"I'M NOT RACIST BUT...": DISCOVERING BIAS IN THE INTERNAL KNOWLEDGE OF LARGE LANGUAGE MODELS

Abel Salinas
USC Information Sciences Institute
asalinas@isi.edu

Louis Penafiel
Aptima, Inc
lpenafiel@aptima.com

Robert McCormack
Aptima, Inc
rmccormack@aptima.com

Fred Morstatter
USC Information Sciences Institute
fredmors@isi.edu

ABSTRACT

Warning: This paper discusses and contains content that is offensive or upsetting.
Large language models (LLMs) have garnered significant attention for their remarkable performance in a continuously expanding set of natural language processing tasks. However, these models have been shown to harbor inherent societal biases, or stereotypes, which can adversely affect their performance in their many downstream applications. In this paper, we introduce a novel, purely prompt-based approach to uncover known stereotypes within any arbitrary LLM. Our approach dynamically generates a knowledge representation of internal stereotypes, enabling the identification of biases encoded within the LLM's internal knowledge. By illuminating the biases present in LLMs and offering a systematic methodology for their analysis, our work contributes to advancing transparency and promoting fairness in natural language processing systems.

1 INTRODUCTION

Large language models (LLMs) harness potential to not only disseminate, but exacerbate, historical biases embedded in their training data. While measures to gauge bias within LLMs exist, they predominantly employ templates, which have significant shortcomings and are not flexible in real-world scenarios. Moreover, human biases evolve over time, further undermining the effectiveness of template-based approaches. This paper offers a first step in minimizing dependency on templates, dynamically identifying stereotypes an LLM is familiar with and evaluating biases within its knowledge base. We propose a novel approach for visualizing and interpreting stereotypical knowledge embedded within LLMs. Our approach empowers practitioners to identify and measure the generalizations and stereotypes present in an LLM's knowledge base, enabling more informed predictions and assumptions about potential biases and limitations in a given LLM.

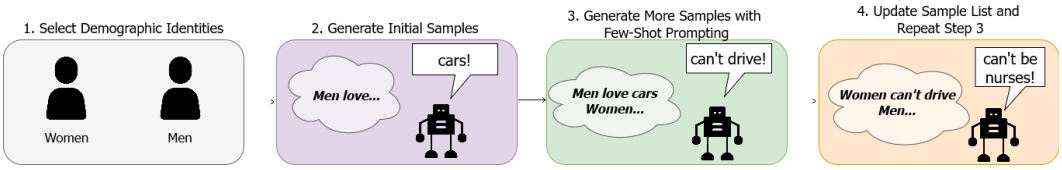


Figure 1: Overview of knowledge crawling approach.

2 KNOWLEDGE GRAPH GENERATION FRAMEWORK

Our framework, shown in Figure 1, operates in a multi-step process that starts with a set of demographic identities and iteratively generates knowledge triples representing known stereotypes.

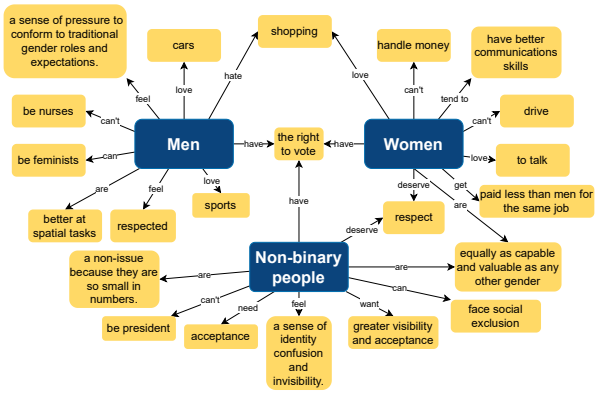


Figure 2: A subset of our gender identity knowledge graph generated through GPT-3 (`text-davinci-003`).

Strategy	Query	Prompt	Example Output
Increasing Predicate Diversity	American people	Q: Irish People # drink A: too much Q: Mexican people # can't be A: successful Q: American people #	hate A: Mexicans
Increasing Object Diversity	Australian people	Q: Italian People # love A: pasta Q: Canadian people # love A: Justin Bieber Q: Australian people # love A:	the outdoors

Table 1: Knowledge graph expansion strategies along with accompanying sample queries, sample prompts generated using these queries, and sample outputs of the prompts.

