When in Doubt, Cascade: Towards Building Efficient and Capable Guardrails

Manish Nagireddy[1], **Inkit Padhi**[1], **Soumya Ghosh**[2], **Prasanna Sattigeri**[1]

[1]IBM Research
[2]Merck Research Labs

Abstract

Large language models (LLMs) have convincing performance in a variety of downstream tasks. However, these systems are prone to generating undesirable outputs such as harmful and biased text. In order to remedy such generations, the development of guardrail (or detector) models has gained traction. Motivated by findings from developing a detector for social bias, we adopt the notion of a use-mention distinction - which we identified as the primary source of underperformance in the preliminary versions of our social bias detector. Armed with this information, we describe a fully extensible and reproducible synthetic data generation pipeline which leverages taxonomy-driven instructions to create targeted and labeled data. Using this pipeline, we generate over 300K unique contrastive samples and provide extensive experiments to systematically evaluate performance on a suite of open source datasets. We show that our method achieves competitive performance with a fraction of the cost in compute and offers insight into iteratively developing efficient and capable guardrail models.

Warning: This paper contains examples of text which are toxic, biased, and potentially harmful.

1 Introduction

Large language models (LLMs) contain high potential for a variety of real-world applications, due to their versatility, adaptability, and ease of use, along with their continuously improving performance (OpenAI 2022; Bommasani 2023; Nayak 2019; Perspective API 2021). Yet, their deployment, especially in critical domains such as healthcare and finance, poses significant risks (IBM AI Risk Atlas). A new host of challenges arises with the generative capabilities of these models, as they may produce convincing output, but this output may often be layered with issues around toxicity, bias, hallucinations, and more.

In order to combat the harmful generations from these models, the concept of *guardrail* (or *detector*) models have gained popularity for several reasons. These guardrail models can be more efficient, modular, and scalable (Achintalwar et al. 2024; Rebedea et al. 2023; Inan et al. 2023) than the LLMs whose output they operate on. In this work, we focus on the problem of detecting whether an LLM's textual output contains social bias.

Motivation Social bias can be defined as discrimination for, or against, a person or group, or a set of ideas or beliefs, in a way that is prejudicial or unfair (Webster et al. 2022; Bommasani and Liang 2022). Importantly, text which contains social bias may not contain any explicit or profane content, but may still propagate discrimination (e.g., "I don't want to hire this individual as a babysitter because they have facial scars."). Driven by a clear and present need to automatically detect whether LLM-generated text contains such harmful content, we developed a `social-bias-detector`. To do so, we gathered a collection of open source datasets, with commercially permissible licenses, and used a combination of four datasets as an attempt at a holistic collection of training data. We provide the specific datasets, as well as the hyper-parameters used during training, in the Appendix[1]. From an architectural standpoint, the social-bias-detector is an encoder-only model with just over 100M parameters that was obtained by fine-tuning BERT (Devlin et al. 2019).

Despite reasonable performance on our evaluation sets (specific numbers in the Appendix), we discovered a high false positive issue with this model. Here, a false positive refers to the classification of benign text as harmful. In order to investigate further, we devised an experiment to test the hypothesis that there existed a mismatch between the training dataset (which were largely human-generated or curated) and the distribution of text that is generated by an LLM. Three of the authors manually annotated outputs from the `llama-2-7b-chat`[2] model and we conjectured that our model's sub-optimal performance was due to the excessively intricate and evasive answers, which tend to be generated by highly aligned and verbose models such as Llama 2 (Touvron et al. 2023). However, concurrent to our work, we came across the true reason for our detector's subpar performance - the *use-mention distinction* (Gligoric et al. 2024). We will revisit this paper later in our experimental results, in Section 3.

In the context of social bias detection, the *use-mention distinction* can be thought of as the difference between using text for ill-intent and simply mentioning text without this

[1]For the full supplementary material, please refer to the version of our work available at https://arxiv.org/abs/2407.06323

[2]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

malicious undertone. To elaborate, the text "All Muslims are terrorists" is an example of use, whereas the text "The notion that all Muslims are terrorists is a harmful stereotype" is an example of mention. For further contextualization of the use-mention distinction, as well as the impact of this phenomenon on downstream social bias detection, refer to Section 6. Extended analysis revealed that many of misclassifications from our detector were the result of failing to differentiate use from mention. In particular, a substantial quantity of the responses from the Llama-2 model were of the following flavor: "It is not accurate to say that {toxic_statement}," which precisely maps to a *mention*.

Our main contributions are:

1. A fully extensible and reproducible synthetic data generation pipeline (Section 2) which leverages taxonomy guided instructions in order to generate high quality labeled and contrastive data at scale. Using this pipeline, we create a novel and diverse dataset of over 300K unique samples intended to equip guardrail models with use-mention distinction capabilities.
2. A suite of detectors with parameter count in the $30 - 110$M range along with extensive experiments (Section 3), demonstrating competitive performance (Section 4) with state-of-the-art models while *being more cost effective and compute efficient* (Section 5). We reduce false positive rate from just over 82% to 33%, with minimal impact on false negative rate, *without increasing the size of the underlying model*.

