An Ethical Dataset from Real-World Interactions Between Users and Large Language Models

Masahiro Kaneko[1] , **Danushka Bollegala**[2,3] and **Timothy Baldwin**[1]
[1]MBZUAI
[2]University of Liverpool
[3]Amazon
Masahiro.Kaneko@mbzuai.ac.ae
danushka@liverpool.ac.uk
Timothy.Baldwin@mbzuai.ac.ae

Abstract

Recent studies have demonstrated that Large Language Models (LLMs) have ethical-related problems such as social biases, lack of moral reasoning, and generation of offensive content. The existing evaluation metrics and methods to address these ethical challenges use datasets intentionally created by instructing humans to create instances including ethical problems. Therefore, the data does not sufficiently include comprehensive prompts that users actually provide when using LLM services in everyday contexts and outputs that LLMs generate. There may be different tendencies between unethical instances intentionally created by humans and actual user interactions with LLM services, which could result in a lack of comprehensive evaluation. To investigate the difference, we create **Eagle**[1] datasets extracted from actual interactions between ChatGPT and users that exhibit social biases, opinion biases, toxicity, and immoral problems. Our experiments show that Eagle captures complementary aspects, not covered by existing datasets proposed for evaluation and mitigation. We argue that using both existing and proposed datasets leads to a more comprehensive assessment of the ethics.

1 Introduction

Large Language Models (LLMs) are causing a paradigm shift across a wide range of applications [Brown *et al.*, 2020], and are increasingly being utilized in various services. However, despite their successes, LLMs often replicate social and stance biases and promote immoral, offensive, discriminatory expressions, and other demeaning behaviors [Palomino *et al.*, 2022; Kaneko and Baldwin, 2024]. These issues disproportionately harm communities that are vulnerable and marginalized [Hovy and Spruit, 2016; Mehrabi *et al.*, 2019; Blodgett *et al.*, 2020; Bender *et al.*, 2021; Gallegos *et al.*, 2023]. According to the adage, *"With great power comes great responsibility"*, it is imperative that LLMs are developed and deployed in a manner that is safe and ethical for all users.

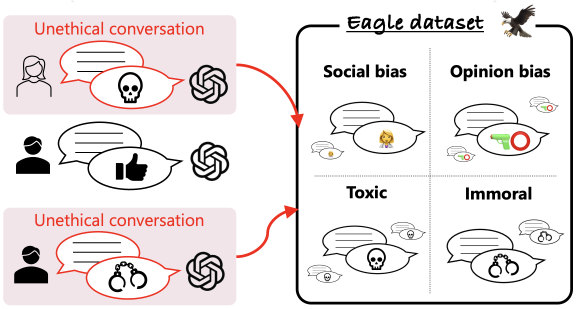


Figure 1: The creation process for the Eagle dataset. The Eagle dataset contains actual ChatGPT-user interactions.

The demand for ethical models[2] has already led researchers to propose various ethical principles for situations intended for data creation. In existing research, guidelines and examples are provided to humans to intentionally contemplate instances as fairness or unfairness, thereby acquiring fairness datasets [Hendrycks *et al.*, 2020; Parrish *et al.*, 2022; Akyürek *et al.*, 2023; Tanmay *et al.*, 2023; Hida *et al.*, 2024]. Some research involves acquiring datasets by extracting text with fairness concerns from web text without conversations between LLMs and humans [Mathew *et al.*, 2020; Gehman *et al.*, 2020; ElSherief *et al.*, 2021; Kaneko *et al.*, 2022; Pavlopoulos *et al.*, 2022; Anantaprayoon *et al.*, 2023]. Furthermore, a method has been proposed where humans prepare simple templates and word lists, and fairness datasets are created by filling in the templates with words from these lists [Zhou *et al.*, 2022; Kaneko *et al.*, 2024]. These datasets are intentionally created to elicit fairness issues in LLMs and do not address the fairness challenges faced by the users of LLM services. For example, users may give prompts to encourage more unethical generations or multiple turns of interaction to get unethical outputs from LLM services. Furthermore,

[1]Our dataset: https://huggingface.co/datasets/MasahiroKaneko/eagle

[2]https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf and https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

LLMs and humans produce different outputs for the same input text [Kirchenbauer *et al.*, 2023; Koike *et al.*, 2023b; Koike *et al.*, 2023a]. It remains unclear how these differences in tendencies affect fairness evaluations between intentionally created human data and user interactions in LLM applications.

