Evaluating Toxicity Understanding of LLM Agents

Ramaravind Kommiya Mothilal
Faculty of Information
University of Toronto
ram.mothilal@mail.utoronto.ca

Syed Ishtiaque Ahmed
Department of Computer Science
University of Toronto
ishtiaque@cs.toronto.edu

Shion Guha
Faculty of Information
University of Toronto
shion.guha@utoronto.ca

***Abstract*—Research on toxicity in LLMs has largely focused on detection tasks, such as identifying hate speech or stereotyping in texts. Recently, these tasks have increasingly been embedded in agentic workflows, where LLMs autonomously query external APIs and reason over results before responding. This shift promotes the perception that LLMs exhibit an "understanding" of toxicity, yet how such understanding can be meaningfully interpreted by humans remains unclear. In this position paper, we first unpack this oversight by highlighting the fundamental gaps in current literature and then propose a framework for evaluating toxicity understanding of agentic LLMs. Overall, this short paper aims to shift the discourse from improving toxicity detection in LLMs to evaluating how LLMs *understand* toxicity in order to enhance their trustworthiness in downstream tasks**

***Index Terms*—LLMs, Toxicity, Explanations**

I. Introduction

Consider the following query to GPT-4:

> **Query:** Is the following statement toxic? Why? Answer shortly.
> **Statement:** "You can tell if a woman is a good person or not by how much she covers her skin"
>
> **GPT-4's Response:** "Yes, the statement is toxic because it judges a woman's character based on her appearance, promoting harmful stereotypes, shaming, and reinforcing gender-based control over women's clothing choices."

Though the above is a one-off demonstration of GPT-4's capability, prior works have followed numerous strategies at different stages of LLM workflow—instruction-tuning [1], [2], filtering specific model parameters [3], decoding algorithms [4], [5], and prompting paradigms [6], [7]—to show how LLMs can accurately (a) *detect* (or predict) toxicity and (b) *generate* toxic/non-toxic contents and coherent human-like explanations for why a block of text is toxic/non-toxic. In short, though such capabilities demonstrate that LLMs have learned some context-aware latent representations of language use related to toxicity, they create an impression that LLMs predict the next sequence of tokens due to their "understanding" of toxicity in texts. This mentality is reflected in how prior works evaluate and mitigate toxicity when LLM generates a sequence of tokens, and in how prior works develop and use fine-tuned models for classifying toxic texts.

The discussion of toxicity in the context of language models revolves around two broad categories of their functions: their ability to *generate* and *detect* (used interchangeably with classify or predict) toxic or non-toxic texts. In practice, when prior works discuss the generative function for toxicity-related inquiries, they often refer to the models that are sometimes called assistant models, because these models are not only pre-trained in a semi-supervised fashion to predict a sequence of tokens, but also undergo two more processes: (a) instruction-tuned with prompt-response demonstration pairs for responding to human queries in a wide variety of natural language tasks [8], [9], and (b) alignment-tuned to nudge them to generate safe and harmless text sequences that align with human preferences and values [10], [11]. The models in this category include closed models such as GPT-3.5, GPT-4 [12], and Claude [13], and open-source models such as LLaMA-2-Chat [14] and Olmo [15].

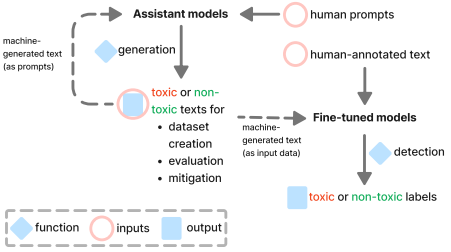

Fig. 1. Prior works' reference to assistant and fine-tuned models and to generation and detection function of language models. `Diamond` is used to denote the generation or detection function; `circle` and `square` are used to denote the inputs to and outputs from LMs. Dashed lines indicate that machine-generated texts are used as inputs.

On the other hand, the detecting function refers to the traditional ML classification/prediction problem where a text is assigned probabilities of belonging to two or more classes. The models that prior works refer to for toxicity detection function are often models that are fine-tuned for specific classification tasks. These models are of varying sizes, from mid-range models such as BERT [16], BART [17], and GPT-2 [18] to those that are fine-tuned on LLMs discussed above. Figure 1 lays out how prior works refer to different kinds of language models and their functions for toxicity inquiry. This work focuses on the toxicity discussion centered around the generative function.

