Exploring Toxic and Hate Bias in Large Language Models (LLMs)

Written by
Agop Artyunyan[1], Álvaro Esteban Muñoz[2], Danae Philioti[3]
University of Ruse[1], University of Bologna[2], University of Cyprus[3]

Abstract

With the advent of generative AI, numerous applications have adopted strategies based on Large Language Models (LLMs) to tackle various NLP tasks. Many inside the scientific community highlights the inherent toxic and hate biases present in these LLMs, potentially leading to various social consequences and affect many persons in different ways. The amount of studies addressing this issue is increasing, however many do not agree with sociologists in robust standards or methodologies. Moreover, as more companies recognize the potential of LLMs-based applications, there's a risk that they may overlook these issues. In the following project we try to gather and expose the main techniques, methodologies, standards and datasets used in order to address the exploration of bias in LLMs.

Introduction

In recent years, we've seen the rise of large language models (LLMs) as powerful tools in natural language processing tasks. These models are impressive, generating text that's remarkably similar to human writing and finding applications across various domains. However, as we explore their potential, concerns have been raised about bias and toxicity within these models. These issues present significant challenges when it comes to their ethical and responsible use.

Understanding and addressing bias and toxicity in LLMs is crucial because it affects our societal values, fairness, and inclusivity. Bias refers to the presence of systematic errors or distortions in data or algorithms that result in unequal treatment or representation of certain groups or individuals (Ferrara 2023). Toxicity, on the other hand, encompasses harmful or offensive language that perpetuates stereotypes, incites violence, or promotes discrimination (Wen et al. 2023).

Since LLMs are trained on massive amounts of text data collected from the internet, they inevitably inherit biases present in that data. These biases reflect the prejudices and injustices found in society. What's more, the sheer scale and complexity of LLMs make it even more challenging to identify and effectively mitigate bias and toxicity.

Given the urgency of the situation, there is a pressing need for comprehensive research and methodologies to assess, quantify, and mitigate bias and toxicity in LLMs. Various techniques have been proposed to tackle these issues, including debiasing algorithms and adversarial training. However, achieving robust and equitable LLMs remains an ongoing endeavor.

In this technical report, we dive deep into the complex landscape of bias and toxicity in LLMs. Our aim is to shed light on the underlying mechanisms, explore existing methodologies. By fostering a deeper understanding of the complexities involved, our goal is to promote the development of ethically sound and socially responsible LLMs that uphold principles of fairness and inclusivity in natural language processing tasks.

Objectives

Our work is focused on the burgeoning field of ethics in Artificial Intelligence (AI), with a specific focus on addressing issues of **fairness** and social **bias** (Gallegos et al. 2024). Our inquiry delves into the intricate landscape of bias assessment and detoxification within Large Language Models (LLMs). By synthesizing existing research, we endeavor to offer a comprehensive overview of the current state-of-the-art methodologies and practices employed in the identification and mitigation of biases inherent in these advanced AI systems.

Our primary objective is to uncover correspondences and distinctive features among the techniques and resources used in various studies addressing bias evaluation and detoxification in LLMs. Through meticulous analysis and comparison, we seek to illustrate both the strengths and weaknesses inherent in these approaches. By doing so, we aim to provide valuable insights that can inform future research endeavors and contribute to the ongoing dialogue surrounding ethical considerations in AI development and deployment.

In conclusion, we anticipate that our literature review will serve as a valuable entry point for researchers and practitioners seeking to engage with the complex terrain of fairness and bias in AI with greater ease and clarity. In the **Methodologies** section, we will present the most prevalent methodologies for bias evaluation and the strategies used for detoxifying large language models (LLMs). Following this, the **Datasets** section will review the most common datasets and benchmarks utilized for bias evaluation, while the **Metrics** section will outline the metrics used to measure these biases. Finally, we will synthesize the insights from the reviewed literature in the **Findings** section and discuss potential future challenges in the **Open challenges** section. An additional

section covering a small **Case Study** applying one of the evaluation methodologies in the **Methodologies** section was added for research purposes.

Methodologies

In this section, we dive into the methodologies employed to address and assess the performance and safety of Large Language Models (LLMs). The methodologies are categorized into two primary domains: i) *Detoxification Methods*; and ii) *Evaluation Methods*. Each category encompasses a range of techniques that serve distinct purposes in the lifecycle of LLMs, from mitigating harmful outputs to rigorously evaluating their effectiveness and bias.

Detoxification Methods

Here we expose some of the main methodologies followed for detoxification of LLMs. These methods can be splitted depending on what part of the lifecycle they are intervening.

Pre-processing Methods These methods seek to balance the bias in the dataset before feeding it to the model so the prevent it from learning toxic and discriminant patterns.

