Commonsense or Stereotype: Investigating Gender Bias in COMET-ATOMIC$^{20}_{20}$

Xinyu Huo, Jiapei Yu, Christy Zhang, Sara Zhang
University of British Columboa

Abstract

This study investigates whether a state-of-the-art generative commonsense reasoning model encodes gender bias in its outputs. We focus on COMET-ATOMIC$^{20}_{20}$ and systematically analyze its generative behavior using a controlled dataset constructed by combining prompts adapted from the WinoBias benchmark with a curated list of male, female, and unisex names. Each prompt differs only in the gender of the subject, enabling isolated evaluation of gender effects. To evaluate the model's responses, we conduct sentiment analysis, lexical bias analysis, and agreement score analysis, and apply statistical testing to assess the significance of observed differences. Our findings uncover consistent and sometimes counterintuitive patterns of gender bias, raising critical concerns about the reliability and fairness of generative commonsense inference systems in downstream applications. [1]

1 Introduction

Language models have dramatically improved natural language processing tasks, but research consistently demonstrates that they often carry undesirable societal biases. Gender bias is one prominent concern, where models tend to produce stereotypical associations. For example, as shown in Figure 1, GPT-4o mini tends to link "she" more strongly with "nurse" and "he" with "doctor". While language models such as BART and GPT offer remarkable linguistic fluency, they may perpetuate unfair or offensive content if biases remain unchecked.

Moving beyond standard language models, generative commonsense models aim to encode everyday knowledge about cause-and-effect, motivations, and social norms. A leading example is COMET-ATOMIC$^{20}_{20}$, a generative model trained on the ATOMIC$^{20}_{20}$ knowledge graph (Hwang et al., 2021a). COMET-ATOMIC$^{20}_{20}$ takes an event like "PersonX sends PersonY a message" and generates plausible if-then inferences, such as PersonX's intent (xIntent) or the possible effect on PersonY (xEffect). Although such generative capabilities are useful for downstream applications, they also pose risks: if COMET-ATOMIC$^{20}_{20}$ encodes stereotypes, it could embed biased "commonsense" into large language models at a foundational level.

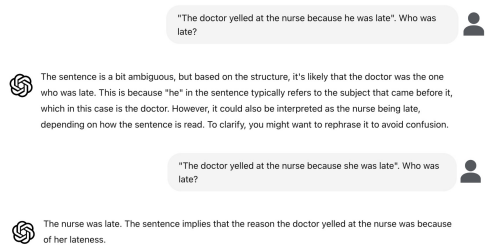

Figure 1: GPT-4o mini's response for a grammatically ambiguous sentence.

While biases in pretrained language models and static commonsense knowledge bases have been extensively studied, there is limited research specifically examining whether generative commonsense models themselves exhibit gender bias in the inferences they produce. Our project aims to address this gap by systematically analyzing COMET-ATOMIC$^{20}_{20}$'s outputs. In principle, commonsense inferences should remain consistent regardless of a person's identity or gender. If, however, substituting "John" for "Mary" or "PersonX" leads to systematically different outputs, it would suggest that the model encodes and amplifies gender stereotypes rather than producing neutral commonsense knowledge.

To investigate this, our work makes the following contributions:

1. We conduct the first systematic audit of gender bias in COMET-ATOMIC$^{20}_{20}$'s generated inferences, evaluating both the BART and GPT-2

[1] The code for this project can be found at https://github.com/christyyz/CPSC532-Project.

based models.

2. We design a controlled experiment where COMET-ATOMIC$^{20}_{20}$ is prompted with identical events while varying only the subject's name, using male, female, unisex, or neutral references.

3. We perform a comprehensive evaluation using sentiment analysis, output agreement scoring, and lexical bias analysis, with statistical significance testing applied to sentiment and agreement measures.

2 Related work

The foundational work in gender bias in natural language processing began with the study of word embeddings. Researchers showed that embeddings like Word2Vec associated male terms with technical roles and female terms with domestic roles (Bolukbasi et al., 2016). For example, "receptionist" was closely associated with female terms, while preserving desired associations like "queen" to "female". This early research not only highlighted the existence of bias but also introduced initial debiasing techniques, sparking broader conversations around fairness in NLP.