2 Approach

In order to address the use-mention distinction problem that affected our social-bias-detector, we sought to leverage synthetic data generation to augment our model.

2.1 Taxonomy-Driven Synthetic Data Generation

We opted to utilize taxonomy driven generation, as this allowed us to explore a rich and structured manner through which to generate data at scale. First, we generated and iteratively updated a taxonomy of social biases. We started by prompting the Mixtral mixture-of-experts model[3], under greedy decoding, with the instruction: `Generate a high level taxonomy for social bias that shows various groups and associated biases and stereotypes.` After post-processing, we were able to arrive at a viable taxonomy with around 50 social biases across a variety of dimensions. Some examples include Anti-Asian bias, Transphobia, Ethnocentrism (the belief that one's own culture is superior to others), etc. The full taxonomy that we utilized is available in the Appendix. Importantly, the taxonomy provides over 10 high-level topics (e.g., Racial and Ethnic Bias, Gender Bias, Classism, etc.), and a variety of sub-topics for each of these groups. In particular, we were encouraged to see that our generated taxonomy included demographic groups that are not as mainly focused on, but are still extremely relevant for the purposes hate speech detection (Nagireddy et al. 2024). One such example is Polyphobia, or prejudice against individuals who are attracted to multiple partners.

[3]https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

> You are asked to come up with a set of {num_samples} diverse examples for the topic of {topic}, specifically focusing on {sub_topic}.
>
> Please follow these guiding principles when generating responses:
>
> 1. Use proper grammar and punctuation.
> 2. The examples should be diverse in both style and form.
> 3. Each of the examples should be in a JSONL format.
> 4. Each example contains a biased statement and an unbiased statement.
>
> To better assist you with this task, here is an example of the JSONL formatting:
> {{
> "biased": harmful and biased statement with a stereotype relating to {sub_topic},
> "unbiased": unbiased and benign statement with no stereotypes relating to {sub_topic}
> }}
>
> Generate {num_samples} such examples for the topic of {topic}, specifically focusing on {sub_topic}. Remember to follow the principles mentioned above.
>
> Return your responses below in JSONL format.

Figure 1: Prompt for taxonomy-guided and contrastive synthetic data generation

Once we generated this taxonomy, we proceeded with tailored synthetic data generation. Here, we directly leveraged the taxonomy and generated data for each leaf node (which is each sub-topic in our taxonomy). Using the prompt in Figure 1, we repeatedly instructed the aforementioned Mixtral mixture-of-experts model to generate pairs of output - jointly generating one biased and one unbiased statement. For reproducibility, we generated under nucleus sampling (Holtzman et al. 2020) with `top_p` at 0.95, `top_k` at 100, `temp` at 0.7, and `max_new_tokens` set to 1024. This method was desirable for a couple of reasons. First, the generation directly utilized the taxonomy, as each call for a generation contained both the root node (i.e., the over-arching topic) as well as the leaf node (i.e., the specific bias) directly in the context for the prompt. Next, requiring JSON formatting in the output allowed for easy post-processing of the LLM-generated text. Finally, and perhaps most interestingly, asking for paired output resulted in the generation

of contrastive data. Specifically, we observed that each pair of {biased, unbiased} text precisely mapped to {use, mention} examples! Refer to Table 1 for the flavor of synthetic data obtained from this procedure.

In total, we generated just under 1 million pairs of data, and filtered it down to just over 300 thousand individual samples after removing duplicates. This took around 1 day to generate given our model hosting options. We plan to openly release this dataset.

2.2 Our Suite of Social-Bias-Detectors

social-bias refers to the models trained on the four human curated datasets mentioned in Section 2 and elaborated upon in the Appendix. In particular, we have three such models - each trained on top of a different base model. social-bias is trained on top of BERT[4], social-bias-distil is trained on top of a transformer architecture (Trivedi et al. 2023) which provides most of the accuracy of a BERT-like model (Devlin et al. 2019), while being seven times faster on a CPU and two times faster on a GPU. Finally, social-bias-toxigen is trained on top of toxigen_hatebert[5](Hartvigsen et al. 2022). Note that these three models were trained with only human curated datasets.

social-bias-use-mention refers to the models trained on the synthetic data that was generated according to Section 2.1. In particular, we have four such models. We note that social-bias-use-mention, social-bias-use-mention-distil, and social-bias-use-mention-toxigen were trained on the entire set of unique instances from the synthetically generated data, with the base models following the same convention as above. Additionally, we trained social-bias-use-mention-onetrial, which saw only one iteration of synthetic data in training. We trained this model in order to provide a frame of reference for the amount of synthetic data that may be required for various levels of performance. Note that these four models have only seen synthetic data in training.

Finally, we trained social-bias-onetrial-concat, which contained all four human curated datasets as well as one trial of synthetic data. We trained this model as another comparison point to determine if combining human curated and synthetic data resulted in better performance.

We provide full details about each of the above models and their training data in the Appendix.