- **Dataset Filtering**. This method focuses on filtering the dataset to prevent the LLM from learning toxic patterns. (Welbl et al. 2021). A common approach is to use a classifier-based method, such as the Perspective API[1], to score toxic samples and then filter them by setting a threshold. However, this technique can reduce the model's ability to detect toxicity and bias (Xu et al. 2021). Furthermore, it has been shown that scoring toxicity based on the Perspective API introduces a bias against African American English (AAE) speakers (Sap et al. 2019).
- **Data Augmentation**. Enhances the ability of the model to detect toxicity and hate by augmenting the data with additional information. (Prabhumoye et al. 2023) proposes two methods, MEDA and INST, which augment the data by providing its toxicity score, computed using the Perspective API (MEDA), or by adding an instruction to the prompt that clarifies whether the text is toxic based on the same score (INST). E.g.

  "Instruction: Complete the following text in a toxic manner. Text:"

In-processing Methods

- **Plug-and-Play Language Model (PPLM)**. PPLM adds an additional layer for toxic discrimination on the hidden representation of an LLM (Dathathri et al. 2020). Essentially, this approach adds a constraint so that the model seeks, not only to reduce the loss function, but also the toxicity of the hidden representation. Nevertheless its effective is proven, (Xu et al. 2021) showed this approach also increases representational bias towards dialect speakers (AAE).
- **Fine-tuning**. Guide the model towards generating less toxic outputs by fine-tuning it on a curated version of the dataset (Gehman et al. 2020). However, (Wang et al. 2022) identified a trade-off between the effectiveness of the detoxification and model's quality.

After-processing Methods

- **Test-time Filtering**. Similar to the *Dataset Filtering* approach, this method aims to filter toxic samples at test time. Like *Dataset Filtering*, it requires a classifier for the automatic annotation of toxic samples. To avoid over-reliance on the Perspective API, (Welbl et al. 2021) employs a different BERT-based model for toxic detection.
- **Reinforcement Learning**. Essentially, a Reward Model is trained from human feedback on the given replies by the original model. Once the reward model is ready, it will be used to guide the original model towards the desired behavior (Ouyang et al. 2022). In the work of (Faal, Schmitt, and Yu 2022), an approach using toxicity-based scores to reward the model.
- **Ensemble models**. Combine collaborative effort from different agents to enhance fairness. (Li et al. 2024) uses prompt engineering to create different personalities out of the same LLM and reach the most fair prompt.

Evaluation Methods

These methods aim to systematically detect and assess biases. The following are some of the main methodologies used for evaluating bias in LLMs:

Intrinsic Evaluation Methods

- **The Matched Guise Technique for Dialect Evaluation**. The Matched Guise Technique (MGT) offers a way to assess how well Large Language Models (LLMs) handle different dialects. This technique involves creating carefully crafted text samples. Each pair presents the same content but uses contrasting dialects, often comparing a specific dialect like African American English (AAE) with Standard American English (SAE). Human evaluators then judge the source dialect (AAE or SAE) for each sample, unaware of which was generated by the LLM. By analyzing how often humans correctly identify the LLM-generated text, we can gauge the model's ability to mimic and adapt its outputs to different dialectal styles.

Extrinsic Evaluation Methods Extrinsic evaluation methods are used to assess the performance of a model within the context of a larger system or real-world application.

- **Task-Based Evaluation**. Task-based evaluation is a critical approach for assessing the practical capabilities of LLMs. Unlike metrics like perplexity that focus on statistical properties, this method directly evaluates how well the model performs on specific tasks involving real-world language usage.

  How it Works:

  - Task Selection: Researchers define a set of tasks that represent real-world language applications. These tasks

[1]https://www.perspectiveapi.com/

can be diverse, ranging from question answering and summarization to dialogue generation and creative writing.

- Task Execution: The LLM is presented with these tasks and asked to complete them. The outputs are then compared against pre-defined criteria or human-generated references.

- Evaluation: Based on the comparison, researchers evaluate the LLM's performance on each task. This evaluation can involve metrics like accuracy, fluency, coherence, or adherence to task instructions.

- **Human Feedback**. Human evaluation is considered the gold standard for assessing their performance. It involves human annotators engaging in tasks like pairwise comparisons of LLM responses to determine if they meet human preferences.

  - The Sample-Efficient Evaluation Approach aims to streamline the evaluation process of Large Language Models (LLMs) by automating the selection of minimal yet informative testing samples. This method draws inspiration from software testing and computational vision domains to identify a minimum set of samples that can serve as counterexamples for falsifying an LLM. By focusing on higher difficulty in falsification as an indicator of LLM superiority, the approach aims to optimize human annotation budget by selecting representative LLMs and assessing their capabilities across multiple levels such as understanding, reasoning, writing, and coding.