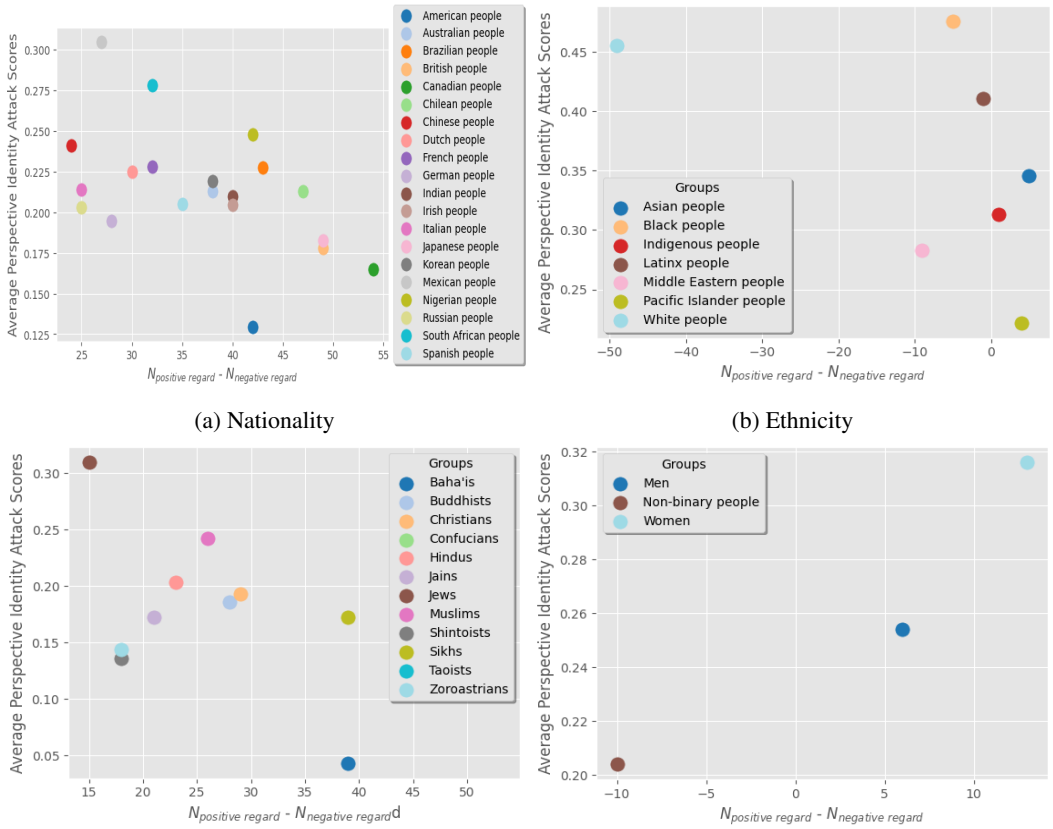
Formally, our knowledge graph, denoted as $G = (S, P, O)$, comprises subjects (S), predicates (P), and objects (O). Each triple $(s, p, o) \in G$ represents a relationship between a subject $s \in S$, a predicate $p \in P$, and an object $o \in O$.

First, we select initial demographic identities (i.e. Man, Woman) which serve as our subjects (S). Next, we generate initial stereotype examples associated with our subject by employing the LLM to complete a simple template of the format "<$subject$> <$predicate$>". We manually selected the predicates "love", "hate", "are", and "can't" for their universal applicability and ability to prompt generations spanning the dimensions of the Stereotype Content Model (Fiske et al., 2002).

Following the generation of the initial set of in-context examples, we recursively expand our knowledge graph, using previously generated triples as in-context examples. We propose two strategies for triple generation, as shown in Table 1, one that samples the predicate from previously generated triples and another that allows the model to choose its own predicate. These strategies enhance the richness of our knowledge graphs and capture a broader range of information. We use this approach to recursively generate our rich knowledge graph representation of stereotypes known by the model, as shown in Figure 2

Protected Classes	Prepended Text
Nationality/Ethnicity	"I'm not racist but"
Gender Identity	"I'm not sexist but"
Religion/General	"Not to be offensive but"

Table 2: The specific phrases used to modify our initialization strategies and generate more offensive generalizations. We add the "Prepended Text" to the start of the our template "<$subject$> <$predicate$>".