The Cascade Approach We also experimented with what we refer to as the cascade approach, where we utilize two models in a sequential manner. Given some text, we first identify a model, m_{bias}, to serve as a preliminary arbiter of whether this text contains bias or not. Then, if the output from m_{bias} is the harm label, we also run the text through another model, m_{use}. This model determines if the text is a case of *use* or *mention*, as defined in Section 1. If the output from m_{use} is *use*, then we assign the label of bias to the text. Otherwise, we assign the label of not_bias - indicating that either m_{bias} assigned a label of not_bias or m_{use} assigned a label of *mention* to this text. Refer to Algorithm 1 for specific pseudo-code of the cascade approach.

We hypothesized that the cascade would be a "best-of-both-worlds" approach where m_{bias} would do a decent job on aggregate, due to the high quality human curated datasets that it has seen. Then, to precisely combat the use-mention distinction issue, where m_{bias} would incorrectly flag text that is a mention as harmful, we will run any harm-labeled instances from m_{bias} through m_{use}, a model that is specifically trained, by way of our synthetic data, to distinguish use from mention.

We have four cascade approaches, defined as follows:

Algorithm 1: Cascade method

m_{bias}: a model which labels text as *social_bias* or *not_bias*
m_{use}: a model which labels text as *use* or *mention*
$data$: some dataset with a list of responses from an LLM
$labels$: list of labels for each response in $data$
procedure CASCADE(m_{bias}, m_{use}, $data$)
    $labels \leftarrow \{\}$ ▷ initialize label list
    for each response r in $data$ **do**
        $label_{m_{bias}} \leftarrow m_{bias}(r)$ ▷ run *r* through m_{bias}
        if $label_{m_{bias}}$ is *social_bias* **then** ▷ need to run through use-mention model
            $label_r \leftarrow m_{use}(r)$ ▷ run *r* through m_{use}
        else
            $label_r \leftarrow label_{m_{bias}}$ ▷ label is *not_bias* from m_{bias}
        end if
        $labels \leftarrow labels + label_r$ ▷ append label
    end for
    return $labels$
end procedure

3 Experiments

In this section, we first describe the evaluations that we conducted and then provide empirical evidence for the cascade approach (defined in Algorithm 1) and its utility.

3.1 Evaluation Datasets and Baselines

Collecting Evaluation Data Due to the fact that our primary axis for evaluation is the use-mention distinction, we sought to construct an evaluation set which reflected this goal. Even though datasets for such a task are not widely available (Gligoric et al. 2024), we leveraged the rich literature on counter-speech and counter-narratives in the context of hate speech. In particular, similar to (Gligoric et al. 2024), we take the Knowledge-grounded hate countering (Chung, Tekiroğlu, and Guerini 2021) and the Multi-Target CONAN (Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco 2021) datasets. The knowledge-grounded dataset contains 195 hate speech and counter-narrative pairs covering multiple hate targets (islamophobia, misogyny, antisemitism, racism, and homophobia). The counter narratives are written by an expert

[4]https://huggingface.co/google-bert/bert-base-uncased
[5]https://huggingface.co/tomh/toxigen_hatebert

biased $\longrightarrow$ use	unbiased $\longrightarrow$ mention
All Romani people are fortune tellers and rely on trickery to make a living.	Some individuals from the Romani community may practice fortune telling, while others have different occupations and interests.
You can't trust someone with schizophrenia because they hear voices and can't tell what's real.	People with schizophrenia may experience auditory hallucinations, but they can still be trustworthy and reliable individuals with the right treatment and support.
Pagans are superficial and materialistic.	Pagans have different values and priorities, and some may emphasize spiritual growth, social justice, and environmental sustainability.

Table 1: Example of Contrastive Synthetic Data

name	m_{bias}	m_{use}
`cascade-orig`	`social-bias`	`social-bias-use-mention`
`cascade-onetrial`	`social-bias`	`social-bias-use-mention-onetrial`
`cascade-distil`	`social-bias-distil`	`social-bias-use-mention-distil`
`cascade-toxigen`	`social-bias-toxigen`	`social-bias-use-mention-toxigen`

Table 2: Cascade Approach Model Combinations

who is tasked with composing a suitable counter-narrative response to a given hate speech using the corresponding knowledge as much as possible (Chung, Tekiroğlu, and Guerini 2021). The Multi-Target CONAN dataset consists of 5003 hate Speech and counter-narrative pairs covering multiple hate targets, including disabled, Jews, LGBT+, migrants, Muslims, people of color (POC), women. The dataset is constructed using a novel human-in-the-loop data collection methodology (Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco 2021).

We combined both datasets and labeled each pair of {hate speech, counter narrative} as {bias, not bias}. In addition, we noted that a label of *bias* most precisely meant a label of *use* and similarly for *not bias* and *mention*. This is because counter-narratives are written such that they directly counteract the harmful hate speech (Gligoric et al. 2024), which results in them being excellent examples of mentions.

Experimental Setup We provide two sets of evaluations below. First, in order to compare with the results from (Gligoric et al. 2024), we utilize the same evaluation set of 180 total examples (also taken from the above two datasets), including both hate speech and counter-narratives. Second, we combine the entirety of the Knowledge-grounded hate countering and Multi-Target CONAN datasets to arrive at an evaluation set of size $10,396$[6].