Explainability

Explainability in language models is crucial for understanding and trust. It allows users to see how and why models make certain decisions, which is vital in high-stakes applications like medical diagnosis, legal judgments, and hiring. By making the reasoning processes of models transparent, we can identify and mitigate biases, improve model design, and ensure ethical use. The study categorizes the explanations provided by models into seven main types, each illustrating how LLMs rationalize their choices. These explanations, drawn from actual model responses, reveal significant insights into the models' decision-making processes and potential biases.

The study (Kotek, Dockum, and Sun 2023) categorizes the explanations provided by the models into seven main types, each illustrating how LLMs rationalize their choices. These explanations, drawn from actual model responses, reveal significant insights into the models' decision-making processes and potential biases.

- Explanations provided by LLMs for their biased choices often appear authoritative but are factually inaccurate, misleading users about the reasoning process of LLMs.

Types of Explanations

1. **Contextual**: The model cites the situational context as the basis for its choice, favoring interpretations that align with plausible scenarios.
   Example: "In theory, 'he' could refer to the nurse, but it's more logical for 'he' to refer to the doctor, given the professional responsibility context."
2. **Grammatical (Subject)**: The model suggests that pronouns typically refer to the sentence's subject.
   Example: "'He' likely refers to the doctor, as it's the sentence's subject."
3. **Grammatical (Object)**: The model points to the proximity principle, where the pronoun refers to the nearest noun, usually the object.
   Example: "'She' probably refers to the nurse, given its immediate mention before the pronoun."
4. **Gender Bias**: The model's explanation is explicitly rooted in gender stereotypes.
   Example: "'She' can't refer to the doctor, implying the nurse is the only viable female referent."
5. **Ambiguity Acknowledgment**: The model recognizes the sentence's ambiguity, suggesting the pronoun could refer to either noun.
   Example: "The sentence shows pronoun-antecedent ambiguity, making 'he' potentially refer to both the doctor and the teaching assistant."
6. **Repetitive**: The model simply repeats its previous answer without further elucidation.
   Example: "'She' refers to the nurse, aligning with the initial antecedent usage."
7. **Confused**: The explanation appears illogical or confused.
   Example: "'She' cannot refer to the groundskeeper as the sentence treats them as separate from the florist."

- **Observations on Explanations**: The diversity and nature of these explanations shed light on the models' underlying logic and biases. Notably, all models occasionally provided explanations indicative of explicit gender bias. Such explanations often hinged on presumptive gender associations with certain nouns or introduced illogical scenarios to justify gender-stereotypical choices.

  Furthermore, a significant portion of grammatical justifications, particularly those asserting unambiguous pronoun references to subjects or objects, were factually inaccurate due to the inherent ambiguity in the tested sentences. This suggests that models may use grammatical rationales as a facade to mask deeper, bias-influenced decision-making processes.

These methodologies play a crucial role in detoxifying, detecting, and evaluating bias in Large Language Models, providing researchers with tools and techniques to address the challenges associated with implicit toxicity and biased language generation.

Datasets

So far we have discussed about main methodologies used in the scientific community for detection, evaluation and even detoxification of LLMs. However, none of this methodologies can be applied rigorously without a proper dataset. Assessing LLMs presents a significant challenge due to their open-ended language nature. Unlike more straightforward

tasks, like text classification or Part-of-speech tagging, there is no such a "correct answer" in open-ended language tasks, instead we must meticulously analyze the generated output and assess its alignment with desired text characteristics using various metrics. The main strategy followed to build these datasets is to mine pieces of text, either in an automatic or manual manner, but in both carefully framed, that the model will need to complete or classify depending on the followed evaluation method. In this section, we expose the main datasets used for this purpose.

Automatically-crafted Datasets Crafting a dataset big enough to train an LLM is a hard and expensive task, hence many researchers defined strategies to automatically mine the text from the web. These strategies varies depending on the type of bias we aim to evaluate, we look for pieces of text which satisfy a set of requirements like "written in Afro-American English (AAE)" (Blodgett, Green, and O'Connor 2016) or "mentioning a religious belief" (Dhamala et al. 2021). Clearly, this requirements need to be matched somehow either using traditional NLP techniques (i.e. regular expressions), as performed in (Dhamala et al. 2021) for selecting sentences mentioning occupations, or classifier-based approaches, like in (Blodgett, Green, and O'Connor 2016) to identify AAE written text. One popular classifier used for building datasets, particularly for assessing toxic bias, is the *Perspective API*. This tool was employed, for instance, in the creation of the widely recognized REALTOXICITYPROMPTS dataset (Gehman et al. 2020). The process involves setting a toxicity threshold with the Perspective API and then filtering the prompts to retain only those considered as toxic.

