a small number of outliers. The majority of female-perspective novels actually exhibit *negative* bias scores, with the exception that novels with a 1p-F narrator mostly exhibit positive agency bias.

Figure 6 illustrates an asymmetry between female and male characters in terms of their contributions to agency bias and appearance bias. Although novels with more mentions of female characters exhibit higher levels of female agentivity and lower levels of appearance bias, no such relationship exists between the number of mentions of male characters and our bias metrics. This suggests that the majority of female characters are objectified, while only the most important female

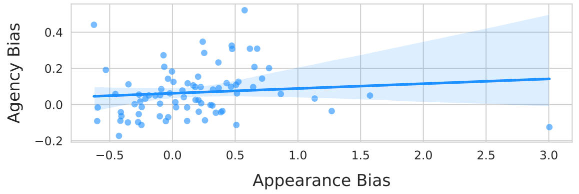

Figure 5: Our two bias metrics are weakly correlated across our novels dataset ($\rho = .104$).

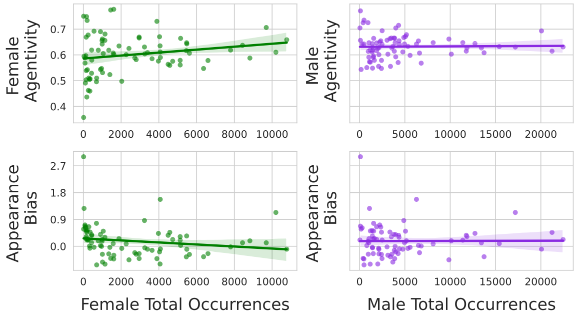

Figure 6: Novels with more mentions of female characters (x-axis) exhibit higher female agentivity and lower appearance bias.

characters assume male levels of agency and non-appearance-related properties. In contrast, even minor male characters that occur relatively infrequently assume high levels of agency, with little attention paid to their appearance.

5. Related Work

Gender and Agentivity. Gender asymmetries in grammatical agentivity have been well-documented in text corpora. Macaulay and Brice (1997) report statistics about the agentivity of gendered entities in example sentences from a linguistics textbook. Using their results, we compute an agency bias of 1.012, almost twice as high as the highest agency bias measured in our novels dataset (.521). A similar study of linguistics journal articles (Kotek et al., 2020) yields a much less extreme agency bias of .067, and da Cunha and Abeillé (2022) report similar results about grammatical subjecthood in English and French corpora. Another line of work has analyzed the kinds of events that agents initiate in films (Rashkin et al., 2018), Wikipedia entries (Sun and Peng, 2021), and fairy tales (Toro Isaza et al., 2023). Using methods similar to ours, these studies find that female agents are more likely to initiate events related to family, appearance, and sexuality, while male agents are more likely to initiate events related to work and violence.

Word Embeddings and Culture. Word embeddings are often used as summaries of cultural knowledge captured by a text corpus. Hamilton et al. (2016), for example, use embedding spaces trained on corpora from different time periods to track semantic change. Caliskan et al. (2017) and Garg et al. (2018) show that word embedding spaces capture psychologically verified gender and racial stereotypes as well as disparities in demographics and labor statistics. An application of these methods to literature is presented by Adukia et al. (2022), who show that children's books associate female characters with traits related to appearance and family relations.

Bias and Objectification. Surveys have found that most papers on NLP gender bias are not grounded in any explicit theory of gender (Devinney et al., 2022) or bias (Blodgett et al., 2020), though much work has been done on gender stereotypes in NLP models. Female objectification is, however, featured hate speech detection benchmarks (e.g., Fersini et al., 2022).

6. Conclusion

This paper has developed a quantitative framework for studying female objectification in text corpora, based on our agency bias and appearance bias metrics. Our analysis of 19th and 20th century novels has found evidence of systematic female objectification, driven by the fact that popular novels from this time period are mostly written from a male perspective. Although many of the novels in our dataset do exhibit negative agency bias and appearance bias, our aggregate results suggest that female objectification is the cultural norm for English-language novels of this time period, featured in the majority of commonly downloaded male-perspective novels as well as a large minority of female-perspective novels. Our examination of frequency effects on agentivity and appearance bias suggests that agency, individuality, and subjectivity are reserved for the most important female characters, even though these attributes are readily available to male characters, both major and minor.