Baselines In addition to the reported numbers with three of the GPT models from (Gligoric et al. 2024), we also report numbers using Llama-Guard[7] and Llama-Guard-2[8]. We use these two models as both a competitive baseline as well as a point of comparison due to the size and latency of these models. To reiterate, the detectors are on the order of 100M parameters, Llama-Guard models are either 7B or 8B parameters, and GPT-4 is rumored to be orders of magnitude larger (OpenAI et al. 2024).

4 Experimental Results

In this section, we provide results on the aforementioned datasets, with all of our detectors and baselines. We report three metrics of interest: false positive rate (FPR), false negative rate (FNR), and average error rate (Avg Err) - similar to (Gligoric et al. 2024). For clarification, the average error rate is the average of the FPR and FNR. To further contextualize these metrics, a false positive refers to incorrectly providing a label of `bias` to benign text. Moreover, because all of the examples of benign text are examples of *mentions* and all the examples of harmful text are *uses*, a false positive represents incorrectly flagging a mention as use. Symmetrically, a false negative refers to incorrectly flagging a use as mention. For our purposes, we note that both false positives and false negatives are important. False positives represent improper moderation by flagging benign text whereas false negatives represent missed detection by failing to flag harmful text.

We provide results below and note for all of the detectors (which are encoder-only models), generation is deterministic. For Llama-Guard models, we use a standard template[9] and greedy decoding.

5 Discussion

5.1 Results for 180 Examples Test Set

First, we comment on results for the evaluation set of 180 examples from (Gligoric et al. 2024). Despite the small size of this test, we wanted to demonstrate performance because

[6]This comes from combining 195 pairs with 5003 pairs from the Knowledge-grounded hate countering and Multi-Target CONAN datasets, respectively.

[7]https://huggingface.co/meta-llama/LlamaGuard-7b

[8]https://huggingface.co/meta-llama/Meta-Llama-Guard-2-8B

[9]A sample notebook for inference with Llama-Guard models is available from the HuggingFace pages above

model	FPR	FNR	Avg Err	# params
gpt-3.5-instruct-turbo*	25.56	13.33	19.44	?
gpt-3.5-turbo (ChatGPT 3.5)*	11.11	22.22	16.67	?
gpt-4*	8.89	20.00	14.44	?
Llama-Guard	12.22	20.00	16.11	7B
Llama-Guard-2	***4.44***	***26.67***	***15.56***	8B
toxigen-hatebert	53.33	12.22	32.78	110M
social-bias	90.00	18.89	54.44	110M
social-bias-use-mention	34.44	23.33	28.89	110M
social-bias-use-mention-onetrial	32.22	37.78	35.00	110M
social-bias-onetrial-concat	72.22	20.00	46.11	110M
cascade-orig	32.22	33.33	32.78	110M x 2
cascade-onetrial	26.67	47.78	37.22	110M x 2
social-bias-distil	95.56	5.56	50.56	39M
social-bias-use-mention-distil	26.67	24.44	25.56	39M
cascade-distil	**24.44**	**28.89**	**26.67**	39M x 2

Table 3: Results on Evaluation Set with 180 examples, * denotes results taken from (Gligoric et al. 2024)

model	FPR	FNR	Avg Err	# params
Llama-Guard	9.27	5.44	7.36	7B
Llama-Guard-2	***1.83***	***16.62***	***9.22***	8B
toxigen-hatebert	45.29	45.29	26.39	110M
social-bias	82.40	4.96	43.68	110M
social-bias-use-mention	36.63	4.10	20.36	110M
social-bias-use-mention-onetrial	43.31	9.37	26.34	110M
social-bias-onetrial-concat	70.55	3.96	37.25	110M
cascade-orig	**32.69**	**8.31**	**19.84**	110M x 2
cascade-onetrial	35.30	13.49	24.39	110M x 2
social-bias-distil	96.58	1.77	49.17	39M
social-bias-use-mention-distil	38.15	5.10	21.62	39M
cascade-distil	37.26	6.68	21.97	39M x 2

Table 4: Results on Combined Evaluation Set with 10K examples

(Gligoric et al. 2024) provided us with points of comparison to the GPT-family of models. In particular, we observe that our cascade models provide competitive performance with the GPT models. We make particular note of `cascade-onetrial`. Recall that the model which determines use or mention (m_{use}) for `cascade-onetrial` has only seen one iteration of synthetic data generation, making it extremely desirable, as the burden of training data is substantially reduced. In absolute terms, (m_{use}) for `cascade-onetrial` saw around 1K examples in training, whereas (m_{use}) for `cascade-distil` saw around 190K examples in training (refer to the Appendix) for full details). Therefore, we are able to see that for this (albeit limited) set of data, the cascade approach is able to perform on par with some of the most widely used and largest models, while having a fraction of the computational cost. We do notice that both Llama-Guard models surpass even GPT-4 by a significant margin. Nevertheless, for this set of 180 examples, the cascade approach takes under 1 minute on a single A100 GPU, and is even able to be run on a CPU (taking a few hours). In comparison, access to GPT models is limited to querying via expensive external APIs, and inference with Llama-Guard models is not possible on a CPU.