Though, we must underscore that text extracted from the internet is not free from bias. This implies our strategy for text mining must be as robust as possible, and that inherent biases have to be meticulously balanced to avoid imprecisions in our evaluations or detoxification methods. Furthemore, possible biases from classifier-based approaches used in the text extraction have to be also considered.

Hand-crafted Datasets Even thought we previously mentioned the difficulty and cost of building a dataset for open-ended language model evaluation, there exist actually some benchmarks that were crafted in a manual or semi-manual way. For example, the WinoBias dataset (Zhao et al. 2018) relies on experts to construct test case scenarios where entities referred by different occupations were paired. Another example is the StereoSet dataset (Nadeem, Bethke, and Reddy 2020), which employs an automated approach to extract a collection of "target terms" representing different social groups from WikiData triples. Subsequently, it leverages Amazon Mechanical Turk (AMT) crowdworkers to generate fill-in-the-blank sentences associating these target terms with both stereotypical and anti-stereotypical descriptors. Evaluations on the generated samples are then conducted by other crowdworkers.

It's crucial to recognize that biases can infiltrate handcrafted datasets as well, often arising from variations in opinions among annotators and experts. This appreciation has led many experts to advocate for a reevaluation of how we define the gold standard for datasets, proposing a perspectivism approach. (Cabitza, Campagner, and Basile 2023) propose not only collecting a single label for each sample but rather gathering as many annotations as possible. This stance gains support from (Nghiem and au2 2024), who introduced HateCOT, a dataset for offensive content detection that supplements human annotations with automatically generated explanations. While these approaches are not immune to bias, they strive for a more balanced perspective.

We have to highlight other datasets frequently used, such as the Common Crawl, OpenWebTextCorpus, the PubMed data, Wikipedia, Arxiv or Github. However, these datasets are not properly cleaned or prepared for bias evaluation, and they must also be meticulously preprocessed. A comparison of the different mentioned datasets can be seen on figure [1]

Metrics

Analyzing large language models (LLMs) involves various metrics to assess their performance, biases, and overall behavior. These metrics provide insights into how well the models perform specific tasks and how they handle sensitive issues such as gender and racial biases. The evaluation of an LLM is strongly dependant on the desired bias to be analyzed, hence most of the papers define their own ad-hoc metrics. To categorize the metrics presented on the different studied papers we rely on the taxonomy presented by (Gallegos et al. 2024).

Generated Text-Based metrics

These metrics are implemented by feeding prompts into the LLMs and analyzing the generated responses. For example, the study might use prompts like "Complete the sentence: The name of the doctor was..." and evaluate the gender or race of the name generated by the model. By running such prompts multiple times and across various professions, the study assesses the distribution of social groups assignments.

- **Gender Parity Score (GPS)**:
  - **Description**: This metric measures the balance between male and female pronouns in the model's outputs. A balanced model should ideally have a GPS close to 1, indicating an equal distribution of male and female pronouns.
  - **Importance**: The GPS helps identify if a model favors one gender over another in its responses, thus providing a clear indicator of gender bias.
- **Stereotype Score**:
  - **Description**: This score evaluates the extent to which the model's predictions align with stereotypes. For instance, it checks if the model associates certain occupations or roles with a specific gender or race based on societal stereotypes rather than actual data.
  - **Importance**: By quantifying how closely model outputs match stereotypical gender roles, the Stereotype

	BOLD	AAE	WinoBias	StereoSet	REALTOXICITYPROMPTS
Nº Samples	23,679	2.2 million	3160	16,995	99,016
Extraction method	Automatic, Classic NLP	Automatic, Classifier-based	Manual, Experts	Hybrid, Crowsourcing	Automatic, Classifier-based
Bias type	Professional, gender, racial, religious, political	Dialectal	Gender	Professional, gender, racial, religious	Toxic
avg. length (words)	6-9	21	7-11	11.70	11.85 (tokens)

Table 1: Comparison of the mentioned datasets

Score reveals the model's tendency to reinforce harmful biases.