Female objectification is not limited to literature or the arts. A possible application of our methods in NLP is the evaluation of female objectification in generated text, pre-training corpora, or NLP research papers. For example, agency and appearance bias scores could be reported in data cards (Mitchell et al., 2019; Gebru et al., 2021). We explore such possibilities in future work.

7. Acknowledgments

We thank the reviewers for their feedback on this paper.

8. Limitations

Quality of NLP Tools. Accurate estimation of agency bias using our analysis pipeline requires access to high-quality named entity recognizers, semantic role labelers, and gender classifiers. Agency bias results may be sensitive to differences between model instances for any of these components.

Embedding Space Fine-Tuning Epochs. The fine-tuning of embedding spaces for appearance bias calculation is hyperparameterized by the number of epochs to fine-tune for. In our study, the number of epochs was chosen in order to ensure that word embeddings are fine-tuned for the same number of training steps across novels. In other settings, appearance bias scores may not be comparable if they are computed using embeddings that have undergone different amounts of fine-tuning.

Theory of Gender. In this study, we have assumed a binary, cisnormative theory of gender. We make this assumption because, to our knowledge, none of the novels in our dataset feature a non-cisgender author or character (though there are instances of authors and characters of unknown gender). Although the definitions of agency and appearance bias do not assume cisnormativity, they do assume binarity of gender. This is a limitation of gender bias metrics more generally, since most bias metrics are designed to capture comparisons between two groups.

9. Ethical Considerations

We do not foresee any ethical issues arising from this study.

10. Bibliographical References

Anjali Adukia, Patricia Chiril, Callista Christ, Anjali Das, Alex Eble, Emileigh Harrison, and Hakizumwami Birali Runesha. 2022. Tales and Tropes: Gender Roles from Word Embeddings in a Century of Children's Books. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3086–3097, Gyeongju, South Korea. Association for Computational Linguistics.

Apoorv Agarwal, Jiehan Zheng, Shruti Kamath, Sriramkumar Balasubramanian, and Shirin Ann Dey. 2015. Key Female Characters in Film Have More to Talk About Besides Men: Automating the Bechdel Test. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 830–840, Denver, CO, USA. Association for Computational Linguistics.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In *Natural Language Processing and Information Systems*, volume 10859 of *Lecture Notes in Computer Science*, pages 57–64, Cham, Switzerland. Springer International Publishing.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem With Bias: Allocative Versus Representational Harms in Machine Learning. Oral presentation at the 9th Annual Conference of the Special Interest Group in Computing, Information, and Society, Philadelphia, PA, USA.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. *Computing Research Repository*, arXiv:2005.14165 [cs].

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

Kate Crawford. 2017. The Trouble with Bias. Invited talk at the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA.

Yanis da Cunha and Anne Abeillé. 2022. Objectifying Women? A Syntactic Bias in French and English Corpora. In *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*, pages 8–16, Marseille, France. European Language Resources Association.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of “Gender” in NLP Bias Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, pages 2083–2102, New York, NY, USA. Association for Computing Machinery.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, WA, USA. Association for Computational Linguistics.

Barbara L. Fredrickson and Tomi-Ann Roberts. 1997. Objectification Theory: Toward Understanding Women’s Lived Experiences and Mental Health Risks. *Psychology of Women Quarterly*, 21(2):173–206.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Computing Research Repository*, arXiv:1803.09010 [cs].

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Hadas Kotek, Sarah Babinski, Rikker Dockum, and Christopher Geissler. 2020. Gender representation in linguistic example sentences. *Proceedings of the Linguistic Society of America*, 5(1):514–528.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-Order Coreference Resolution with Coarse-to-Fine Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2 (Short Papers), pages 687–692, New Orleans, LA, USA. Association for Computational Linguistics.

Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Online. Association for Computational Linguistics.

Monica Macaulay and Colleen Brice. 1997. Don’t Touch My Projectile: Gender Bias and Stereotyping in Syntactic Examples. *Language*, 73(4):798–825.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. In *ICLR 2013 Workshop Proceedings*, Scottsdale, AZ. OpenReview.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119, Montreal, Canada. Curran Associates, Inc.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* ’19, pages 220–229, New York, NY, USA. Association for Computing Machinery.