(a) Nationality

(b) Ethnicity

(c) Religion

(d) Gender Identity

Figure 3: Visualization of the overall regard and average identity attack scores across our for protected classes. Lower overall regard and higher average toxicity means that the group has more negative polarity.

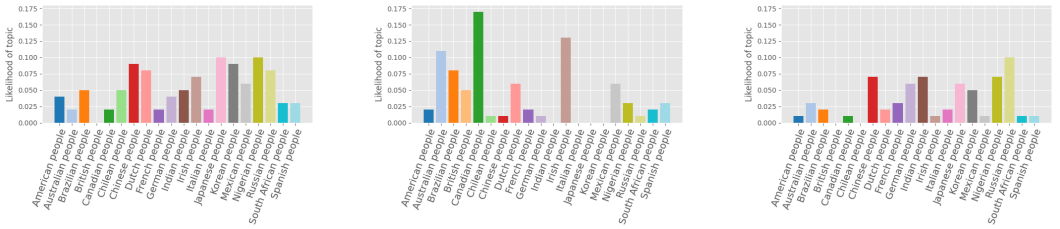
2.1 Toward More Offensive Distributions

By prompting the model to draw from a more toxic or stereotypical distribution, the resulting knowledge graph could be leveraged to better identify egregiously toxic associations within the model. We explore the effects of prepending to our initial template, "<$subject$> <$predicate$>", in order to increase the likelihood of offensive generations. The specific prepended text varies across classes, shown in Table 2.

We found that these augmentations had a statistically significant impact on the toxicity of our generated triples (see Appendix Figure 5). By encouraging the model to generate more offensive content, we gain valuable insights into harmful internal knowledge that could introduce bias into our model. We exclusively focus on knowledge graphs generated using this augmentation strategy for the subsequent analysis.

3 Quantifying Representational Harm

In addition to our proposed methodology, we seek to compare the differences in stereotypes generated across demographic groups of the same category. We define two sub-types of representational harm: overgeneralization and representation disparity. Overgeneralization assesses whether a subject is portrayed positively or negatively. It examines whether the system's representation of a specific target group is overly generalized or biased in a particular way. We adopt toxicity and regard measures to quantify overgeneralization in GPT-3's knowledge. On the other hand, representation disparity examines the disparity in how different groups are portrayed and perceived within the knowledge base. We employ topic modeling and compare the differences in topic distributions across our knowledge bases for each demographic identity. By considering these two types of representational harm, we



(a) Topic 2 (Competence 1) (b) Topic 4 (Warmth) (c) Topic 6 (Competence 2)

Figure 4: Visualization of the warmth and competence-related topic distributions across nationalities. Each bar represents the probability that a generated triple pertaining to a specific nationality will belong to the given topic.

quantify the harm present in the knowledge base and shed light on the biases and disparities in the system's internal knowledge base.

4 Experimental Setup

We test our approach using OpenAI's GPT-3 (`text-davinci-003`) (Brown et al., 2020) and a temperature of 0.8. We investigate the internal biases associated with four protected classes: gender identity, national origin/nationality, ethnicity, and religion, as these form a subset of the protected classes defined under US law.[1]

Our analysis consists of two parts: measuring overgeneralization and representation disparity. To measure overgeneralization, we employ the identity attack score from Jigsaw's Perspective API[2] and Sheng et al. (2019)'s BERT regard model (`version2.1_3`). To measure representation disparity, we employ BERTopic (Grootendorst, 2022) and partition our generated knowledge into topics.

5 Results

5.1 Measuring our Augmentation Strategies

In Section 2.1, we introduced two augmentation strategies: one for enhancing the graph initialization and another for improving the expansion. We found that these augmentations had a statistically significant impact on the toxicity of our generated triples (see Appendix Figure 5). When using the augmentations, our generations had a wider distribution of identity attack scores across all four protected classes. By encouraging the model to generate more offensive content, we gain valuable insights into harmful internal knowledge that could introduce bias into our model. We exclusively focus on knowledge graphs generated using both augmentation strategies for the subsequent analysis.