- **Word Association Tests (WATs)**:
  - **Description**: WATs assess biases in word associations related to social groups. This involves analyzing how the model associates certain words or phrases with male or female pronouns. For example, it might look at whether words like "nurse" are more frequently associated with female pronouns and "engineer" with male pronouns. Notice it defers from WEATs (Word Embeddings Association Tests) (Caliskan, Bryson, and Narayanan 2017) in the analyzed structure, here we use the model's output while in WEATs we use the word's embedding.
  - **Importance**: WATs provide insight into the subtle biases embedded in the model's language understanding, reflecting deeper societal biases present in the training data.
- **Reward**:
  - **Description**: Measures the average reward of responses based on the reward model trained to prefer implicit toxic outputs.
  - **Importance**: To evaluate how well the reinforcement learning model induces implicit toxicity.
- **Distinct-n:**
  - **Description**: Computes the percentage of unique n-grams among all n-grams generated.
  - **Purpose**: Higher distinct values indicate greater diversity in the generated responses, which can be an indicator of nuanced and varied implicit toxicity.
- **Human annotation**:
  - **Description**: Human annotations shows how much differ model's output from human's expected one. Model's output are labeled or evaluated by humans directly under some pre-defined guidelines.
  - **Importance**: Human annotations are very flexible and show, in a very accurate way, what would be human's behavior. However, they are not scalable and may reproduce the biases present in society.
- **Classifier-Based toxicity**:
  - **Description**: Model outputs are labeled using a classifier trained on assessing toxicity scores. One of the most common classifiers used is the previously mentioned *Perspective API*. However, this metric may be extended to any other toxicity classifier.
  - **Importance**: This metric has a very important advantage, it is highly scalable. We can extent this metric to thousands of samples, which cannot be done for human annotations. Nevertheless, this scalability comes with a trade-off for flexibility. Keep in mind relying to much in classifier-based metrics may also introduce biases from the applied model.

Probability-Based metrics

- **Discovery of Correlations (DisCo)**:
  - **Description**: This metric evaluates the probability of generating a token from a predefined set of adjectives in the position of a masked token. The masked token is embedded in a template sentence, e.g:

    The person is `<X>`. *The person is:* `[MASK]`

    Usually `<X>` is a description or possible sentence said by a person. The original approach presented in (Caliskan, Bryson, and Narayanan 2017) differs from the one used in the analyzed paper of (Hofmann et al. 2024) in how the final metric is computed since Hofmann applies the log ratio of the two social groupds analyzed in the study.
  - **Importance**: As the Stereotype Score, DisCo helps us understand how strong the model associates specific societal stereotypes with certain social groups. The main difference relies on the measured input, while the Stereotype scores is conditioned by model's output, DisCo computes directly the probability over the whole vocabulary, which makes this approach much less dependant on the generated text, possibly uncovering covert biases.
- **Log-Probability Bias Score (LPBS)**:
  - **Description:** Using same concept of the template as *DisCo* normalizes tokens predicted probability with the models prior probability. Then the bias is measured by difference of probability scores for two binary opossing social group words (Kurita et al. 2019). On the approach of (Hofmann et al. 2024) this difference is computed without normalization.
  - **Importance**: LPBS exhibits similar correlations as DisCo. However, while this metric computes a unique value showing the strength of the association of stereotypical set of terms compared to the non-stereotypical one, DisCo focuses on specific terms.

A summary of the mentioned metrics can be seen grouped by paper in table 2. We can notice two important points on this table: i) Most of the papers just apply Generated Text-Based metrics, which, as stated on (Blodgett et al. 2020) is what most of the papers do; they focus on assessing bias in the model's prediction while forgetting about possible biases in datasets or even the metrics themselves. ii) Many papers also rely only on Classifier-based metrics; even if they mention classifier-based metrics may introduce bias, they do not apply a wider range of metrics in order to improve the robustness of their analysis.

Findings

In our exploration, we have delved into the primary methodologies for evaluating bias and mitigating its effects in large language models (LLMs). We have observed that LLMs, pre-trained on datasets embedded with societal biases, tend to reflect those biases in their outputs. These biases span various domains, including gender (Kotek, Dockum, and Sun 2023), race (Hofmann et al. 2024), religion (Dhamala et al. 2021) and political position (Bang et al. 2024). LLMs can sometimes produce toxic prompts (Gehman et al. 2020), which may be implicit and undetectable ((Wen et al. 2023), (Hofmann et al. 2024)).

The necessity for robust benchmarking and measuring of biases in open-ended language generation is evident (Dhamala et al. 2021). Several methodologies exist to detoxify these models. Such methods include augmenting the data with information regarding its toxicity (Prabhumoye et al. 2023; Zhao et al. 2018), fine-tuning them on toxic detection (Nghiem and au2 2024), applying reinforcement learning to reward non-toxicity (Faal, Schmitt, and Yu 2022) or building ensemble models for generation of a more fair response (Li et al. 2024). Although all these methods have been proven to be effective at balancing the biases, most of them come at the trade-off of inserting other biases or lowering model's performance (Welbl et al. 2021; Wang et al. 2022; Xu et al. 2021).

Open challenges

In light of recent findings from multiple studies on the biases and potential harms associated with large language models (LLMs), it is imperative to chart a comprehensive course for future research to address these critical issues. These studies (Kotek, Dockum, and Sun 2023; Hofmann et al. 2024; Wen et al. 2023), provide a robust foundation upon which to build these future efforts.

These discussions underscore the critical need for ongoing research and development to address the complex issues surrounding LLMs. The continuous evolution and increasing capabilities of these models necessitate a focused effort on identifying and mitigating biases to enhance their fairness and ethical deployment. Drawing insights from the aforementioned studies, several key areas for future research emerge that aim to mitigate the potential risks and maximize the societal benefits of LLMs.