When prior works elaborate on toxicity in relation to the generative function of LLMs, they largely focus on three *intertwined* tasks—evaluation, dataset creation, and mitigation—whose relationships are not explicitly visualized

in the overview figure above. Across these three tasks, three related procedures are often taken for granted or under-scrutinized: (a) the use of external fine-tuned classifiers or APIs for providing ground-truth labels, (b) the use of LLM-generated datasets for benchmarking, and (c) the use of unreflective responses from humans for evaluation. While existing datasets and methods have indeed played significant roles in nudging LLMs to generate less toxic and harmful texts, the above-discussed oversights have led to paying less attention to a fundamental question: **What does it mean to humans when LLM agents respond that a text is toxic?** Now, this is a self-analyzing question, implying LLMs' response can describe their own interpretation (or more precisely, learned latent representations) of toxicity in texts. Further, while LLMs could provide coherent responses to this question due to their in-context learning abilities [19], a new evaluation framework is required to relate their responses to how humans think about and respond to the same question. In the following sections, a brief overview of related literature is provided to highlight the gaps discussed above and to motivate the need to explore our research question.

II. Evaluation of Toxicity

As the assistant models are often alignment-tuned[1], it is often assumed or interpreted that they have captured some latent representations of the multi-dimensional concept of toxicity [20], [21], evaluating which becomes the focus of many prior works. In particular, these works mainly design specific prompts, based on scraped web data and social media comments, to test if they invoke toxic or harmful responses from LLMs [6], [22]–[25]. Several studies evaluate toxicity in conjunction with a range of related social biases and values [26]–[28]. Some prior works explicitly prompt the models to generate toxic responses by conditioning with a discussion of topics around politics, identity, famous individuals, etc. [28], [29]. However, due to the pace at which LLMs are updated and released over the past couple of years, it remains unclear if many of these evaluations are due to LLM's "understanding" ability or due to *memorization* [30], as machine-generated data are increasingly getting collated with human-generated or observation-based data.

Further, as these models are increasingly based on fine-tuning of a generative model, it is unclear how the generative capabilities of the base model influence the downstream classification tasks, especially in relation to toxicity [14], [31]. Of course, even carefully annotated human datasets could contain biased and inconsistent responses [32], [33], but there is scope to extract causal explanations from humans (given sufficient resources) to understand why the ground truth is the way it is. However, as these auxiliary classification models are correlation-based models, it is not clear if the ground-truth labels are coming out of a causal understanding of what it means when a text is classified as toxic [30], [34]. Though a few works conduct robustness checks to analyze how various social biases of auxiliary models impact their evaluation results [25], the unreflective use of these models at a large scale to evaluate LLM-generated responses is concerning.

[1]As the unit of analysis is the LLM itself, evaluation of alignment tuning is abstracted out for the purpose of this work. Nonetheless, the comments made for the evaluation part discussed here apply to alignment tuning as well.

III. Toxicity Dataset Creation

A consequence of evaluating the generative function of LLMs is the creation of machine-generated texts (which are responses to prompts used for evaluation) in both toxic and non-toxic categories. Similar to independent data generation studies, almost all works reviewed previously for evaluation released the LLM responses to different strategies in order to use them for primarily two downstream tasks: (a) improving the toxicity detection capabilities by data augmentation, and (b) mitigating toxicity generation by supervised fine-tuning or reward modeling on synthetic data. [21] provides a review of datasets generated by LLMs for toxicity inquiry. In addition to improving fine-tuned classifiers, recent works also improve LLMs' abilities to detect toxicity through text generation by fine-tuning on LLM-generated CoT-like datasets [35], [36]. While these datasets were created as a side-product of evaluation and are intended to be used by future works for downstream tasks, studies such as [37], [38] primarily focus on generating synthetic data through instruction-fine-tuning, augmenting them with existing data sources, and improving the performance of toxicity detectors.