5.2 Results for Full (10K Examples) Test Set

Next, we comment on results for the combined evaluation set of 10K examples. Here, we first point to the stellar performance from both Llama-Guard models. However, as previously mentioned, inference with Llama-Guard models is not possible on a CPU and takes several hours on an A100 GPU for a test set of this size. For our detectors, we observe that all of the detectors which have *only seen human curated data* (`social-bias`, `social-bias-distil`, and `social-bias-toxigen`) have extremely high false positive rates. This implies that they perform poorly at distinguishing between use and mention, which reaf-

firms our initial motivations described in Section 1. Even when we add a little bit of synthetic data and combine it with the human curated data, as in the case of `social-bias-onetrial-concat`, we see a rather high false positive rate. We also find better, but not optimal, performance in the detectors which have *only seen synthetic data* (`social-bias-use-mention`, `social-bias-use-mention-distil`, and `social-bias-use-mention-toxigen`). These models tend to have lower false positive rates, but still not the best performance - implying that human curated data still contains some signal or information which is not present in the synthetic data.

This brings us to the performance of our cascade approaches. We observe the lowest FPR among all of our detectors in `cascade-orig`. On aggregate, the cascade approaches have lower false positive rates, at the cost of slightly higher false negative rates. This is a direct consequence of the approach itself (defined in Algorithm 1), as the method only alters labels that have already been labeled as 1 or harm (by m_{bias}). However, we see that the cascades are still under-performing Llama-Guard-2. We find that `cascade-orig` is around 10% worse in terms of average error rate than Llama-Guard-2, but interestingly, the cascade approaches do better in terms of false negative rate - which suggests that they are better than Llama-Guard-2 at understanding which data points are strictly cases of hate speech or *use*. Nevertheless, we observe that there is still a decent gap in the false positive rates between any cascade approach and the Llama-Guard models - implying underperformance in discerning between use and mention, with the cascade approaches having a tendency to provide a label of `bias`, leading to inaccurately labeling true *mentions* as *uses*. We observe that `cascade-onetrial` performs slightly worse than the best cascade, `cascade-orig`, but not too dramatically so. This is important again for the same computational efficiency reasons that were previously mentioned.

Significance of Low FNR Despite the gap in false positive rates, we bring attention to the low false negative rates for our approaches. In fact, *all* of the detectors and cascades have lower FNRs than Llama-Guard-2. Importantly, false negatives are directly representative of a detector's utility as a guardrail. A high false negative rate implies that the detector fails to flag text that is harmful, thereby allowing toxic content to be displayed to an end user. Hence, it is imperative to retain a lower FNR such that the detectors function as effective guardrails. We are encouraged to see as much as a 50% decrease in FPR, at minimal expense to FNR, when comparing `social-bias` with `cascade-orig` - indicating the effectiveness of our synthetic data.

5.3 Extended Analysis on Llama-Guard-2

Ignoring the increased computational overhead, we further explore the differences in performance (with respect to false positives) when comparing our detectors with the Llama-Guard models. At an instance level, we explored the differences in predictive quality between Llama-Guard and our detectors. For example, the sample *Did you know that children with Down syndrome, if surrounded by regularly-developed peers, receive a fundamental boost in their development? This "mix" fosters a positive learning atmosphere and increases engagement and attention in all students. This mean that children with Down syndrome would be a burden to society if isolated, not the contrary.* was incorrectly flagged as `unsafe` by Llama-Guard-2, but properly classified as `not_bias` by our `cascade-orig` method. Another such sample is *Even if all British Muslims had cancer, you are safe: you cannot catch it from anyone! Cancer cannot be spread, since it is a non contagious disease. Also, it is worth considering that illnesses are not connected to religions.*

In both of these cases, we observe the presence of *mentioning* text, which is throwing off even Llama-Guard-2. Hence, the reason for the `cascade-orig` method being able to properly classify these instances stems from the deliberate synthetic data generation and subsequent training for the use-mention distinction, but also highlights the need for continued improvement on our detectors as well to properly capture this phenomenon.

6 Related Work

Social Bias in Language Models Social bias can be defined as discrimination for, or against, a person or group, or a set of ideas or beliefs, in a way that is prejudicial or unfair (Webster et al. 2022; Bommasani and Liang 2022). Numerous studies have demonstrated that generative models exhibit undesirable behavior that amplifies social bias (Blodgett et al. 2020; Parrish et al. 2022; Smith et al. 2022; Selvam et al. 2023; Dhamala et al. 2021; Nagireddy et al. 2024). Additionally, the deficiencies of current datasets (Blodgett et al. 2021) and opacity of defining what constitutes social bias in language models and their measures (Blodgett et al. 2020; Selvam et al. 2023; Achintalwar et al. 2024) have emerged.