Enhanced bias detection and mitigation are critical areas for future research in the development of large language models (LLMs). Recent studies have highlighted the persistence of biases and the inadequacy of current mitigation strategies, necessitating more advanced techniques to detect and address these issues effectively. There is a need for more sophisticated methods to identify biases in LLM outputs (Kotek, Dockum, and Sun 2023). Despite advancements in training methodologies, gender stereotypes remain prevalent. Future research should focus on refining bias detection metrics to capture subtle and overt biases more accurately. This involves developing new algorithms and tools that can analyze and quantify biases in model outputs during both pre-training and fine-tuning phases (Kotek, Dockum, and Sun 2023).

Beyond detection, implementing **robust bias correction algorithms** is essential. Current approaches often fall short in fully addressing biases, as evidenced by the persistent gender stereotypes in LLM outputs(Kotek, Dockum, and Sun 2023). Future work should explore the integration of fairness-aware training methods that adjust for biases during model development. This includes leveraging techniques such as adversarial training, data augmentation, and post-processing corrections to minimize biased behaviors in models.

Building on the critical need for **advanced bias detection and mitigation techniques**, addressing dialect and sociolinguistic biases is equally crucial for ensuring fairness in AI systems. Hofmann underscores the significant impact of dialect on the fairness of AI decisions on its work (Hofmann et al. 2024). Future research should prioritize comprehensive strategies to handle dialectical variations and prevent LLMs from perpetuating sociolinguistic prejudices.

Expanding training datasets to include a wider range of dialects and linguistic variations is essential. This diversity will help models better understand and represent different sociolinguistic groups, ensuring that LLMs are exposed to and can accurately process various dialects, thus reducing biases arising from linguistic differences. Additionally, developing robust evaluation frameworks that account for sociolinguistic diversity is necessary (Hofmann et al. 2024). These frameworks should assess the model's performance across different dialects to identify and mitigate biases, creating metrics and evaluation processes that can detect and measure biases related to dialects, ensuring equitable treatment across diverse linguistic groups.

Implementing **fairness-aware training methods** that adjust for dialectical differences during model development is also critical. Collaborating with sociolinguistic experts can further enhance these efforts. Their insights can guide the creation of more inclusive and representative models, leveraging expertise from sociolinguistics to inform model training and evaluation, ensuring that the models are culturally and linguistically sensitive.

Another significant challenge is **implicit toxicity** (Wen et al. 2023; Hofmann et al. 2024). As we continue to enhance large language models, it is crucial to improve mechanisms for detecting subtle toxic outputs. Future research should focus on refining the quality of reward models used in reinforcement learning, enhancing human feedback loops, and developing sophisticated automated systems capable

Paper	GDPS	Stereotype Score	WATs	DisCo	LPBS	Reward	Distinct-n	Human Annotation	CLS-based Toxicity
(Hofmann et al. 2024)	–	–	–	✓	✓	–	–	–	–
(Kotek, Dockum, and Sun 2023)	✓	✓	✓	–	–	–	–	–	–
(Wen et al. 2023)	–	–	–	–	–	✓	✓	✓	✓
(Dhamala et al. 2021)	✓	–	–	–	–	–	–	–	✓
(Gehman et al. 2020)	–	–	–	–	–	–	–	–	✓
(Welbl et al. 2021)	–	–	–	–	–	–	–	✓	✓
(Prabhumoye et al. 2023)	–	–	–	–	–	–	–	–	✓
(Faal, Schmitt, and Yu 2022)	–	–	–	–	–	–	–	–	✓

Table 2: Comparison of metrics applied among the studied papers.

of identifying and rectifying implicit toxicity in real-time. Moreover, it is essential to broaden the scope of toxicity detection to encompass more nuanced forms of harmful content, ensuring that LLMs do not inadvertently perpetuate or amplify undesirable elements.

Integrating **logical reasoning frameworks** with fairness-aware training methods has shown promise in significantly reducing bias in large language models (LLMs). Future research should focus on merging these logical learning approaches with methods designed to ensure fairness, aiming to produce models that are both less biased and more transparent. This dual approach not only mitigates biases but also enhances the interpretability and ethical deployment of LLMs.

Evaluating the **real-world impact** of LLMs across different domains, such as healthcare, legal, and educational settings, is crucial. Future studies should concentrate on cross-domain evaluations to understand how biases in LLMs manifest in various applications and develop tailored mitigation strategies. Collaboration with domain experts will be essential to ensure the ethical deployment of AI technologies. Such cross-domain studies can reveal unique biases specific to each field and help in developing comprehensive strategies to address them. Additionally, as LLMs continue to evolve, continuous monitoring and dynamic updates to bias detection and mitigation strategies are imperative.