As models grew larger and more complex, the rise of large language models (LLMs) introduced new layers of complexity and risk. Despite architectural advancements, stereotypical associations persisted. To systematically benchmark these biases, Nadeem et al. (2020) introduced StereoSet, a large-scale dataset designed to evaluate both the stereotypical bias and language modeling ability of popular models such as BERT and GPT-2. Their findings revealed strong gender, race, professional, and religon stereotypes embedded in LLM outputs. Extending beyond benchmarking, Ferrara (2023) explored the ethical risks of deploying biased LLMs in real-world applications, emphasizing the importance of ongoing fairness research.

Beyond broad model-level stereotypes, recent studies have shown that even subtle cues, such as names, can significantly affect model behavior. Shwartz et al. (2020) found that model predictions could shift dramatically based purely on the name provided, creating unintended associations. Similarly, An and Rudinger (2023) demonstrated that demographic attributes of a name (race, ethnicity, and gender) and name tokenization length influence the behaviour of models including BERT, RoBERTa and GPT-2. These findings are particularly relevant to our research, where we systematically vary male, female, and unisex names to probe behavioral differences.

This observed sensitivity to names raises a natural question regarding how commonsense reasoning models handle gender.

Commonsense reasoning has typically been represented through static knowledge graphs such as ConceptNet (Speer et al., 2018) and ATOMIC (Sap et al., 2019). To improve the coverage and generative capabilities of commonsense reasoning, Bosselut et al. (2019) introduced COMET, a transformer model trained to generate commonsense inferences from natural language prompts. Building on this, Hwang et al. (2021a) released COMET-ATOMIC$^{20}_{20}$, trained on ATOMIC$^{20}_{20}$. It includes 51 inferential relation types, which significantly broadens the commonsense coverage compared to the earlier COMET model.

COMET has been leveraged in prior studies as a tool for bias detection. For example, Huang et al. (2021) used COMET to infer social implications from narrative texts, uncovering implicit gender bias in story generation systems. Mehrabi et al. (2021) investigated biases associated with COMET by analyzing its outputs. However, their evaluation only involved giving COMET target events drawn directly from its training set, such as ConceptNet, potentially limiting their ability to assess COMET's true generative inference behavior.

Our work extends the analysis of gender bias to COMET-ATOMIC$^{20}_{20}$, evaluating the model's generative behavior on unseen events involving gendered and unisex subject names. We design a controlled experimental setup leveraging benchmark datasets and curated name lists, which we describe in detail in the following section.

3 Experiments

This section outlines the construction of the dataset and the design of our experimental procedure. Each data instance comprises a base event, which consists of two components: a subject name and an associated action. For example, "Alex argued with the designer" is one base event sentence. To represent gendered subject names, we chose to use the 100 commonly used female names and 100 male names. (Social Security Administration, 2023). Unisex names were generated using GPT-4o by prompting for names commonly associated with

both genders. We also use the gender-neutral placeholder "PersonX", consistent with how entities are referred to in ATOMIC$^{20}_{20}$ for addtional comparative analysis.

The action components of the base events were drawn from WinoBias (Zhao et al., 2018), a benchmark dataset developed to evaluate gender bias in coreference resolution systems. WinoBias contains events that reflect occupational and gender-based stereotypes. To tailor this dataset for our purposes, we performed a series of preprocessing steps: we removed the original subjects, sub-events, and accompanying reasonings, retaining only the main event content. Duplicate events were also eliminated, resulting in a total of 400 unique base events.

To support commonsense inference, we employed the 51 predefined relation types defined by the COMET-ATOMIC$^{20}_{20}$ framework (Hwang et al., 2021b). These relation types encode various forms of commonsense knowledge, such as probable causes, effects, intentions, and attributes linked to events.

To construct the final dataset, we randomly assigned a subject name from our curated name lists to each base event and paired it with all 51 relation types. This process generated 20,400 event–relation pairs per gender group (female and male), yielding a total of 40,800 data points. In addition, we created a comparable set using unisex names and the placeholder “PersonX” to facilitate gender-neutral comparisons.