Use-Mention Distinction As discussed briefly at the end of Section 1, the *use-mention distinction* can be thought of as the difference between using text for ill-intent and simply mentioning text without this malicious undertone. Although a subtle difference, this phenomenon is essential for many downstream applications. For example, a significant amount of content online falls under the umbrella of mentions, such as counter-speech, media reporting, education, and legal settings (Gligoric et al. 2024; Mun et al. 2023; Kirk et al. 2022; Henderson et al. 2022; Wexler, Robbennolt, and Murphy 2019). Of particular use to us is the notion of counter-speech, which refers to speech produced by users of online platforms to counteract harmful speech of others (Gligoric et al. 2024). By definition, counter-speech is an example of *mention*, and thus detection systems which are unable to distinguish use from mention risk contributing to downstream harm - such as improper removal of counter-speech. This in turn reduces opportunities to rectify false narratives and risks further censoring those already most affected by harmful language (Gligoric et al. 2024). Recently, there has also been work which documents the real experiences of counter-speakers and the ways in which to address

existing barriers (Mun et al. 2024). An exciting by-product of our work was our own introduction to this rich space, and we believe that the area of counter-speech will continue to provide essential help in the plight of detecting and mitigating harmful content produced by generative models.

Guardrails for LLMs There is a growing set of rich literature on guardrail models and how they can be more efficient, modular, and scalable (Achintalwar et al. 2024; Rebedea et al. 2023; Inan et al. 2023). NeMo Guardrails (Rebedea et al. 2023) is an open-source toolkit for adding modular and programmable guardrails to LLM-based conversational systems. Llama-Guard (Inan et al. 2023) is an LLM-based input-output safeguard model which has a customizable taxonomy of harms. Finally, the authors in (Achintalwar et al. 2024) go over insights from developing a host of guardrail-like models, which they refer to as detectors.

Synthetic Data Generation Central to our work is the leveraging of capable large language models to generate synthetic data which can be used to train our guardrail models. In particular, we leverage taxonomy-guided data generation. Previous work such as GLAN (Li et al. 2024) utilizes a pre-curated taxonomy of human knowledge and capabilities as input and generates large-scale synthetic instruction data. However, as pointed out by (Sudalairaj et al. 2024), this method relies using the proprietary GPT-4 (OpenAI et al. 2024) as the teacher model, which imposes restrictions on the downstream commercial usability of the generated data. Concretely, the terms of use of proprietary models obfuscate the viability of training commercially viable models using data that is generated from proprietary or otherwise closed models - which many times forbid using their model to improve other models [10]. Thus, similar to (Sudalairaj et al. 2024), we utilize the open source Mixtral[11] model to generate our data. Additionally, the taxonomy guides the generation of the data, as it enables targeted coverage around the individual leaf nodes of the taxonomy (Sudalairaj et al. 2024).

7 Limitations and Future Work

7.1 Limitations

Coverage of Synthetic Data We acknowledge that our taxonomy (even the full version in the Appendix) cannot cover all groups which are salient for social bias. This taxonomy is a point-in-time artifact, and we note that social bias dynamically evolves over time. Because social harms are the product of context-dependent classification systems with deep historical roots and are socially and morally charged, careful attention must be paid to the choices made (such as the groups included in a taxonomy) during development of the detectors (Achintalwar et al. 2024). Concretely, we understand that generated counter-speech may not always be reflective of real-world harm. One way to combat this, though expensive, is via human annotations with true counter-speakers (Mun et al. 2024). Another direction is to leverage LLMs themselves to provide feedback on a given counter-speech example (Jones et al. 2024). We are experimenting with both avenues in the near future.

We also point out that closed-source LLMs tend to generate high quality data, but their proprietary nature is prohibitive for downstream use. Hence our decision to opt against use of GPT models for generation.

7.2 Future Work

Calibration and Confidence-Based Improvements In order to combat overconfidence, we are considering conformal prediction approaches (Vovk, Gammerman, and Saunders 1999). These approaches quantify uncertainty in a model's prediction by constructing *predictive sets* (as opposed to singleton labels, in the case of our detectors) with guaranteed frequentist coverage probabilities. Specifically, we are looking into the regularized adaptive prediction sets approach (Romano, Sesia, and Candes 2020; Angelopoulos et al. 2021) which, in addition to providing coverage guarantees, produces larger (or non-singleton in the case of our detectors) prediction sets for difficult instances and smaller (singleton) sets for easier to classify examples. Then, on the instances for which our models produce these larger sets, we are exploring is the idea of a "confident-cascade" wherein we adopt a collaborative method and offload these inputs to Llama-Guard. We conjecture this may provide a reasonable trade-off between computational efficiency and performance.

Novel Development Currently, detectors are being trained in a supervised fine-tuning method with the standard binary cross-entropy loss. Given that our synthetic data is generated in a contrastive manner, we will consider training with contrastive loss (Khosla et al. 2021). Additionally, we are looking into directly training a multi-head detector, which eliminates the need for two models in our cascade approach.