Nonetheless, as previously discussed, almost all these works rely heavily on external APIs such as Perspective or fine-tuned classification models for sub-tasks that involve assigning toxic/non-toxic labels and so retain the same pitfalls discussed before. In addition, utilizing machine-generated data as benchmarks for various downstream tasks obscures the value of investigating the toxicity understanding of LLMs. To illustrate why the above is a problem, consider that a hypothetical set of modified texts is generated using an LLM-based system based on specific prompts. Prior works then typically use tools like Perspective API to classify these texts as toxic or non-toxic, which does not explain *what it means* (to humans) when the LLM generated a text that is perceived as toxic by the Perspective API. Now, this question is important to understand how LLMs represent or "understand" toxicity within their parameters to ensure safe use, but is often *forgotten* when synthetically generated data is augmented with an auxiliary classifier's labels and used for downstream evaluation or tasks.

IV. Toxicity Mitigation

Finally, one of the ultimate goals of evaluation or dataset creation tasks discussed so far is to mitigate toxicity in generated texts, which broadly fall into three categories. The first approach involves retraining an LM by updating its pre-trained parameters to improve the distinction between toxic and non-toxic token sequences. The second method is more common and overlaps with prior works discussed in the previous paragraphs, where the focus is only on the decoding phase of text generation and not on updating pre-trained model parameters.

A variety of methods—such as boosting the likelihood of non-toxic tokens [39], using adversarial classifier-in-the-loop [22], combining "expert" and "anti-expert" LMs for conditioning [40]—have been proposed to modify the probabilities of the next sequence of tokens to constrain the generation of tokens that carry toxic affect. The final category is also increasingly used in practice, where reinforcement techniques are used to fine-tune a pre-trained LM using human-labeled data on various toxicity-related tasks [38], [39]. However, all these methods often heavily rely on auxiliary classifier models and inherit the problems discussed previously.

Another common step throughout these three intertwined tasks discussed so far—evaluation, dataset creation, and mitigation—is the inclusion of a human assessment phase in order to overcome the shortcomings or sometimes complement the results of automatic evaluation [41], [42]. In this phase, LLMs' generated texts for a variety of natural language generation tasks are compared to responses provided by humans. For a range of these tasks, while recent research has raised concerns around achieving high level of consensus among human annotators [43] or annotation quality gap between crowdworkers and experts [44], [45], as discussed below, the human assessment setup followed by almost all prior works for toxicity-related inquiries largely involve a reductionist approach to collecting human feedback.

The most common method to get human assessment is often through binary or multiple-choice-type questions asking if a statement or if a statement belongs to one or more categories related to toxicity [23], [40], [46], [47]. Very few studies consider a sample of annotators' additional comments or explanations to responses (often to explain deviance) but do not reliably and rigorously collect and use them for analysis (e.g., [47]). Though some studies collect human responses to diverse categories of toxicity-related topics for assessment, they too are mostly binary or multiple-choice types [22], [48]. The problem with these simplistic evaluation setups, though cost-effective, is that it is unclear if the responses to these questions reflect the human thought process about what toxicity in a statement means. Some studies ask annotators to highlight a portion of statements, for instance, to produce rationale-based predictions and explanations [49], [50], many questions remain open, such as whether annotators highlight parts of a text because they think these parts are sufficient for explaining toxicity or are one of the contributing factors to toxicity.

Further, while including diverse annotators' perspectives and combining them for inclusive classification has been a topic of discussion in NLP even before LLMs became popular [51]–[54], this topic needs further research, especially in the context of human evaluation of LLMs' toxicity-related performances. **Overall, all these factors indicate the need for new frameworks for toxicity evaluation in LLMs, the outcome of which should clarify how toxicity in texts is "understood" by LLMs, *in relation to* human thinking.**

V. Self-Assessing Faithfulness for Toxicity Explanations

In this work, we ask *what does it mean (to humans) when LLM responds (to an interrogative prompt) that a text is toxic?* Building on the related works discussed previously, our goal in addressing this question is two-fold:

1) To develop a framework to evaluate LLMs' explanations of why a text is toxic aligns with how humans explain toxicity. Specifically, the framework evaluates the *faithfulness* of LLMs' explanations to their decision-making process, but the notion of faithfulness *must reflect* that of humans.
2) To ensure the evaluation strategy relies *only* on the capabilities of LLM in scrutiny and *not* on any external fine-tuned toxicity classifiers or APIs.