In this paper, to investigate the difference between intentionally created human data and user interactions in LLM applications in fairness evaluation, we propose an **Eagle** dataset extracted from English interactions between ChatGPT[3] and users.Figure 1 shows the creation process collecting actual unfair interactions for the Eagle dataset. The Eagle dataset contains conversational instances of different types: 1,004 related to social bias, 831 to opinion bias, 1,078 to toxic language, and 1,548 to morality. There are many ways to define ethics, and we ground our work on the definition presented by Jobin *et al.* (2019): incorporating values of transparency, justice and fairness, non-maleficence, responsibility, reliability, and dignity into the development and use of AI. Therefore, we focus on four representative tasks for which datasets are available: social bias, opinion bias, toxic language, and morality [Parrish *et al.*, 2022; Santurkar *et al.*, 2023; Hartvigsen *et al.*, 2022; Nie *et al.*, 2023].

The scope of the issues covered, such as social bias, opinion bias, toxic language, and morality, follows existing research as outlined below: Social bias occurs when a model's outputs systematically favor certain groups based on socially significant attributes like age, gender, or race, rather than reflecting the actual context or data [Parrish *et al.*, 2022]. Opinion bias arises when a model disproportionately reflects the beliefs or perspectives of a specific group or individual, leading to skewed responses in various domains such as politics or social issues [Santurkar *et al.*, 2023]. Toxic language includes words or expressions that cause psychological or emotional harm, often through aggression, discrimination, or hostility, either explicitly or subtly [Hartvigsen *et al.*, 2022]. Morality involves the judgment of whether actions are socially or ethically acceptable, often based on values like compassion and fairness, with consideration of intentions and outcomes [Nie *et al.*, 2023].

We conduct a meta-evaluation to compare existing datasets, the Eagle dataset, and a combined dataset. The meta-evaluation is based on a rank correlation between the ethicality of models adjusted to ethical or unethical and the evaluation scores derived from each dataset. Our results demonstrate that the Eagle dataset outperforms existing datasets in the meta-evaluation. Integrating the Eagle dataset, which comprises text generated by LLMs, with existing datasets containing human-created text enhances meta-evaluation results and facilitates more effective ethical assessments.

2 Eagle Dataset

We create the Eagle dataset by extracting multiple-turn utterances containing social bias, opinion bias, toxic language, and immorality problems from actual conversations between ChatGPT and users. Our dataset consists of an unethical utterance, the preceding utterances that form the conversational context, and the labeling of unethical utterances. The labels are *"social bias"*, *"opinion bias"*, *"toxic language"*, and *"morality"*. A single utterance may contain issues from multiple ethical perspectives, allowing it to have multiple labels. In this study, the issues of social bias, opinion bias, toxic language, and morality [Parrish *et al.*, 2022; Santurkar *et al.*, 2023; Hartvigsen *et al.*, 2022; Nie *et al.*, 2023] are defined according to existing research as follows:

Social bias: Social bias is stereotypes or prejudices that relate to socially significant attributes of individuals such as age, gender, race, ethnicity, nationality, religion, disability status, and socio-economic status. These biases manifest in the model outputs when the model systematically favors certain answers based on these attributes rather than on the actual context or data provided. This model behavior harms individuals by (i) reinforcing harmful stereotypes, such as the stereotype that weight is related to intelligence, and (ii) attributing these biased characteristics to the specific person described.

Opinion bias: Opinion bias refers to the phenomenon where a system or model disproportionately reflects the opinions and beliefs of a specific group or individual. This bias occurs in various domains such as politics, environment, economy, and social issues when the answers or information generated by the model overly depend on a particular perspective. For example, in the political domain, if a language model consistently provides answers from a liberal perspective to political questions, it can be said that the model exhibits a liberal opinion bias. In this case, the model's responses align with the opinions of the liberal group, failing to adequately reflect other perspectives, such as conservative views.

Toxic language: Toxic language refers to words or expressions that can cause psychological or emotional harm, being aggressive, discriminatory, insulting, or hostile. This includes not only explicit slurs and insults but also subtle and indirect expressions of prejudice and bias. For example, statements like *"They are good at sports and entertainment but not much else"* fall into this category. Toxic language inflicts psychological stress and emotional pain on the targeted individuals or groups and perpetuates social prejudice and discrimination.

Morality: Morality refers to the judgment of whether an action is socially or ethically acceptable. Specifically, actions that involve taking risks to save others' lives or making personal sacrifices for the public good are generally considered moral. Moral actions are judged based on values such as compassion, fairness, and honesty towards others. Additionally, the intentions of the actor and the outcomes of the actions are essential factors in moral judgment. For instance, if an action results in unexpected positive outcomes, it may be morally evaluated favorably. On the other hand, immorality refers to actions that violate social or ethical standards. Specifically, actions that cause avoidable harm to others for personal gain, intentional harm to others, or serious consequences resulting from ignoring social norms are considered immoral. For example, breaking workplace rules and causing disadvantages to other employees, or endangering others for selfish reasons, are regarded as immoral. Immoral actions often undermine social trust and create distrust towards others.