For inference generation, we employed two generative language models: the COMET-ATOMIC$^{20}_{20}$ models, based on BART and GPT-2XL. In both cases, we used pretrained weights and adapted code from the official COMET-ATOMIC$^{20}_{20}$ GitHub repository (Hwang et al., 2021b). Each model generated a single output inference for every event–relation pair, enabling comparative analysis of model behavior across name groups.

4 Results

To evaluate the outputs generated by the BART and GPT2-XL based COMET-ATOMIC$^{20}_{20}$ models, we conducted sentiment analysis, agreement score evaluation, and lexical bias analysis on outputs generated with female and male names. For reference, we also included outputs generated with unisex names and the placeholder PersonX.

4.1 Sentiment Analysis

In the sentiment analysis experiment, we extracted positive, neutral, and negative sentiment scores for each generated output. These scores represent the model’s confidence in assigning each sentiment category to the text.

Table 1a presents the average sentiment scores for outputs generated by the BART model. Outputs associated with male names exhibit slightly higher positive and negative sentiment scores, whereas those with female names tend to receive more neutral sentiment scores on average. Table 1b reports the results of statistical tests comparing outputs with female and male names. All p-values are below 0.05, indicating that the differences in sentiment scores are statistically significant across all categories.

Table 2a presents the results from the GPT2XL model. Similar to BART, outputs generated with female names received higher neutral sentiment scores compared to those with male names. However, in this case, the negative sentiment scores are identical for both male and female names. While the absolute difference in positive sentiment scores is not large, it is more pronounced than in the BART results, suggesting that GPT2XL assigns a stronger positive tone to outputs associated with male names. The t-tests reported in Table 2b show that the differences in positive and neutral sentiment scores are statistically significant.

4.2 Agreement Score Analysis

An agreement score analysis is conducted to evaluate how similarly the models generate outputs when prompted with male or female names, relative to a neutral reference. Using cosine similarity, we compared each output to its corresponding neutral version and performed paired t-tests to assess differences in similarity scores. As shown in Table 3, for both models and across both neutral references (PersonX and a unisex name), outputs generated with male names are consistently more similar to the neutral outputs than those generated with female names. All differences are statistically significant, highlighting a consistent gender-based variation in model behavior.

4.3 Lexical Bias Analysis

To investigate potential lexical bias in the models’ outputs, we applied the LIWC-22 analysis tool on each inference generated with female and male

	Positive	Neutral	Negative
Female Name	**0.123**	**0.811**	**0.061**
Male Name	**0.127**	**0.802**	**0.067**
Unisex Name	0.118	0.811	0.068
PersonX	0.096	0.841	0.063

(a) Average Sentiment Scores of Outputs Generated by BART, Grouped by Gendered Names

	t-value	p-value
Positive	-3.147	0.002
Neutral	5.688	0.000
Negative	-5.106	0.000

(b) Statistical Comparison of Sentiment Scores Between Female and Male Names

Table 1: Sentiment Analysis Results and Statistical Comparisons for BART

	Positive	Neutral	Negative
Female Name	**0.131**	**0.766**	**0.102**
Male Name	**0.152**	**0.745**	**0.102**
Unisex Name	0.139	0.753	0.107
PersonX	0.180	0.741	0.079

(a) Average Sentiment Scores of Outputs Generated by GPT2XL, Grouped by Gendered Names

	t-value	p-value
Positive	-10.768	0.000
Neutral	8.587	0.000
Negative	0.103	0.918

(b) Statistical Comparison of Sentiment Scores Between Outputs with Female and Male Names

Table 2: Sentiment Analysis Results and Statistical Comparisons for GPT2XL

	t-value	p-value
Bart PersonX	-3.766	0.000166
Bart Unisex	-6.362	2e-10
GPT2 PersonX	-7.229	5e-13
GPT2 Unisex	-2.438	0.0148

Table 3: Agreement Scores Between Gendered and Neutral Outputs Across Models

names by the BART and GPT2-XL models. The LIWC-22 tool provides scores across a broad range of psychologically and linguistically relevant categories. We conducted independent t-tests to determine whether the differences were statistically significant in each of the categories. Table 4 presents the categories in both models that are statistically significant and can infer bias, where the description and most frequent examples are obtained from LIWC-22 user manual (Boyd et al., 2022).