Laura Mulvey. 1975. Visual Pleasure and Narrative Cinema. *Screen*, 16(3):6–18.

Martha C. Nussbaum. 1995. Objectification. *Philosophy & Public Affairs*, 24(4):249–291.

OpenAI. 2023. GPT-4 Technical Report. Technical report, OpenAI, San Francisco, CA, USA.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Computing Research Repository*, arXiv:2203.02155.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense Inference on Events, Intents, and Reactions. In *Proceedings of the 56th Annual Meeting of the Association* *for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Peng Shi and Jimmy Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *Computing Research Repository*, arXiv:1904.05255.

Smriti Singh, Tanvi Anand, Arijit Ghosh Chowdhury, and Zeerak Waseem. 2021. “Hold on honey, men at work”: A semi-supervised approach to detecting sexism in sitcoms. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 180–185, Online. Association for Computational Linguistics.

Smriti Singh, Amritha Haridasan, and Raymond Mooney. 2023. “Female Astronaut: Because sandwiches won’t make themselves up there”: Towards Multimodal misogyny detection in memes. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 150–159, Toronto, Canada. Association for Computational Linguistics.

Jiao Sun and Nanyun Peng. 2021. Men Are Elected, Women Are Married: Events Gender Bias on Wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 350–360, Online. Association for Computational Linguistics.

Dawn M. Szymanski, Lauren B. Moffitt, and Erika R. Carr. 2011. Sexual Objectification of Women: Advances to Theory and Research 1ψ7. *The Counseling Psychologist*, 39(1):6–38.

Paulina Toro Isaza, Guangxuan Xu, Toye Oloko, Yufang Hou, Nanyun Peng, and Dakuo Wang. 2023. Are Fairy Tales Fair? Analyzing Gender Bias in Temporal Narrative Event Chains of Children’s Fairy Tales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6509–6531, Toronto, Canada. Association for Computational Linguistics.

Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. 2023. Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’23, pages 1174–1185, New York, NY, USA. Association for Computing Machinery.

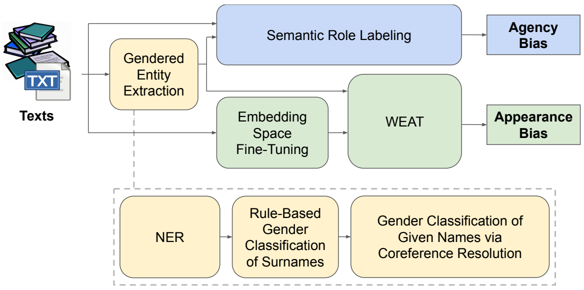
Texts
TXT
Gendered Entity Extraction
Semantic Role Labeling
Agency Bias
Embedding Space Fine-Tuning
WEAT
Appearance Bias
NER
Rule-Based Gender Classification of Surnames
Gender Classification of Given Names via Coreference Resolution

Figure 7: Pipeline for calculating the agency bias and appearance bias of a text.

Female Honorifics	Madam, Madame, Mademoiselle, Miss, Mlle, Mme, Mrs.
Male Honorifics	M., Monsieur, Mr., Sir

Table 4: Titles used in the honorific heuristic.

11. Language Resource References

Robert Parker and David Graff and Junbo Kong and Ke Chen and Kazuaki Maeda. 2011. *English Gigaword Fifth Edition*. Linguistic Data Consortium: LDC2011T07, 5, ISLRN 911-942-430-413-0.

A. Implementation Details

This appendix describes the pipeline, illustrated in Figure 7, that was used to calculate the agency bias and appearance bias of novels for the main experiment described in Section 3 and Section 4.

A.1. Gendered Entity Extraction

Both bias metrics rely on a *gendered entity extraction* procedure that consists of the entity extraction and gender classification steps described in Subsection 2.1.2 and Subsection 2.2.2. For each text, this procedure produces an initial set of gendered entities appearing in the text. These entities, among others, are used in the calculation of agency bias and appearance bias.

Entity Extraction. The entity extraction step is implemented using spaCy’s named entity recognizer.[5] We extract all `PERSON` entities that appear in the text at least three times.