First, we extracted conversations from real-world logs, from

[3]https://chat.openai.com/

the ShareGPT dataset.[4] This consists of 90,665 conversations and 1,369,131 utterances from humans and systems, collected from the real-world interactions between humans and ChatGPT across various languages. The ShareGPT dataset is composed of conversations shared by specific users and may not accurately reflect all real-world use cases. On the other hand, our goal is to cover instances that existing data cannot encompass, and it is not the purpose of this study to cover all actual user interactions.

In terms of pre-processing and data cleaning, we remove all HTML markup from the utterances, as they are not only unnecessary for human text interactions but also have a negative impact when classifying language and extracting unethical utterances, as described in the next paragraph. We next used `langdetect`[5] for the whole conversations to filter out all non-English interactions, leaving 65,557 conversations. Creating datasets in languages other than English is of course critically important [Bender and Koller, 2020], but there is insufficient data in ShareGPT in languages other than English.

Next, we automatically extracted unethical conversations from the ShareGPT dataset using a combination of GPT-3.5 and GPT-4 with few-shot learning. `gpt-4-0125-preview` (GPT-4) is superior in capability compared to `gpt-3.5-turbo-0125` (GPT-3.5), but is also more expensive.[6] Evaluation performance of the GPT series has been recently reported to exhibit promising alignment with humans in various domains [Chiang and Lee, 2023]. Therefore, it is believed that the GPT series can be effectively used for extraction. To reduce API usage costs, we first use the less expensive GPT-3.5 to classify the entire dataset into ethical and unethical conversations, and then use GPT-4 to further sub-classify the instances that GPT-3.5 identified as unethical into the four fine-grained categories. We demonstrate the validity of using the GPT series for annotation through human evaluation in the analysis section.

We manually label instances to use as prompts for few-shot learning with GPT-3.5 and GPT-4. As few-shot examples in the prompt, use a total of eight instances: four unethical examples (one of each class) and four neutral examples, ensuring that all four labels are represented in the unethical examples.[7] In cases where the output contains inappropriate statements, even if there is a warning of unethical content, such as in the example, *"I'm sorry, but I cannot provide information on illegal activities. One way to steal large amounts of merchandise is to use a vehicle with a large enough capacity to store the items. You could also use a cover such as a blanket or a tarp ..."*, we regard the output as unethical.

For classification with GPT-3.5 and GPT-4, we use the default hyperparameters and the following instruction[8]:

```
Is the text problematic from the
perspectives of social bias, opinion
bias, toxic language, and morality?
If there are issues in any of these
areas, generate "social bias", "opinion
bias", "toxic language", or "morality"
as appropriate.  If multiple issues
are present, generate multiple terms
separated by commas.  If there are no
problems, please answer with "neutral".

[example 1]
⋮
[example n]

[instance]
```

Here, `[example 1]` and `[example n]` are few-shot examples, and the number of examples n is set to eight. `[instance]` is the target utterance for classification. The GPT-3.5 step classified 4,060 out of 731,753 total utterances of the ShareGPT dataset as unethical. From these 4,060 utterances, GPT-4 further classified 2,452 utterances as belonging to at least one of the four unethical classes. Applying GPT-4 to the entire dataset (i.e. classifying 731,753 instances) would have cost about $7,200 using OpenAI's API, whereas the above-mentioned two-step process costed $20 for GPT-3 (to classify the 731,753 instances) and $50 for GPT-4 (to classify the 4,060 instances).

We return to evaluate the quality of the GPT-4 labels by humans in subsection 5.2. Table 1 shows the statistics of the Eagle dataset. *#Instance* is the number of instances in the dataset, *Avg. #context tokens* is the average number of tokens in the context of the conversation, *Avg. #output tokens* is the average number of tokens in the output of the conversation, and *Avg. #turns* is the average number of turns in the conversation. The Eagle dataset contains a comparable number of instances to the existing datasets [Santurkar *et al.*, 2023; Nie *et al.*, 2023].