Table 4a shows that family, friend, swear, and emo_anger categories present meaningful variation in their usage across gendered outputs. The results reveals that lexical choices in BART-generated text reflect traditional gender stereotypes. Words related to family, such as mother and baby are significantly more frequent in female-associated outputs, whereas terms like friend and dude are more common in male-associated outputs. This pattern suggests an implicit alignment between women and caregiving or familial roles, and men with more casual or socially oriented relationships.

Contrary to common stereotypes, the data shows that swear words are significantly more frequent in female-associated outputs. On the other hand, words associated with anger, such as mad, angry, and hate, appear more frequently in male-associated outputs. This finding suggests that while male outputs may be more associated with emotionally aggressive expressions, female outputs surprisingly contain more profanity.

Table 4b presents the LIWC categories that reveal gender-related lexical bias in GPT2XL-generated outputs. In contrast to BART, GPT2-XL output shows that anger-related words appear more frequently in female-associated outputs, challenging traditional stereotypes. While the lack category indicates that male outputs more often include words signaling deficiency or unmet needs, other categories, such as power, fulfill, reward, and prosocial, highlight themes of agency, achievement, success, and social support more frequently appeared in male-associated outputs.

Moreover, the greater frequency of moral and polite language in female-associated outputs suggests that GPT portrays women as more likely to use polite or morally evaluative language. This reflects social expectations around gendered communication styles. Finally, the money and home categories again reflect traditional gender stereo-

types, with female outputs emphasizing domestic environments and male outputs more frequently referencing economics and the public sphere.

5 Conclusions

This study offers compelling evidence that generative commonsense reasoning models, such as COMET-ATOMIC, are not neutral arbiters of social knowledge but instead exhibit systematic gender bias in their inferences. Through carefully controlled experiments that isolate gender as the only variable, by substituting male, female, and unisex names in otherwise identical event prompts, we uncover consistent differences in the outputs generated by both the BART and GPT-2XL variants of COMET-ATOMIC$^{20}_{20}$. Our results reveal that these differences are not merely stylistic or random, but statistically significant across multiple dimensions, including sentiment, lexical choices, and inference similarity to gender-neutral references.

Importantly, these biases are not uniformly aligned with traditional stereotypes, which reveals a deeper complexity. For instance, while some outputs reinforce familiar tropes such as associating women with family and caregiving, others challenge expectations, as seen in the unexpectedly higher frequency of swear words in female-associated outputs. This duality highlights that the biases encoded in language models may not be simple reflections of societal norms but rather convoluted artifacts shaped by training data distributions and model architectures.

What distinguishes our work is its direct interrogation of generative behavior, rather than representational or static associations. Prior studies have either used models like COMET as tools for analyzing bias in external corpora or assessed bias in fixed knowledge graphs. Our contribution lies in showing that even when commonsense reasoning is generated on the fly, it can reproduce, and sometimes amplify gendered patterns. In doing so, we bridge the gap between token-level bias studies and the broader, more impactful realm of reasoning-level bias in generative AI systems.

Ultimately, our findings raise important questions about the trustworthiness of commonsense inference models in downstream applications. If these models encode subtle yet systematic gender differences in how they interpret human actions, intentions, and emotions, then their use in safety-critical domains such as education, hiring, or healthcare demands greater scrutiny. Addressing these biases is not just a matter of technical correction, it is a prerequisite for building AI systems that reason fairly about the world.

6 Future Work

While our current study focuses on binary gender representation and explores potential bias in commonsense inference generation, several directions remain for future exploration.

First, a more granular analysis of relation types could yield deeper insights. For example, relation categories such as xEffect and oEffect, which refer to effects on the subject versus the object were not distinguished by the gender of the action recipient in our current setup. Future work could incorporate this distinction, as well as evaluate relation groups thematically, such as physical actions versus social interactions, to identify patterns of bias across different contexts.

Second, improvements in the selection of unisex names would enhance the robustness of comparison across gender categories. Rather than relying on external lists and placeholders, future work could algorithmically identify names that are frequently used across both male and female populations, using data-driven overlap to construct a more representative unisex name set.

Finally, we acknowledge that our study simplifies gender into a binary classification for the purpose of controlled experimentation. This approach does not encompass the full spectrum of gender identities. Future research should aim to include a broader and more inclusive range of gender representations to better reflect real-world diversity and mitigate limitations introduced by binary assumptions.