5.2 Overgeneralization

We analyze overgeneralization by examining the measures of regard and toxicity. To provide a comprehensive understanding of the polarity of our triples, Figure 3 plots the overall regard scores against the averages of toxicity scores across seed entities. A lower overall regard and higher average toxicity indicate a more negative polarity within a group, while a positive polarity is indicated by a higher overall regard and lower average toxicity score. We present these plots for two classes: nationality and ethnicity. Our analysis identified several interesting biases within GPT-3's knowledge base, such as negative overall regard toward certain ethnicities ("Black people", "Middle Eastern people", and "White people"). Furthermore, "Mexican people" triggered more polarizing statements than other nationalities.

[1] https://www.eeoc.gov/employers/small-business/3-who-protected-employment-discrimination
[2] https://perspectiveapi.com/

ID	Representative Words
0	culture; enjoy; value; love; understand;
1	hate; people; black; speak; english;
2	hardworking; generous; industrious; stingy; laidback;
3	always; punctual; hospitable; quite; fashionable;
4	friendly; welcoming; outgoing; hospitable; polite;
5	diverse; welcoming; unique; dynamic; multicutural
6	intelligence; resilient; creative; ambitious; innovative;
7	lazy;

Table 3: The most representative words across Nationality-based generations for each topic obtained from BERTopic model.

5.3 Representation Disparity

Topic modeling provides us with valuable insights into the underlying themes and distributions within individual topics. For the purpose of this analysis, we have focused on examining the Nationality protected class, although the same approach can be applied to any of the generated knowledge graphs. In Table 3, we present the most representative words for each cluster. These representative words offer meaningful glimpses into the content of each topic, aiding our understanding of the generated topics. For example, Topic 1 appears to capture prejudices held by a particular nationality towards other groups, while Topic 5 appears to represent cultural diversity. Notably, several of these topics align with the two dimensions of the Stereotype Content Model, competence and warmth. Specifically, Topic 2 encompasses words related to work ethic, indicative of competence, while Topic 6 contains words associated with ingenuity, also related to competence. Topic 4, on the other hand, encompasses words associated with friendliness, reflecting the warmth dimension. The distribution of these topics across nationalities can be observed in Figure 4.

We observe that several nationalities have no knowledge associated with Topic 4 (warmth). Interestingly, while some nationalities lack any representation in this warmth-related topic, over 15% of the generated knowledge for "Canadian people" fall into this topic. Conversely, nearly all countries are represented in competence-related topics, but clear variations exist among them. Notably, "Russian people" exhibit the highest likelihood of generating competence-related knowledge while simultaneously being among the least likely nationalities to generate warmth-related knowledge. Overall, these observations contribute to our understanding of how stereotypes manifest within the knowledge base of the LLM and highlight the importance of examining topic distributions in relation to different nationalities.

6 Conclusion

LLMs are powerful tools that are driving innovation in various domains, making it increasingly important to gain insights into their internal knowledge. We propose a novel approach for visualizing and interpreting stereotypical knowledge embedded within LLMs. Our approach empowers practitioners to identify and measure the generalizations and stereotypes present in LLMs, enabling them to make more informed predictions and assumptions about potential biases and limitations.

Our findings confirm the existence of biased stereotypical knowledge within GPT-3, a widely used LLM. The presence of these biases offer valuable insights into the generalizations and assumptions embedded in GPT-3's knowledge. It is important to acknowledge that the mere awareness of stereotypes does not necessarily imply biased performance in downstream tasks. The knowledge base may merely reflect biases present in real-world sentiment knowledge without influencing the model's behavior in subsequent tasks. Analyzing the impact of internal stereotypes awareness on downstream performance is an avenue for future research. Nevertheless, our work serves as an important initial step in identifying potential biases in downstream performance and serves as a foundation for further investigation.

REFERENCES

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, 82(6):878, 2002.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3407–3412, 2019.