8 Conclusion

In this work, we began with insights from the development of a social bias detector. Post deployment, we recounted our realization that issues arose due to the use-mention distinction. Motivated by this discovery, we described an extensible and reproducible synthetic data generation pipeline, which leverages taxonomy guided instructions in order to generate high quality labeled and contrastive data at scale. We then documented the training procedures of various models which utilized this synthetic data and introduced the cascade approach. Next, we outlined the extensive experiments performed with these models on a variety of evaluation datasets. We revealed that the cascade approach provided competitive performance on these evaluation sets, in addition to being more substantially more cost effective and compute efficient. We hope that our findings contribute to the growing body of work on building efficient and capable guardrails for large language models.

[10] See point v under License Rights in https://llama.meta.com/llama3/license/

[11] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

References

Achintalwar, S.; Garcia, A. A.; Anaby-Tavor, A.; Baldini, I.; Berger, S. E.; Bhattacharjee, B.; Bouneffouf, D.; Chaudhury, S.; Chen, P.-Y.; Chiazor, L.; Daly, E. M.; de Paula, R. A.; Dognin, P.; Farchi, E.; Ghosh, S.; Hind, M.; Horesh, R.; Kour, G.; Lee, J. Y.; Miehling, E.; Murugesan, K.; Nagireddy, M.; Padhi, I.; Piorkowski, D.; Rawat, A.; Raz, O.; Sattigeri, P.; Strobelt, H.; Swaminathan, S.; Tillmann, C.; Trivedi, A.; Varshney, K. R.; Wei, D.; Witherspooon, S.; and Zalmanovici, M. 2024. Detectors for Safe and Reliable LLMs: Implementations, Uses, and Limitations. *arXiv preprint arXiv:2403.06009*.

Angelopoulos, A. N.; Bates, S.; Jordan, M.; and Malik, J. 2021. Uncertainty Sets for Image Classifiers using Conformal Prediction. In *International Conference on Learning Representations*.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Blodgett, S. L.; Lopez, G.; Olteanu, A.; Sim, R.; and Wallach, H. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Bommasani, R. 2023. AI Spring? Four Takeaways from Major Releases in Foundation Models.

Bommasani, R.; and Liang, P. 2022. Trustworthy Social Bias Measurement.

Chung, Y.-L.; Tekiroğlu, S. S.; and Guerini, M. 2021. Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.; and Gupta, R. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*, 862–872.

Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Gligoric, K.; Cheng, M.; Zheng, L.; Durmus, E.; and Jurafsky, D. 2024. NLP Systems That Can't Tell Use from Mention Censor Counterspeech, but Teaching the Distinction Helps. arXiv:2404.01651.

Hartvigsen, T.; Gabriel, S.; Palangi, H.; Sap, M.; Ray, D.; and Kamar, E. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3309–3326. Dublin, Ireland: Association for Computational Linguistics.

Henderson, P.; Krass, M.; Zheng, L.; Guha, N.; Manning, C. D.; Jurafsky, D.; and Ho, D. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 29217–29234. Curran Associates, Inc.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.

IBM AI Risk Atlas. 2023. Https://www.ibm.com/docs/en/watsonx-as-a-service?topic=ai-risk-atlas.

Inan, H.; Upasani, K.; Chi, J.; Rungta, R.; Iyer, K.; Mao, Y.; Tontchev, M.; Hu, Q.; Fuller, B.; Testuggine, D.; and Khabsa, M. 2023. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. arXiv:2312.06674.

Jones, J.; Mo, L.; Fosler-Lussier, E.; and Sun, H. 2024. A Multi-Aspect Framework for Counter Narrative Evaluation using Large Language Models. arXiv:2402.11676.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2021. Supervised Contrastive Learning. arXiv:2004.11362.

Kirk, H.; Birhane, A.; Vidgen, B.; and Derczynski, L. 2022. Handling and Presenting Harmful Text in NLP Research. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 497–510. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Li, H.; Dong, Q.; Tang, Z.; Wang, C.; Zhang, X.; Huang, H.; Huang, S.; Huang, X.; Huang, Z.; Zhang, D.; Gu, Y.; Cheng, X.; Wang, X.; Chen, S.-Q.; Dong, L.; Lu, W.; Sui, Z.; Wang, B.; Lam, W.; and Wei, F. 2024. Synthetic Data (Almost) from Scratch: Generalized Instruction Tuning for Language Models. arXiv:2402.13064.

Mun, J.; Allaway, E.; Yerukola, A.; Vianna, L.; Leslie, S.-J.; and Sap, M. 2023. Beyond Denouncing Hate: Strategies for Countering Implied Biases and Stereotypes in Language. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 9759–9777. Singapore: Association for Computational Linguistics.

Mun, J.; Buerger, C.; Liang, J. T.; Garland, J.; and Sap, M. 2024. Counterspeakers' Perspectives: Unveiling Barriers and AI Needs in the Fight against Online Hate. In *Proceedings of the CHI Conference on Human Factors in Comput-*

ing Systems, CHI '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400703300.

Nagireddy, M.; Chiazor, L.; Singh, M.; and Baldini, I. 2024. SocialStigmaQA: A Benchmark to Uncover Stigma Amplification in Generative Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19): 21454–21462.

Nayak, P. 2019. Understanding searches better than ever before.

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. [https://openai.com/blog/chatgpt/Online].