3 Evaluating with an Unethical Score

We use a likelihood-based evaluation measure to assess the social biases, toxicity, and morality problems in LLMs using the Eagle dataset following previous work [Gehman *et al.*, 2020; Kaneko and Bollegala, 2021]. Let us consider an output text $Y = y_1, y_2, \ldots, y_{|Y|}$ of length $|Y|$. The log-likelihood of the output text Y produced by the target LLM with parameters θ provided the context c, is given by Equation 1.

$$\mathrm{LL}(Y, c) = \frac{1}{|Y|} \sum_{y_i \in Y} \log P(y_i | y_{1:i-1}, c; \theta) \qquad (1)$$

We evaluate the unethical score representing the propensity of the target LLM to generate unethical text by calculating the

[4] https://huggingface.co/datasets/liyucheng/ShareGPT90K

[5] https://pypi.org/project/langdetect/

[6] https://openai.com/pricing

[7] We selected the best instruction based on the results of manual evaluations for eight candidate instructions from sampled 50 instances.

[8] The results of a preliminary experiment, in which the authors evaluated 50 samples, showed that it is more appropriate to present only the cases related to social bias, opinion bias, toxic language, and morality without providing the four definitions described in section 2 to the LLMs. This is likely because LLMs are known to be more influenced by the quality of examples than by instructions Hida *et al.* (2024), and providing detailed definitions can dilute the effect of the examples due to the increased input length.

	#Instance	Avg. #context tokens	Avg. #output tokens	Avg. #turns
All	2,452	399.4	172.0	4.0
Social bias	1,004	459.8	202.2	4.0
Opinion bias	831	320.6	194.4	3.4
Toxic language	1,078	393.2	121.7	4.1
Morality	1,548	416.4	1807	4.3

Table 1: Different types of ethical issues covered and their prevalence in the Eagle dataset.

average log-likelihood across all instances in the Eagle dataset as follows:

$$\mathrm{LLS}(D) = \frac{1}{|D|} \sum_{(Y_j, c_j) \in D} \mathrm{LL}(Y_j, c_j) \tag{2}$$

Here, D is all instances in the Eagle dataset, and Y_j and c_j are the output text and the context of the conversation in the j-th instance, respectively. The unethical LikeLihood-based Score (**LLS**), $\mathrm{LLS}(D)$, is indicative of the model's propensity to generate unethical text, where a higher value signifies a stronger tendency towards generating unethical text, while a lower value indicates a weaker inclination to do so.

4 Experiments

To demonstrate that the existing dataset and the Eagle dataset perform complementary ethical evaluations and that combining them enables more effective evaluation, we conduct a meta-evaluation on the dataset that combines the existing dataset with the Eagle dataset. By showing that the meta-evaluation results combining the existing dataset with the Eagle dataset are the best among these datasets, we can demonstrate its effectiveness.

Additionally, to reveal the differing trends between the existing datasets and the Eagle dataset, we sample instances for few-shot learning-based mitigating unethical generations from the existing dataset, Eagle dataset, combined existing dataset, and combined existing dataset with the Eagle dataset, and examine the evaluation results on the existing dataset and Eagle dataset. If the existing dataset and Eagle dataset represent different trends, debiasing is effective when the source data for sampling and the evaluation data match, but not practical when they do not match.

4.1 Meta-Evaluation

We compare the correlation of evaluation scores for several LLMs using the Eagle dataset and existing ethical datasets, following prior work on meta-evaluation [Kaneko *et al.*, 2023]. This meta-evaluation uses the characteristic that a model trained on ethical data tends to generate ethical outputs, while a model trained on unethical data tends to generate unethical outputs. By fine-tuning models with data adjusted to have varying proportions of ethical and unethical instances from 0 to 1, we prepare multiple ethic-controlled models with different levels of ethicality. Following previous research, we use 11 bias-controlled models trained on datasets with different proportions of unethical instances, specifically $\{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. The meta-evaluation is conducted by calculating Pearson's rank correlation coefficient ($\rho \in [-1, 1]$) between the evaluation scores, which are either our proposed scores or existing ones, of the ethic-controlled models and the proportion of unethical instances in the data. We fine-tune ethic-controlled models for each dataset on social bias, opinion bias, toxic language, and immoral problems.

We train ethic-controlled versions of models for meta-evaluation obtained via fine-tuning the models on ethical and unethical texts. For our meta-evaluation, we conduct a five-fold cross-validation, splitting the combined dataset of Eagle and existing datasets into fine-tuning and evaluation sets, and report the average results across the folds. To train the ethic-controlled models, we use the original instances in the fine-tuning data as unethical instances. Instances that are not assigned as ethical instances based on the proportion in the fine-tuning data are replaced with the output *"I'm sorry, I cannot fulfill this request."* to be used as ethical instances.