7 Self-evaluation of the project

Overall, we were largely able to follow the scope outlined in our original project proposal. Although the process of implementing the code and generating inferences from both models BART-based and GPT-2XL-based COMET-ATOMIC$^{20}_{20}$ models took longer than anticipated, we successfully completed inference for both, which exceeded our initial expectations and significantly strengthened the comparative aspect of our study.

One of our proposed extensions was to use a large language model to generate a list of names commonly used across both genders. While we

Category	Description / Most Frequent Examples	T-Stat	P-Value
family	parent*, mother*, father*, baby	2.45	0.0142
friend	friend*, boyfriend*, girlfriend*, dude	-2.44	0.0146
swear	shit, fuckin*, fuck, damn	3.23	0.0012
emo_anger	hate, mad, angry, frustr*	-2.38	0.0175

(a) LIWC categories revealing gender-related lexical bias in BART-generated text.

Category	Description / Most Frequent Examples	T-Stat	P-Value
emo_anger	hate, mad, angry, frustr*	2.96	0.0031
lack	don't have, didn't have, *less, hungry	-2.06	0.0397
power	own, order, allow, power	-3.12	0.0018
fulfill	enough, full, complete, extra	-6.44	1.17e-10
reward	opportun*, win, gain*, benefit*	-5.19	2.16e-07
prosocial	care, help, thank, please	-3.24	0.0012
moral	wrong, honor*, deserv*, judge	2.36	0.0180
polite	thank, please, thanks, good morning	2.24	0.0251
money	business*, pay*, price*, market*	-3.63	0.0003
home	home, house, room, bed	3.61	0.0003

(b) LIWC categories revealing gender-related lexical bias in GPT2-XL-generated text.

Table 4: Comparison of statistically significant LIWC categories in text generated by BART and GPT2-XL with female versus male names.

Note: Words marked with an asterisk (e.g., friend*, price*, *less) indicate wildcard stems. A trailing asterisk matches all words beginning with the stem (e.g., friend, friends, friendship), while a leading asterisk matches words ending with that stem (e.g., hopeless, careless).

implemented this approach, we found that verifying the quality and representativeness of the generated names was challenging. This remains a promising direction for future refinement and evaluation.

Due to time constraints, we were unable to explore one of the planned components of our proposal: incorporating bias detection techniques from prior work, such as the SODAPOP (An et al., 2023) framework for identifying social biases in commonsense knowledge models. Exploring such techniques may enhance our evaluation and offered a more comprehensive understanding of gender bias in model-generated inferences.

References

Haozhe An, Zongxia Li, Jieyu Zhao, and Rachel Rudinger. 2023. Sodapop: Open-ended discovery of social biases in social commonsense reasoning models. *Preprint*, arXiv:2210.07269.

Haozhe An and Rachel Rudinger. 2023. Nichelle and nancy: The influence of demographic attributes and tokenization length on first name biases. *Preprint*, arXiv:2305.16577.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Preprint*, arXiv:1607.06520.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *Preprint*, arXiv:1906.05317.

Ryan L. Boyd, Kayla M. Jordan, and James W. Pennebaker. 2022. Liwc-22 manual: Development and psychometrics.

Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *First Monday*.

Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. 2021. Uncovering implicit gender bias in narratives through commonsense inference. *Preprint*, arXiv:2109.06437.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021a. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *Preprint*, arXiv:2010.05953.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021b. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *Preprint*, arXiv:2010.05953.

Ninareh Mehrabi, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021. Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources. *Preprint*, arXiv:2103.11320.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *Preprint*, arXiv:2004.09456.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. *Preprint*, arXiv:1811.00146.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online. Association for Computational Linguistics.

Social Security Administration. 2023. Top names of the period 1923–2022. `https://www.ssa.gov/oact/babynames/decades/century.html`. [Accessed: Mar. 7, 2025].

Robyn Speer, Joshua Chin, and Catherine Havasi. 2018. Conceptnet 5.5: An open multilingual graph of general knowledge. *Preprint*, arXiv:1612.03975.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *Preprint*, arXiv:1804.06876.