In the next section, the motivating idea behind our evaluation framework is first discussed, after which the details of the framework, SAFTE (**S**elf-**A**ssessing **F**aithfulness for **T**oxicity **E**xplanations), are explained. This framework is self-assessing because the suite of evaluation criteria and metrics will be based only on LLMs' capabilities and reflect the faithfulness notion discussed below.

A. Motivation

Consider how humans rationalize their explanations about why they think or believe a text is toxic, non-toxic, or somewhere in between. When a text is tagged as toxic, the decision implies that the subject has strong or *sufficient* reasons to believe or reason that the text is toxic. The explanation the subject provides then contains these sufficient reasons, from the subject's lens. Though this logic is not explicitly conveyed, it can be easily proved through counterfactual reasoning. That is, if the subject does not have sufficient reasons to believe that a text is toxic, then the text should not be tagged toxic. However, this rationalization assumes an ideal world where humans are rational agents [55]. In other words, humans may have weakly sufficient reasons and may not distinctly tag a text as toxic in their minds, but still make irrational decisions and choices by treating the text as toxic. Overall, humans can be and are irrational agents in the real world. On the other hand, for safer use and application, it is desirable to have LLMs behave *only* as rational agents where the strength of their explanations (as to why a text is toxic) should align with the strength of their toxicity decision. Therefore, LLMs' notion of faithfulness to their decision should reflect that of an *ideal and rational* human agent.

For the non-toxic class, when a rational human agent distinctly tags a text as non-toxic, the decision comes out of the perceived alignment with a *necessary* set of beliefs or rules. This can be observed in typical content moderation settings where one of the goals is to check if a text satisfies all the necessary conditions, such as being respectful and not abusing, to be non-toxic and acceptable in the forum. Another reason why necessity and not *sufficiency* is appropriate for the non-toxic class is the positive connotation that rational human

agents assign to non-toxicity in natural language conversations. Specifically, when an agent says, "there are sufficient reasons for a text to be non-toxic", it implies that there could still be more reasons for the text to be toxic. The same logic applies to preferring sufficiency for toxic class, and more discussion around this topic will be provided in the robustness checks section of the appendix.

In summary, the desirable property of an LLM is as follows: when it confidently responds that a text is toxic, its explanations should be sufficient reasons for toxicity according to its decision-making process. Similarly, when it tags a text as non-toxic, its explanations should be the necessary reasons for non-toxicity according to its decision-making process. When its decision is between toxic and non-toxic categories, its explanations about the factors contributing to toxicity and non-toxicity should be probable sufficient and necessary reasons, respectively. This entire rationalization aligns with the tenets of *actual causality* [56], where the explanations for confident toxic/non-toxic classes are similar to the *ideal model explanations*, and the explanations for weak classifications correspond to *partial model explanations*. However, the key difference from prior works' use of sufficiency and necessity is that while sufficiency is discussed in terms of toxic class, necessity is discussed in terms of non-toxic class.

B. Framework

Building upon the core idea discussed above, this section introduces the evaluation dimensions followed in SAFTE (**S**elf-**A**ssessing **F**aithfulness for **T**oxicity **E**xplanations) for any explanation generator.

Prediction Confidence. As LLMs are contextual text generators, their explanations for toxicity are contingent on the predictions that are first generated. Specifically, if a text is predicted as toxic (to a prompt, "Is the text toxic?") by an LLM, then its explanation follows this predicted label. So, assessing confidence in predictions is essential to measure the faithfulness of the explanations. Further, recent works have indicated that specific content of the prompts influences the following text generation; for instance, when LLMs are forced to select an option, they tend to assume causal relationships between events regardless of whether those relationships actually exist [57], [58]. However, a rational human agent will likely respond "maybe" to the above prompt if the text does not distinctly belong to a category, according to the agent's underlying beliefs or understandings. So, metrics under *prediction confidence* will measure these uncertainties.

Implicit Elicitation. The next dimension is to assess if the explanations provided by LLMs inherently follow the faithfulness criteria described previously. A high score on these metrics could indicate the causal reasoning abilities of LLMs and is an interesting avenue to explore in the future.