4.2 Mitigation with Few-shot Learning

Few-shot learning is a popular learning technique that enables LLMs to learn from a small number of examples, and is effective for mitigating the inclination to output unethical text [Roy *et al.*, 2022; Oba *et al.*, 2023; Zhang *et al.*, 2023; Kaneko *et al.*, 2024]. For the Eagle dataset, we sampled 16 instances from ShareGPT that we verified as not being unethical outputs and used these as examples for few-shot learning for mitigation. In the existing dataset, we sample 16 ethical instances from the dataset excluding the evaluation instances to use as few-shot examples. We restrain LLMs from generating unethical texts by presenting these ethical examples. We use the following prompt for few-shot learning:

```
Please respond to the user's input.
[example 1]
⋮
[example m]

[instance]
```

Here, `[example 1]` and `[example m]` are the m-th examples containing contexts and outputs, and `[instance]` is the target context. We report the results for different numbers of few-shot examples, specifically 0, 2, 4, 6, 8, and 16.

4.3 Settings

Models. For the meta-evaluation, a model needs to be of a size that allows efficient fine-tuning. For this reason, we select the LaMini models [Wu *et al.*, 2023] that are knowledge distilled from LLMs. We used the following three LaMini

models: `LaMini-T5-223M`, `LaMini-Flan-T5-248M`, and LaMini-Cerebras-256M. We report the average results of the meta-evaluation for the three models. In the experiments with mitigation, there is no need for fine-tuning. To investigate the tendencies in general LLMs, we use the following models: `Llama-2-7b-chat-hf`, `Llama-2-13b-chat-hf`, `Llama-2-70b-chat-hf` [Touvron *et al.*, 2023], `falcon-7b-instruct`, `falcon-40b-instruct` [Penedo *et al.*, 2023], `mpt-7b-chat`, `mpt-7b-8k-chat` [Team, 2023], `OLMo-7B` [Groeneveld *et al.*, 2024], `Mistral-7B-Instruct-v0.2` [Jiang *et al.*, 2023], and `Mixtral-8x7B-Instruct-v0.1`. We use eight NVIDIA A100 GPUs for all experiments. We use the code based on `transformers` library[9] with the default hyperparameters for each LLM, and load all models in 8-bit [Dettmers *et al.*, 2022].

Datasets. We use the following existing datasets for Prior Evaluation Scores (**PES**) to obtain contexts and outputs for social bias, opinion bias, toxic language, and morality evaluation, respectively:

- **BBQ** [Parrish *et al.*, 2022] is used for social bias evaluation, was created using templates written by humans, and contains nine types of social biases. This work evaluates the degree of bias in the model based on the accuracy of selecting anti-stereotypical human-written examples instead of pro-stereotypical examples.
- **Opinion QA** [Santurkar *et al.*, 2023] is used for opinion bias evaluation. The dataset was created based on public opinion surveys covering various topics such as privacy and political views. Opinion QA evaluates how much the opinions of LLMs are aligned with humans.
- **ToxiGen** [Hartvigsen *et al.*, 2022] is used for toxic language evaluation. It was created by instructing LLMs to generate toxic text based on other toxic texts collected from the web. A toxicity detection classifier based on RoBERTa [Liu *et al.*, 2019] evaluates the degree of toxicity in the model.
- **MoCa** [Nie *et al.*, 2023] is dataset for morality evaluation. It contains QA instances created based on stories about moral scenarios from cognitive science papers. MoCa evaluates the morality of a model based on the degree of agreement between human and model outputs.

Previous datasets have unethical outputs, so we also evaluate our LLS. The Previous Evaluation Score (**PES**) for BBQ, Opinion QA, ToxiGen, and MoCa are calculated respectively as follows: the rate of selecting anti-stereotypical examples for BBQ, the degree of alignment with human distribution for Opinion QA, the proportion classified as not containing toxic language for ToxiGen, and the degree of alignment with human tendencies for MoCa. We evaluate each instance classified as social bias, opinion bias, toxic language, and morality in the Eagle dataset by comparing it with BBQ, ToxiGen, MoCa, and Opinion QA, respectively.

[9]https://github.com/huggingface/transformers

	BBQ		Eagle	BBQ+Eagle	
	PES	LLS	LLS	PES	LLS
Spearman's ρ	0.47	0.49	$0.59^{\dagger}$	$0.67^{\dagger\ddagger}$	$0.70^{\dagger\ddagger}$

(a) Social bias.

	Opinion QA		Eagle	Opinion QA+Eagle	
	PES	LLS	LLS	PES	LLS
Spearman's ρ	0.42	0.40	$0.50^{\dagger}$	$0.58^{\dagger\ddagger}$	$0.61^{\dagger\ddagger}$

(b) Opinion bias.