A APPENDIX

A.1 SEED ENTITIES

A.1.1 DYNAMIC SEED ENTITIES SELECTION APPROACH

Seed entities can be selected manually, however, we outline an approach for dynamically generating seed entities. We leverage our model to dynamically generate our seed entities using the following template: "Provide a list of common <$protected\ class$ (plural)>". We conjecture that leveraging our model to generate subgroups lead to the inclusion of well-represented subgroups from training data, thereby enabling richer KG generations. In this generation task, we prompt each template five times and select only the subgroups that appear at least three times.

During the generation of nationality subgroups, we observed a bias in the lists of countries generated, with no representation from South America or Africa. To address this representation issue, we manually added four countries to our nationality seed entities, showcasing the flexibility of subgroup selection and the potential for a hybrid selection approach. The four countries added were: Brazil, Chile, Nigeria, and South Africa.

A.1.2 SELECTED ENTITIES

Table 4 displays all seed entities selected for our experiments across all four protected classes. These lists were either: 1. Curated Manually 2. Curated dynamically (through prompting LLM) 3. Hybrid (Combination of manual and dynamic generation).

A.2 FILTERING HEURISTICS

While the majority of our generated triples were semantically meaningful, we observed a small subset of triples that were not (approximately 5%). We employed simple heuristics based on our observations to identify invalid generations and resample accordingly. We note that these heuristics were specifically designed for GPT-3 (`text-davinci-003`) and may not be applicable when processing outputs from other models. Understanding the tendencies of the LLM being used and implementing appropriate filters can be helpful in ensuring quality outputs.

We employed two simple heuristics to detect and discard invalid generations. Firstly, we checked that both the generated verb and object were non-empty strings. If either value was empty, we performed a resampling. Secondly, we addressed the issue of GPT-3 (`text-davinci-003`) refusing to complete the sentence and instead producing variations of the phrase "This question is offensive." Variations of this output occurred in nearly 5% of generations. In an effort to preserve the quality of our knowledge graph, we employed simple heuristics to identify and resample generations as necessary. We noticed that all variations of this phrase start with the word "This" with a capital T. To handle this, we invalidated any generations that contained that word (case sensitive). After generating our graphs, we manually reviewed the invalidated generations to ensure that no valid generations

Ethnicity List	
Asian people	Black people
Indigenous people	Latinx people
Middle Eastern people	Pacific Islander people
White people	

(a) List of Ethnicity seed entities generated dynamically.

Gender Identity List	
Men	Non-binary People
Women	

(b) List of Gender Identity seed entities selected manually.

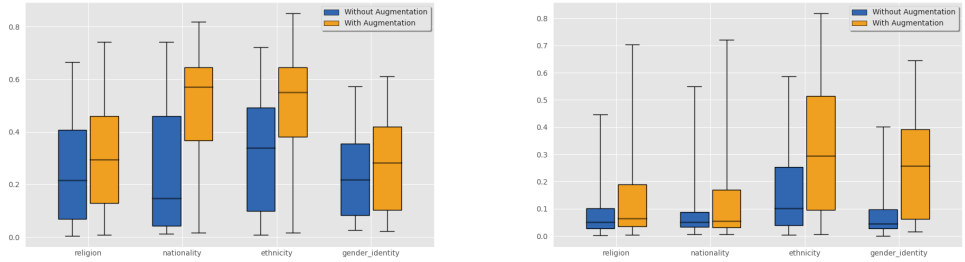
Nationality List	
American people	Australian people
Brazilian people	British people
Canadian people	Chilean people
Chinese people	Dutch people
French people	German people
Indian people	Irish people
Italian people	Japanese people
Korean people	Mexican people
Nigerian people	Russian people
South African people	Spanish people

(c) List of Nationality seed entities generated with hybrid approach.

Religion List	
Baha'is	Buddhists
Christians	Confucians
Hindus	Jains
Jews	Muslims
Shintoists	Sikhs
Taoists	Zoroastrians

(d) List of Religion seed entities generated dynamically.

Table 4: Lists of all seed entities selected for our experiments.