Implementing systems that allow for **real-time feedback** and adjustments can help maintain the fairness and accuracy of LLMs over time. This proactive approach will prevent the entrenchment of biases and ensure that models remain aligned with ethical standards. Regular updates and monitoring can help in quickly identifying and rectifying any new biases that emerge as the models learn and adapt.

Evaluating LLMs using a **perspectivism** approach is also crucial. This approach involves having multiple annotators evaluate the outputs to gather diverse opinions, ensuring a more comprehensive assessment of biases and fairness. By incorporating the perspectives of individuals from different backgrounds, future research can develop a more nuanced understanding of how LLMs perform across various contexts. This method can help identify biases that may be overlooked by a single annotator and contribute to the creation of more equitable and effective AI systems.

Addressing the limitations noted in the study on implicit toxicity, future work should focus on improving the quality of comparison data and designing stronger reward models. Additionally, conducting experiments on extra-large models, such as LLaMA-65B and GPT-3.5-turbo, will provide deeper insights into scaling properties and the potential for implicit toxicity in more powerful models. Understanding these scaling effects is crucial for developing robust mitigation strategies that can be applied to models of varying sizes and complexities. By focusing on these key areas, future research can significantly contribute to the development of fairer, more ethical, and more effective large language models. These efforts will ensure that AI technologies benefit all segments of society equitably and responsibly.

Case study — Matched Guise Probing for Latino American Spanish dialect prejudice

In this section we propose a small case study based on the Matched Guise Probing technique proposed by (Hofmann et al. 2024) on the Latino American Spanish (LESP) compared to the Peninsular Spanish (PESP).

LESP and PESP definition

First of all, we need to somehow define the differences between the two dialects. It is important to notice that this difference is not trivial, many words in LESP come actually from the south of Spain (Andalucía) where the exchange of culture between Latino America and Spain was higher. This gap is even smaller in the canary islands, since this was a critical point in every trip to Latino America. Moreover, the problem had to be simplified due to the difference between many terms used among the different countries in Latino America, this is, the dialect spoken in Argentina is not the same as the one spoken in Mexico, still there are many terms that can be classified as Latino Americans and differ them from the Spanish used in the Iberian Peninsula.

Collected Data

The performance of the experiment needs three type of data elements:

- The LESP and PESP input texts
- The tokens whose associations with LESP vs. PESP we want to analyze
- A set of prompts.

These were obtained relying on some sort of crowdsourcing. The LESP and PESP inputs texts were created thanks to the collaboration of Latino American people that helped to collect some words that differ in their dialects from the

Peninsular Spanish. The tokens are simply the ones generated by the BERT tokenizer when applying it to the adjectives from the Princenton Trilogy (as done by Hofmann on its original work) translated into spanish. Finally the set of prompts are the ones written by Hofmann translated into spanish as well. The translations were performed personally by me and two philologists who assisted me on the task. The total number of pairs is 89, which is actually not a really high amount due to the heavy work of creating them manually. While in english language the gender of words is not a problem, this does not happen with spanish. Hence, for every translation was needed to add another gender, both for the adjectives and the prompts.

Probability computation

The work done by Hofmann takes into account only single-token words, this is not the case for BERT tokenizer in the spanish language, most of the adjectives are composed by more than one token. Hence the computation had to be extended. We can define the probability of generating a word from the list of adjectives given the embedded pair and the model in the following way:

$$P(w|t(u);\theta) = \prod_{x_i \in w} P(x_i|t(u) \oplus x_{i-1};\theta) \tag{1}$$

Notice, some of the adjectives may have more than one gender, therefore the probability of actually generating an adjective needs to be summed in the following way:

$$P(a|t(u);\theta) = \sum_{w \in a} P(w|t(u);\theta) \tag{2}$$

Where x_i denotes each of the tokens inside a word w, and each adjective a may be composed by more than one gender (or word) w. Notice for each word we recompute the probability at each step concatenating ($\oplus$ operator) the previous token so we get the actual probability of the next token of a word given the previous token was generated. Each pair (either in LESP or PESP) is embedded in a prompt, denoted as $t(u)$ and the model is represented by θ.

Metric

For the experiments we analysed the BETO[2] model, a version of BERT trained on a big unannotated corpora in spanish language. The metric to measure the results is the same as the one used in the work of Hofmann, the log ratio (LR) of the probability assigned to each adjective following the LESP and PESP input.

$$LR = \log_{10} \frac{P(a|t(u)}{P(a|t(v)} \tag{3}$$

Where $t(u)$ denotes the embedded pair in Latino American Spanish and $t(v)$ the embedded pair in Peninsular Spanish.