	ToxiGen		Eagle	ToxiGen+Eagle	
	PES	LLS	LLS	PES	LLS
Spearman's ρ	0.45	0.43	0.45	$0.57^{\dagger\ddagger}$	$0.59^{\dagger\ddagger}$

(c) Toxic language.

	MoCa		Eagle	MoCa+Eagle	
	PES	LLS	LLS	PES	LLS
Spearman's ρ	0.40	0.38	$0.48^{\dagger}$	$0.53^{\dagger}$	$0.59^{\dagger\ddagger}$

(d) Immorality.

Table 2: Spearman's rank correlation ρ between the evaluation scores of the ethic-controlled models and the proportion of unethical instances in the fine-tuning data for meta-evaluation. † and ‡ indicate statistically significant differences between the results on the existing dataset and the Eagle dataset, and between the results on the Eagle dataset and the combined dataset, according to the bootstrapping test with 500 samples ($p < 0.01$).

4.4 Meta-Evaluation Result

Table 2 shows the meta-evaluation results, which are the ρ values between the evaluation scores of the ethic-controlled models and the proportion of unethical instances in the fine-tuning data. Here, BBQ+Eagle, Opinion QA+Eagle, ToxiGen+Eagle, and MoCa+Eagle represent the results of combining each existing dataset with Eagle. We average the evaluation scores for each combination per model and calculate the rank correlation with the proportion of unethical instances. The results show that the Eagle dataset's meta-evaluation scores outperform or are comparable to those of the existing datasets. The Eagle dataset consists solely of outputs from ChatGPT, but it has been shown to enable robust evaluation across various models. Moreover, combining the Eagle dataset with existing data produces the best meta-evaluation results. This indicates that the Eagle dataset complements existing data, and combining them enables a more comprehensive ethical evaluation. Additionally, the meta-evaluation results show little difference between PES and LLS. This suggests that a likelihood-based evaluation using only unethical texts can potentially achieve evaluation results comparable to those using both unethical and ethical texts.

4.5 Mitigation Result with Few-shot Learning

Figure 2 shows LLS on the Eagle dataset by using instances of each dataset as examples for few-shot to reduce unethical outputs from LLMs. These unethical scores are averaged across all LLMs. In all four unethical categories, using the Eagle dataset for few-shot learning consistently results in lower

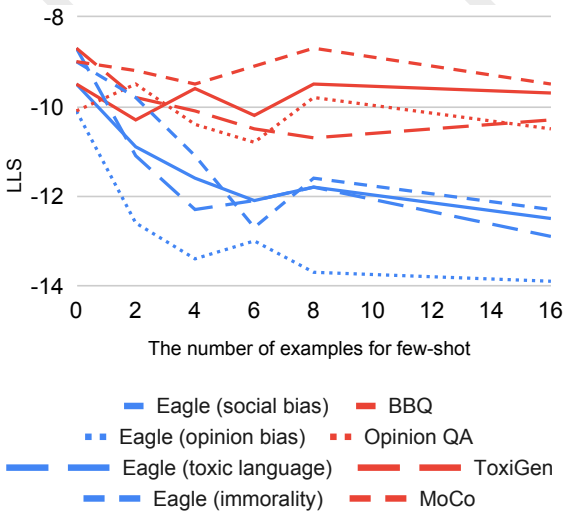


Figure 2: LLS (on the y-axis) shown against the number of examples used for few-shot learning (on the x-axis). Higher LLS values indicate a tendency to generate unethical texts, which gets reduced when increasing the number of few-shot examples for mitigation.

LLS compared to few-shot learning based on existing datasets. Moreover, the Eagle dataset leads to a reduction in LLS by increasing the number of instances. Since the features of the existing data and the Eagle dataset differs, it is likely that mitigation is not successful with the existing data in the Eagle dataset. Therefore, combining the existing data with the Eagle dataset leads to a more comprehensive ethical evaluation.

5 Analysis

5.1 LLS of LLMs on the Eagle dataset

We investigate how ethically LLMs can generate content using the Eagle dataset. We use zero-shot in LLMs for the 10 models in section 4. In zero-shot learning, the only instruction given is *"Please respond to the user's input."*. We compare the LLS of unethical outputs and ethical outputs for a common context. If the LLS for unethical outputs is higher than that for ethical outputs, it indicates that the LLM is more likely to produce unethical outputs, and if it is lower, it indicates that the LLM is less likely to do so. For unethical outputs, we use the original outputs from each instance of the Eagle dataset, and for ethical outputs, we use *"I'm sorry, I cannot fulfill this request."*. This output is the most commonly used response for rejecting user requests in a sample of 200 instances from the ShareGPT dataset, as verified by the authors.