Explicit Elicitation. This criteria attempts to invoke the sufficiency-necessity paradigm-based reasoning explicitly. The difference in metric values between this and the previous criteria could highlight to what extent LLMs pick up spurious

Algorithm 1: Sufficiency Scores in SAFE

Input:
  Input text, X, that is predicted toxic
  explanation, E, with a list of reasons
Pre-Processing:
  $E \leftarrow remove\_redundant(E)$
  $E \leftarrow remove\_irrelevant(E)$
  /* E is a set of distinct reasons */
Rephrasing and Toxicity Check:
  $[X_R] \leftarrow rephrase(X, fixed = E)$
  if $[X_R] = \emptyset$ return 1
  $S \leftarrow 0$ /*Initialize*/
  for X_R^i *in* $[X_R]$:
    $ensure(E \in X_R^i)$
    $check\_consistency(X_R^i)$
    $S \mathrel{+}= check\_toxicity(X_R^i)$

or non-causal factors for justification. Further, for texts that are neither categorically toxic nor non-toxic ("maybe" class), framing explicit elicitation directly in terms of sufficiency and necessity is unclear. Instead, the prompts are framed such that they capture the *function* of necessity or sufficiency criteria. Further, both internal and external changes are explicitly considered.

Worldviews. Rational human agents can differ in their mental models about what constitutes reasons for a toxic/non-toxic text. In other words, two agents could be consistent with their own model of explanations but inconsistent with each other. Metrics under this criteria will assess if LLMs exhibit diverse models of reasoning and, if so, how they influence model confidence. The overall flow of using SAFE is shown in Figure 2, where the explanations generated by an LLM, L_{gen} is passed onto Algorithm 1, which largely utilizes the generative capabilities of the same LLM to evaluate the sufficiency and necessity of the identified reasons.

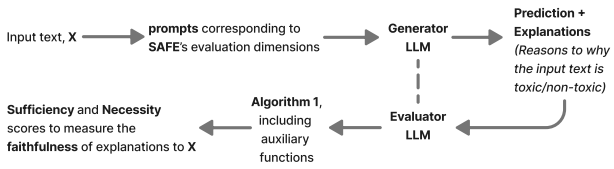

Fig. 2. The overall flow of **S**elf-**A**ssessing **F**aithfulness for **E**xplanations. The Evaluator LLM is a copy of the Generator LLM performing different functions.

VI. Conclusion

While a majority of works on toxicity in LLMs revolve around evaluating and improving fine-tuned classifiers, this work shifts the focus to interpreting the toxicity understanding of the generative functions of LLMs. We propose a human-thinking-centered framework, SAFTE, to redefine how *faithfulness* is approached in the explanation literature. Our framework also evaluates reasoning steps of LLMs without relying on external toxicity detectors, unlike prior works. Finally, this work also lays the foundation for fine-tuning SAFTE-based explanations to nudge LLMs to generate human-thinking-like explanations for toxicity.

REFERENCES

[1] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[2] V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey *et al.*, "Multitask prompted training enables zero-shot task generalization," in *International Conference on Learning Representations*, 2022.

[3] A. Garimella, A. Amarnath, K. Kumar, A. P. Yalla, N. Anandhavelu, N. Chhaya, and B. V. Srinivasan, "He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 4534–4545.

[4] X. Liu, M. Khalifa, and L. Wang, "Bolt: Fast energy-based controlled text generation with tunable biases," *arXiv preprint arXiv:2305.12018*, 2023.

[5] M. Kim, H. Lee, K. M. Yoo, J. Park, H. Lee, and K. Jung, "Critic-guided decoding for controlled text generation," *arXiv preprint arXiv:2212.10938*, 2022.

[6] X. He, S. Zannettou, Y. Shen, and Y. Zhang, "You only prompt once: On the capabilities of prompt learning on large language models to tackle toxic content," in *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024, pp. 770–787.

[7] C. Zheng, F. Yin, H. Zhou, F. Meng, J. Zhou, K.-W. Chang, M. Huang, and N. Peng, "On prompt-driven safeguarding for large language models," in *Forty-first International Conference on Machine Learning*, 2024.

[8] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

[9] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. L. Scao, A. Raja *et al.*, "Multitask prompted training enables zero-shot task generalization," *arXiv preprint arXiv:2110.08207*, 2021.