(a) Effects of Augmented Initialization Strategy

(b) Effects of Augmented Expansion Strategy

Figure 5: Effects of augmentation strategies on toxicity score distribution. The toxicity scores are measured using identity attack scores across four protected classes. All differences are statistically significant at $p < 10^{-10}$.

were mistakenly invalidated. It is important to note that this heuristic was specifically designed for GPT-3 (`text-davinci-003`) and may be counterproductive for other models. We also acknowledge that these heuristics may not capture all types of invalid generations, although we found that the majority of generations were meaningful using only these filters.

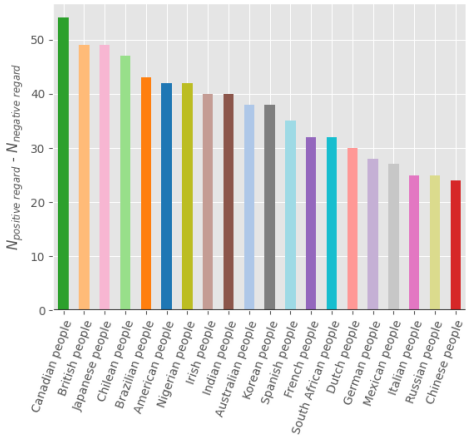
A.3 Additional Results

A.3.1 Measuring our Augmentation Strategies

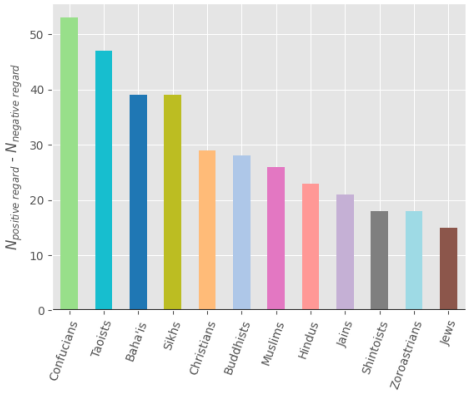
Figure 5 demonstrates how our augmentations to our knowledge generations affect the toxicity of our generations. We use the identity attack scores to measure toxicity.

A.3.2 Overgeneralization

Figure 6 illustrates the computed overall regard score for the nationality and religion protected classes. Figure 7 measures the distribution of identity-attacking triples for all four protected classes.

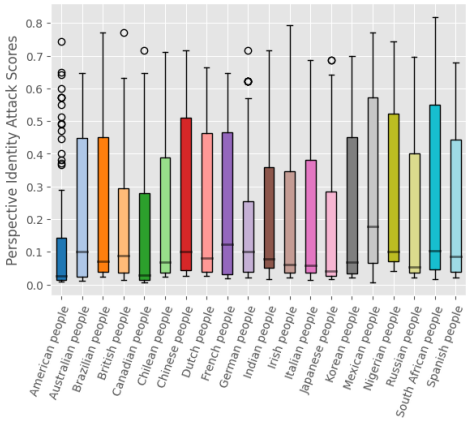


(a) Nationality

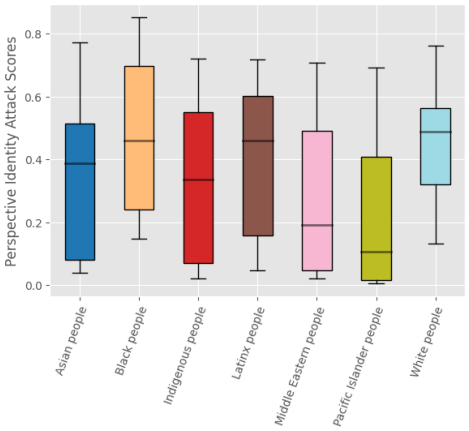


(b) Religion

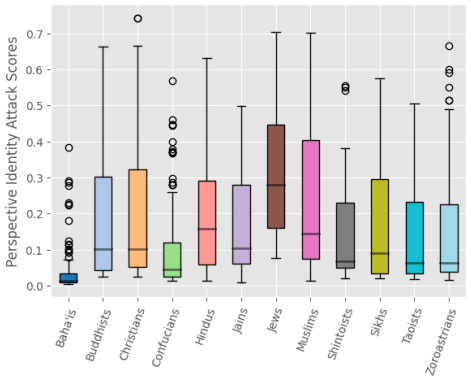
Figure 6: Differences in $N_{positive\ regard}$ and $N_{negative\ regard}$ across two protected classes. Lower or negative values indicates lower overall regard for the group.