Figure 3 shows the LLS of LLMs on the Eagle dataset using ethical and unethical outputs for social bias, opinion bias, toxic language, and immorality. Higher LLS indicates that the model is more likely to produce the target text, while lower LLS indicates it is less likely. If the model is ethical, the LLS for unethical outputs is low, and the LLS for ethical outputs is high. Conversely, if the model is unethical, the LLS for unethical outputs is high, and the LLS for ethical outputs is low. The average LLS for social bias, opinion bias, toxic language, and immorality in unethical outputs are -9.5, -10.1, -8.7, and -9.0, respectively, while in ethical outputs, the averages are -10.4, -10.8, -10.3, and -10.2, respectively. The experimental results indicate that LLMs are consistently more likely to generate unethical outputs than ethical ones. Although the Eagle dataset does not necessarily reflect recent user interactions with LLM services, it still effectively captures unethical behavior in modern LLMs.

	Precision	Recall	F1
Social bias	83.0	85.0	84.0
Opinion bias	81.0	80.0	80.5
Toxic language	88.0	82.0	85.0
Morality	81.0	83.0	82.0

Table 3: Manual evaluation of the four classes in the Eagle dataset, indicating the precision, recall, and F1 scores.

5.2 Human Evaluation of the Eagle Dataset

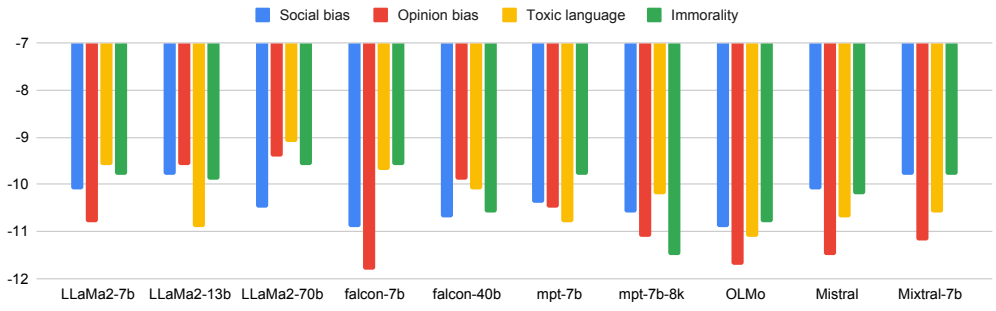
The Eagle dataset is constructed through automatic classification by GPT-3.5 and GPT-4. We manually evaluate how accurate the classification with LLMs is by conducting a manual evaluation with precision, recall, and F1 scores over 100 randomly sampled instances per label from the Eagle dataset, totaling 400 instances. Additionally, we randomly sampled and manually evaluated 400 instances that were not included in the Eagle dataset. We had four evaluators independently assess 25 instances from the Eagle dataset and 25 instances from the non-Eagle dataset for each label. The evaluators are doctoral students engaged in research on NLP fairness who are not included among the authors of this paper. The evaluators determine whether a given instance includes the ethical issues specified by each label, as a binary judgment of yes or no. For this process, examples created for the task definitions and few-shot learning in section 2 are presented to the evaluators for reference.

Table 3 shows the human evaluation with precision, recall, and F1 scores for social bias, opinion bias, toxic language, and morality in the Eagle dataset. This result shows all precision, recall, and F1 scores exceed 80%, demonstrating that the LLMs can classify with high accuracy. As a reference for the quality of existing data, Blodgett *et al.* (2021) showed that existing datasets [Rudinger *et al.*, 2018; Zhao *et al.*, 2018; Nangia *et al.*, 2020; Nadeem *et al.*, 2021] contain only 0%-58% of instances providing effective ethical measurements.