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.; Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaftan, T.; Łukasz Kaiser; Kamali, A.; Kanitscheider, I.; Keskar, N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.; Kirchner, J. H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Łukasz Kondraciuk; Kondrich, A.; Konstantinidis, A.; Kosic, K.; Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike, J.; Leung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.; Litwin, M.; Lopez, T.; Lowe, R.; Lue, P.; Makanju, A.; Malfacini, K.; Manning, S.; Markov, T.; Markovski, Y.; Martin, B.; Mayer, K.; Mayne, A.; McGrew, B.; McKinney, S. M.; McLeavey, C.; McMillan, P.; McNeil, J.; Medina, D.; Mehta, A.; Menick, J.; Metz, L.; Mishchenko, A.; Mishkin, P.; Monaco, V.; Morikawa, E.; Mossing, D.; Mu, T.; Murati, M.; Murk, O.; Mély, D.; Nair, A.; Nakano, R.; Nayak, R.; Neelakantan, A.; Ngo, R.; Noh, H.; Ouyang, L.; O'Keefe, C.; Pachocki, J.; Paino, A.; Palermo, J.; Pantuliano, A.; Parascandolo, G.; Parish, J.; Parparita, E.; Passos, A.; Pavlov, M.; Peng, A.; Perelman, A.; de Avila Belbute Peres, F.; Petrov, M.; de Oliveira Pinto, H. P.; Michael; Pokorny; Pokrass, M.; Pong, V. H.; Powell, T.; Power, A.; Power, B.; Proehl, E.; Puri, R.; Radford, A.; Rae, J.; Ramesh, A.; Raymond, C.; Real, F.; Rimbach, K.; Ross, C.; Rotsted, B.; Roussez, H.; Ryder, N.; Saltarelli, M.; Sanders, T.; Santurkar, S.; Sastry, G.; Schmidt, H.; Schnurr, D.; Schulman, J.; Selsam, D.; Sheppard, K.; Sherbakov, T.; Shieh, J.; Shoker, S.; Shyam, P.; Sidor, S.; Sigler, E.; Simens, M.; Sitkin, J.; Slama, K.; Sohl, I.; Sokolowsky, B.; Song, Y.; Staudacher, N.; Such, F. P.; Summers, N.; Sutskever, I.; Tang, J.; Tezak, N.; Thompson, M. B.; Tillet, P.; Tootoonchian, A.; Tseng, E.; Tuggle, P.; Turley, N.; Tworek, J.; Uribe, J. F. C.; Vallone, A.; Vijayvergiya, A.; Voss, C.; Wainwright, C.; Wang, J. J.; Wang, A.; Wang, B.; Ward, J.; Wei, J.; Weinmann, C.; Welihinda, A.; Welinder, P.; Weng, J.; Weng, L.; Wiethoff, M.; Willner, D.; Winter, C.; Wolrich, S.; Wong, H.; Workman, L.; Wu, S.; Wu, J.; Wu, M.; Xiao, K.; Xu, T.; Yoo, S.; Yu, K.; Yuan, Q.; Zaremba, W.; Zellers, R.; Zhang, C.; Zhang, M.; Zhao, S.; Zheng, T.; Zhuang, J.; Zhuk, W.; and Zoph, B. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Perspective API. 2021. Using Machine Learning to Reduce Toxicity Online. [https://perspectiveapi.com/how-it-works/Online; accessed 21-July-2021].

Rebedea, T.; Dinu, R.; Sreedhar, M. N.; Parisien, C.; and Cohen, J. 2023. NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails. In Feng, Y.; and Lefever, E., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 431–445. Singapore: Association for Computational Linguistics.

Romano, Y.; Sesia, M.; and Candes, E. 2020. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33: 3581–3591.

Selvam, N.; Dev, S.; Khashabi, D.; Khot, T.; and Chang, K.-W. 2023. The Tail Wagging the Dog: Dataset Construction Biases of Social Bias Benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1373–1386. Toronto, Canada: Association for Computational Linguistics.

Smith, E. M.; Hall, M.; Kambadur, M.; Presani, E.; and Williams, A. 2022. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9180–9211. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Sudalairaj, S.; Bhandwaldar, A.; Pareja, A.; Xu, K.; Cox, D. D.; and Srivastava, A. 2024. LAB: Large-Scale Alignment for ChatBots. arXiv:2403.01081.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang,

B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Trivedi, A.; Udagawa, T.; Merler, M.; Panda, R.; El-Kurdi, Y.; and Bhattacharjee, B. 2023. Neural Architecture Search for Effective Teacher-Student Knowledge Transfer in Language Models. arXiv:2303.09639.

Vovk, V.; Gammerman, A.; and Saunders, C. 1999. Machine-learning applications of algorithmic randomness. In *Proceedings of the International Conference on Machine Learning*.

Webster, C. S.; Taylor, S.; Thomas, C.; and Weller, J. M. 2022. Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations. *BJA education*, 22(4): 131—137.

Wexler, L.; Robbennolt, J.; and Murphy, C. 2019. #MeToo, Time's Up, and Theories of Justice. *University of Illinois Law Review*, 2019(1): 45–111.