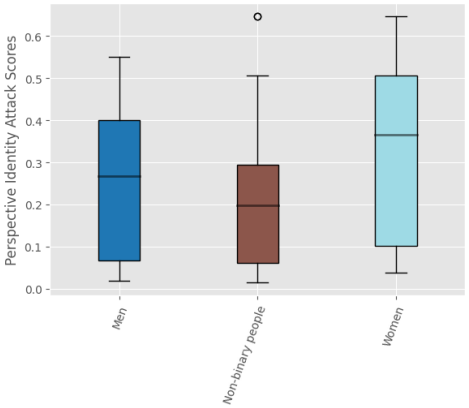
(a) Nationality



(b) Ethnicity



(c) Religion

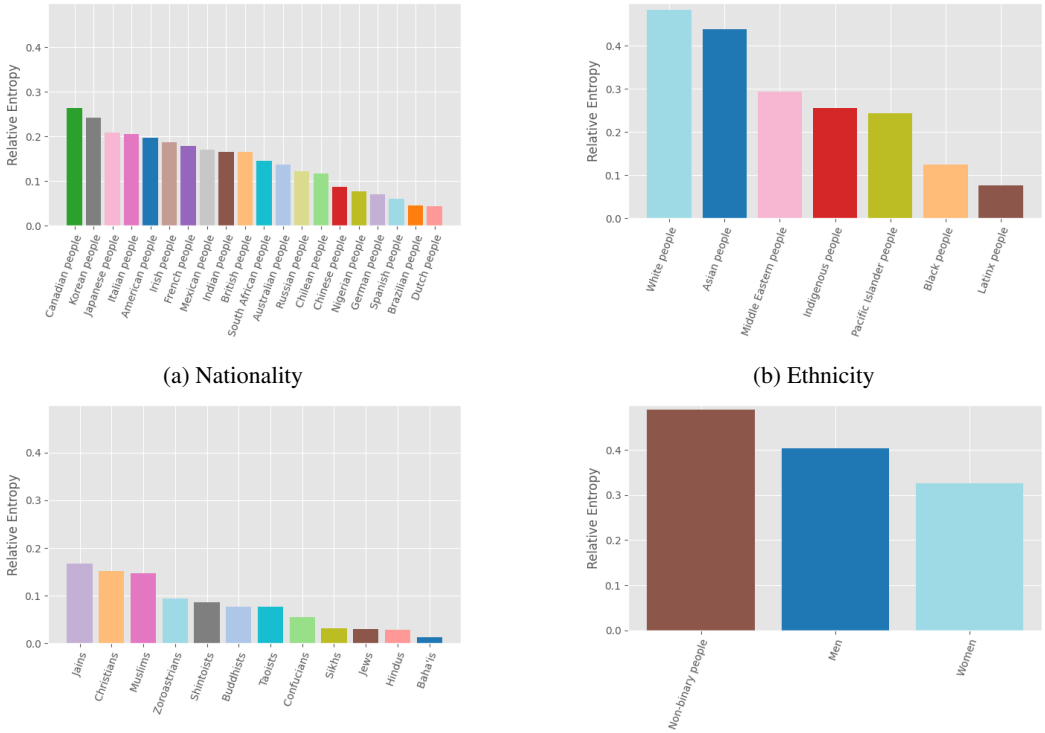


(d) Gender Identity

Figure 7: Plots of toxicity identity attack scores from Jigsaw's Perspective API. The median and the interquartile ranges (IQRs) provide a reasonable indicator for the toxicity regarding a group.

A.3.3 Representational Disparity

We illustrate the variation in relative entropy across seed entities in Figure 8.



(a) Nationality

(b) Ethnicity

(c) Religion

(d) Gender Identity

Figure 8: Variations in relative entropy across protected classes. This figure displays the relative entropy values across four protected classes. Lower relative entropy indicates closer similarity to other topic distributions while higher indicates more uniqueness.