[5] https://spacy.io/models/en#en_core_web_sm

Gender Classification. After the entity extraction step, we assign a gender label to each of the extracted entities using a method similar to the one used in Toro Isaza et al. (2023). Our gender classification procedure consists of two heuristic steps.

- **Honorific Heuristic:** Entities preceded by one of the *gendered honorific titles* in Table 4 are assumed to be "surnames" and assigned a gender label according to their honorific.
- **Coreference Heuristic:** We use Lee et al.'s (2018) coreference resolution model to identify third-person pronouns co-referring with each entity that was not assigned a gender via the honorific heuristic. We label an entity as "female" if the model identifies more instances of *she*/*her*/*herself* as co-referents than *he*/*him*/*himself*, and *vice versa*.

Our gender classification procedure attains an accuracy of 98.2% on a manually labeled validation set consisting from 10 novels from our overall dataset. Entities that could not be assigned a gender label are excluded from the agency bias and appearance bias calculations.

A.2. Agency Bias Calculation

Our implementation of the agency bias metric uses Shi and Lin's (2019) semantic role labeler (SRL)[6] to determine whether gendered entities are agents or patients. This SRL model is a tagging model, which takes a sentence as input and assigns semantic role labels to spans of tokens.

Gendered Argument Extraction. Entities with semantic roles are collectively known as *arguments*. Our agency bias calculation is based on a set of *gendered arguments* extracted as follows. A span of tokens is considered a gendered argument if it has been assigned the label `ARG0` (agent) or `ARG1` (patient) by the SRL model, and one of the following conditions holds.

(a) The span exactly matches one of the entities extracted and gender-classified during the gendered entity extraction step.

(b) The span exactly matches one of the "common gendered entities" appearing in Table 5.

(c) The last word of the span satisfies condition (a) or (b), and all words in the span satisfying condition (a) or (b) have the same gender label.

(d) The span contains at least one word satisfying condition (a) or (b), the last word of the span is a surname identified using the honorific heuristic during gender classification, and all words in the span satisfying condition (a) or (b) have the same gender label.

Agentivity Calculation. Female agentivity is calculated as follows:

$$\text{female agentivity} = \frac{\text{\# of female agents}}{\text{\# of female arguments}}$$

and male agentivity is calculated analogously.

A.3. Appearance Bias Calculation

In addition to the male and female words appearing in Table 1, our WEAT scores also include novel-specific named entities. Since WEAT is a measure of stereotypical associations between embedding vectors, the appearance bias calculation can only include individual vocabulary items, taken out of context. These vocabulary items include all single-token entities from the gendered entity extraction step that are assigned a consistent gender label throughout the novel. Certain two-token entities are also represented in the appearance bias score, according to the following rules.

- If the first token is an honorific title appearing in Table 4, then the second token is included if it never appears after an honorific of the opposite gender label.
- If the first token is not an honorific title appearing in Table 4, then the first token is included if the second item is *not* a surname that has appeared with both a male and a female honorific.

[6] https://docs.allennlp.org/models/main/models/structured_prediction/models/srl_bert/

Common Female Entities	abbess, aunt, bachelorette, baroness, bride, countess, dame, daughter, doe, druidess, duchess, empress, female, females, firewoman, girl, girlfriend, girls, goddaughter, godmother, grandmother, heiress, her, heroine, herself, ladies, lady, madam, mademoiselle, mailwoman, matriarch, miss, miss., mother, mothers, mrs, mrs., niece, nun, policewoman, princess, queen, saleswoman, she, sister, sorceress, stepmother, widow, wife, witch, wives, woman, women
Common Male Entities	abbot, bachelor, baron, boy, boyfriend, boys, brother, druid, duke, earl, emperor, father, fathers, fireman, friar, gentleman, godfather, godson, grandfather, groom, he, heir, him, himself, husband, husbands, king, knight, mailman, male, males, man, men, mister, monsieur, mr, mr., nephew, patriarch, policeman, prince, salesman, sir, son, sorcerer, stag, stepfather, uncle, widower, wizard

Table 5: “Common gendered entities” that are included in the calculation of agency bias, regardless of whether they were extracted during the gendered entity extraction step.