Results

A ranking of the ten top adjectives associated with the LESP dialect are shown on table 3.

[2]https://github.com/dccuchile/beto

attribute	ratio	Log ratio
11	eficiente	0.148125
3	ambicioso	0.135416
1	alerta	0.123672
12	estupido	0.100135
22	perezoso	0.077574
13	fiel	0.039729
33	testarudo	0.036863
20	leal	0.029443
29	sensible	0.023907
4	apasionado	0.021712

Table 3: Top ten adjectives associated with the LESP dialect

Conclusion

As we can notice from the ranking on table 3, both, positive and negative adjectives are present in the top, this could be due to two reasons; i) either the model is not biased towards the LESP dialect, ii) or our experiment needs to be more robust. Given the amount of data used compared to the original experiment performed by (Hofmann et al. 2024), we could agree that another experiment, with a higher amount of pairs, should be performed in order to draw a conclusion. Keep in mind the experiment performed by Hofmann analyses one linguistic feature, the use of the infinite form of the verb *to be* instead of its conjugated version, while here we are analysing many different words in both dialects, hence to make the experiment robust enough we would rather need even more pairs than the ones used by Hofmann, or reduce the linguistic features, for example using just one word that differs in both dialects.

References

Bang, Y.; Chen, D.; Lee, N.; and Fung, P. 2024. Measuring Political Bias in Large Language Models: What Is Said and How It Is Said. arXiv:2403.18932.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. Online: Association for Computational Linguistics.

Blodgett, S. L.; Green, L.; and O'Connor, B. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. arXiv:1608.08868.

Cabitza, F.; Campagner, A.; and Basile, V. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6): 6860–6868.

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

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *International Conference on Learning Representations*.

Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.-W.; and Gupta, R. 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*, FAccT '21. ACM.

Faal, F.; Schmitt, K.; and Yu, J. Y. 2022. Reward modeling for mitigating toxicity in transformer-based language models. *Applied Intelligence*, 53(7): 8421–8435.

Ferrara, E. 2023. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, 6(1): 3.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and Fairness in Large Language Models: A Survey. arXiv:2309.00770.

Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; and Smith, N. A. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369. Online: Association for Computational Linguistics.

Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2024. Dialect prejudice predicts AI decisions about people's character, employability, and criminality. arXiv:2403.00742.

Kotek, H.; Dockum, R.; and Sun, D. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23. ACM.

Kurita, K.; Vyas, N.; Pareek, A.; Black, A. W.; and Tsvetkov, Y. 2019. Measuring Bias in Contextualized Word Representations. In Costa-jussà, M. R.; Hardmeier, C.; Radford, W.; and Webster, K., eds., *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–172. Florence, Italy: Association for Computational Linguistics.

Li, T.; Zhang, X.; Du, C.; Pang, T.; Liu, Q.; Guo, Q.; Shen, C.; and Liu, Y. 2024. Your Large Language Model is Secretly a Fairness Proponent and You Should Prompt it Like One. arXiv:2402.12150.

Nadeem, M.; Bethke, A.; and Reddy, S. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. arXiv:2004.09456.

Nghiem, H.; and au2, H. D. I. 2024. HateCOT: An Explanation-Enhanced Dataset for Generalizable Offensive Speech Detection via Large Language Models. arXiv:2403.11456.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Prabhumoye, S.; Patwary, M.; Shoeybi, M.; and Catanzaro, B. 2023. Adding Instructions during Pretraining: Effective Way of Controlling Toxicity in Language Models. arXiv:2302.07388.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678. Florence, Italy: Association for Computational Linguistics.

Wang, B.; Ping, W.; Xiao, C.; Xu, P.; Patwary, M.; Shoeybi, M.; Li, B.; Anandkumar, A.; and Catanzaro, B. 2022. Exploring the Limits of Domain-Adaptive Training for Detoxifying Large-Scale Language Models. arXiv:2202.04173.

Welbl, J.; Glaese, A.; Uesato, J.; Dathathri, S.; Mellor, J.; Hendricks, L. A.; Anderson, K.; Kohli, P.; Coppin, B.; and Huang, P. 2021. Challenges in Detoxifying Language Models. *CoRR*, abs/2109.07445.

Wen, J.; Ke, P.; Sun, H.; Zhang, Z.; Li, C.; Bai, J.; and Huang, M. 2023. Unveiling the Implicit Toxicity in Large Language Models. *arXiv preprint arXiv:2311.17391*. Warning: This paper discusses and contains content that can be offensive or upsetting.

Xu, A.; Pathak, E.; Wallace, E.; Gururangan, S.; Sap, M.; and Klein, D. 2021. Detoxifying Language Models Risks Marginalizing Minority Voices. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2390–2397. Online: Association for Computational Linguistics.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20. New Orleans, Louisiana: Association for Computational Linguistics.