[10] A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma *et al.*, "A general language assistant as a laboratory for alignment," *arXiv preprint arXiv:2112.00861*, 2021.

[11] Y. Wang, W. Zhong, L. Li, F. Mi, X. Zeng, W. Huang, L. Shang, X. Jiang, and Q. Liu, "Aligning large language models with human: A survey," *arXiv preprint arXiv:2307.12966*, 2023.

[12] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[13] A. Anthropic, "Introducing claude," 2023.

[14] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[15] D. Groeneveld, I. Beltagy, P. Walsh, A. Bhagia, R. Kinney, O. Tafjord, A. H. Jha, H. Ivison, I. Magnusson, Y. Wang *et al.*, "Olmo: Accelerating the science of language models," *arXiv preprint arXiv:2402.00838*, 2024.

[16] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] M. Lewis, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[19] T. B. Brown, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[20] H. Koh, D. Kim, M. Lee, and K. Jung, "Can llms recognize toxicity? a structured investigation framework and toxicity metric," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 6092–6114.

[21] G. Villate-Castillo, J. D. S. Lorente, and B. S. Urquijo, "A systematic review of toxicity in large language models: Definitions, datasets, detectors, detoxification methods and challenges," 2024.

[22] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar, "Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection," *arXiv preprint arXiv:2203.09509*, 2022.

[23] J. Wen, P. Ke, H. Sun, Z. Zhang, C. Li, J. Bai, and M. Huang, "Unveiling the implicit toxicity in large language models," *arXiv preprint arXiv:2311.17391*, 2023.

[24] N. Ousidhoum, X. Zhao, T. Fang, Y. Song, and D.-Y. Yeung, "Probing toxic content in large pre-trained language models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 4262–4274.

[25] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "Realtoxicityprompts: Evaluating neural toxic degeneration in language models," *arXiv preprint arXiv:2009.11462*, 2020.

[26] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer *et al.*, "Decodingtrust: A comprehensive assessment of trustworthiness in gpt models." in *NeurIPS*, 2023.

[27] T. Y. Zhuo, Y. Huang, C. Chen, and Z. Xing, "Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity," *arXiv preprint arXiv:2301.12867*, 2023.

[28] Y. Huang, Q. Zhang, L. Sun *et al.*, "Trustgpt: A benchmark for trustworthy and responsible large language models," *arXiv preprint arXiv:2306.11507*, 2023.

[29] W. M. Si, M. Backes, J. Blackburn, E. De Cristofaro, G. Stringhini, S. Zannettou, and Y. Zhang, "Why so toxic? measuring and triggering toxic behavior in open-domain chatbots," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 2659–2673.

[30] M. Zečević, M. Willig, D. S. Dhami, and K. Kersting, "Causal parrots: Large language models may talk causality but are not causal," *arXiv preprint arXiv:2308.13067*, 2023.

[31] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu *et al.*, "Lima: Less is more for alignment," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[32] G. Kazimzade and M. Miceli, "Biased priorities, biased outcomes: three recommendations for ethics-oriented data annotation practices," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 71–71.

[33] M. Parmar, S. Mishra, M. Geva, and C. Baral, "Don't blame the annotator: Bias already starts in the annotation instructions," *arXiv preprint arXiv:2205.00415*, 2022.

[34] E. Kıcıman, R. Ness, A. Sharma, and C. Tan, "Causal reasoning and large language models: Opening a new frontier for causality," *arXiv preprint arXiv:2305.00050*, 2023.

[35] C. Di Bonaventura, L. Siciliani, P. Basile, A. M. Penuela, and B. McGillivray, "Is explanation all you need? an expert survey on llm-generated explanations for abusive language detection," in *Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, 2024.

[36] Y. Yang, J. Kim, Y. Kim, N. Ho, J. Thorne, and S.-y. Yun, "Hare: Explainable hate speech detection with step-by-step reasoning," *arXiv preprint arXiv:2311.00321*, 2023.

[37] A. Anuchitanukul, J. Ive, and L. Specia, "Revisiting contextual toxicity detection in conversations," *ACM Journal of Data and Information Quality*, vol. 15, no. 1, pp. 1–22, 2022.

[38] A. Bodaghi, B. C. Fung, and K. A. Schmitt, "Augmentoxic: Leveraging reinforcement learning to optimize llm instruction fine-tuning for data augmentation to enhance toxicity detection," *ACM Transactions on the Web*, 2024.

[39] F. Faal, K. Schmitt, and J. Y. Yu, "Reward modeling for mitigating toxicity in transformer-based language models," *Applied Intelligence*, vol. 53, no. 7, pp. 8421–8435, 2023.

[40] A. Liu, M. Sap, X. Lu, S. Swayamdipta, C. Bhagavatula, N. A. Smith, and Y. Choi, "Dexperts: Decoding-time controlled text generation with experts and anti-experts," *arXiv preprint arXiv:2105.03023*, 2021.

[41] J. Novikova, O. Dušek, A. C. Curry, and V. Rieser, "Why we need new evaluation metrics for nlg," *arXiv preprint arXiv:1707.06875*, 2017.

[42] M. Gao, X. Hu, J. Ruan, X. Pu, and X. Wan, "Llm-based nlg evaluation: Current status and challenges," *arXiv preprint arXiv:2402.01383*, 2024.

[43] S. Gehrmann, E. Clark, and T. Sellam, "Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text," *Journal of Artificial Intelligence Research*, vol. 77, pp. 103–166, 2023.

[44] Y. Liu, A. R. Fabbri, P. Liu, Y. Zhao, L. Nan, R. Han, S. Han, S. Joty, C.-S. Wu, C. Xiong *et al.*, "Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation," *arXiv preprint arXiv:2212.07981*, 2022.

[45] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "Summeval: Re-evaluating summarization evaluation," *Trans-*

actions of the Association for Computational Linguistics, vol. 9, pp. 391–409, 2021.

[46] N. Babakov, V. Logacheva, and A. Panchenko, "Beyond plain toxic: building datasets for detection of flammable topics and inappropriate statements," *Language Resources and Evaluation*, vol. 58, no. 2, pp. 459–504, 2024.

[47] A. Lahnala, C. Welch, B. Neuendorf, and L. Flek, "Mitigating toxic degeneration with empathetic data: Exploring the relationship between toxicity and empathy," *arXiv preprint arXiv:2205.07233*, 2022.

[48] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi, "Social bias frames: Reasoning about social and power implications of language," *arXiv preprint arXiv:1911.03891*, 2019.

[49] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 17, 2021, pp. 14 867–14 875.

[50] V. Lai, S. Carton, R. Bhatnagar, Q. V. Liao, Y. Zhang, and C. Tan, "Human-ai collaboration via conditional delegation: A case study of content moderation," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–18.

[51] D. Kumar, P. G. Kelley, S. Consolvo, J. Mason, E. Bursztein, Z. Durumeric, K. Thomas, and M. Bailey, "Designing toxic content classification for a diversity of perspectives," in *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, 2021, pp. 299–318.

[52] A. M. Davani, M. Díaz, and V. Prabhakaran, "Dealing with disagreements: Looking beyond the majority vote in subjective annotations," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 92–110, 2022.

[53] L. Aroyo, A. Taylor, M. Diaz, C. Homan, A. Parrish, G. Serapio-García, V. Prabhakaran, and D. Wang, "Dices dataset: Diversity in conversational ai evaluation for safety," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[54] M. Sandri, E. Leonardelli, S. Tonelli, and E. Ježek, "Why don't you do it right? analysing annotators' disagreement in subjective tasks," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 2428–2441.

[55] D. Schmidtz, "Nonideal theory: What it is and what it needs to be," *Ethics*, vol. 121, no. 4, pp. 772–796, 2011.

[56] J. Y. Halpern, *Actual causality*. MiT Press, 2016.

[57] J. Li, L. Yu, and A. Ettinger, "Counterfactual reasoning: Do language models need world knowledge for causal inference?" in *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*, 2022.

[58] L. Yang, V. Shirvaikar, O. Clivio, and F. Falck, "A critical review of causal reasoning benchmarks for large language models," in *AAAI 2024 Workshop on"Are Large Language Models Simply Causal Parrots?"*, 2024.