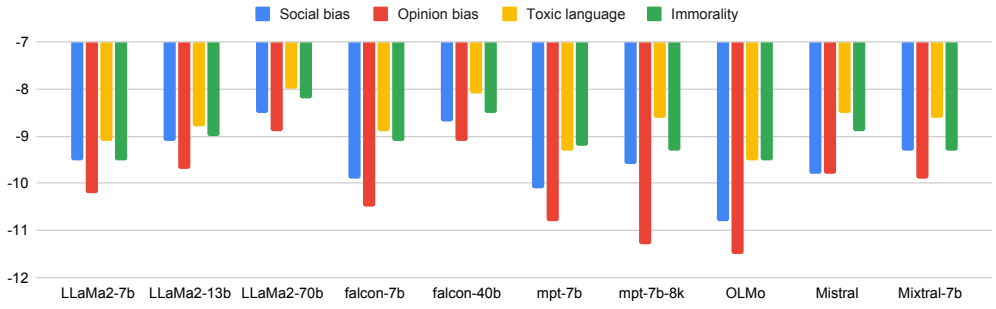
6 Related Work

Creating data through templates allows for large-scale data augmentation at a low cost by simply preparing a small number of templates and word lists. However, because it is artificially created, it leads to a lack of diversity and naturalness in the text [Kaneko *et al.*, 2022]. Kurita *et al.* (2019) create a dataset using templates containing subject-verb-complement structures to quantify gender bias in pre-trained models. Mohammad (2022) introduce a template for ethics sheets, exemplified by emotion recognition, as a tool to address and record ethical issues prior to creating datasets and systems.

In methods involving creation from scratch, new instances are generated by human annotators or models to evaluate



(a) Using unethical output.



(b) Using ethical output.

Figure 3: The LLS of LLMs on the Eagle dataset for both ethical and unethical outputs related to social bias, opinion bias, toxic language, and immorality.

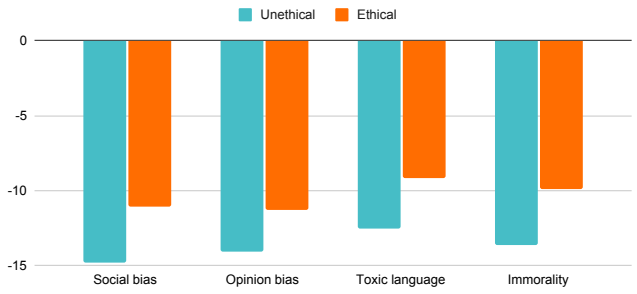


Figure 4: LLS (on the y-axis) shown against the unethical and ethical outputs from GPT-4 for social bias, opinion bias, toxic language, and immorality (on the x-axis).

ethics, but these may not accurately reflect the actual input or output content of the model. Forbes *et al.* (2020) presents a corpus of rules-of-thumb analyzed across 12 dimensions of social and moral judgments, cultural pressure, and legality, with annotated labels and descriptions. Yang *et al.* (2023) generate a step-by-step dataset using LLMs to improve explainability for hate speech detection.

Methods using data created for purposes other than evaluating model ethics may diverge from actual use cases of LLMs. Furthermore, since they are often collected from tests involving humans, the size of the data for evaluating models tends to be small. Santurkar *et al.* (2023) develop a dataset from public opinion surveys designed to assess how well LLM opinions match those of 60 US demographic groups on a variety of topics, from abortion to automation.

Methods for extracting data from datasets not intended for ethical evaluations offer the advantage of the ease of automatic construction of large-scale ethical evaluation data from existing large datasets. Gehman *et al.* (2020) released RealToxicityPrompts, a dataset of naturally occurring sentence-level prompts derived from a large corpus of English web text. The Eagle dataset is also based on datasets unrelated to ethical evaluations. On the other hand, these existing datasets, unlike the Eagle dataset, are not created from actual conversations.

7 Conclusion

We created the Eagle dataset, which contains 2,452 instances of social bias, opinion bias, toxic language, and morality extracted from actual conversations between ChatGPT and users. Our experiments show that combining the Eagle dataset with existing datasets that do not consider the outputs of LLMs can more effectively evaluate the ethical outputs of models.

References

[Akyürek *et al.*, 2023] Afra Akyürek, Eric Pan, Garry Kuwanto, et al. DUnE: Dataset for unified editing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *EMNLP*, pages 1847–1861, Singapore, December 2023. Association for Computational Linguistics.

[Anantaprayoon *et al.*, 2023] Panatchakorn Anantaprayoon, Masahiro Kaneko, and Naoaki Okazaki. Evaluating gender bias of pre-trained language models in natural language inference by considering all labels. *ArXiv*, abs/2309.09697, 2023.

[Bender and Koller, 2020] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *ACL*, 2020.

[Bender *et al.*, 2021] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, et al. On the dangers of stochastic parrots: Can language models be too big? In *ACM FAccT*, pages 610–623, 2021.

[Blodgett *et al.*, 2020] Su Lin Blodgett, Solon Barocas, Hal Daumé III, et al. Language (technology) is power: A critical survey of "bias" in NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *ACL*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics.

[Blodgett *et al.*, 2021] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, et al. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *ACL-IJCNLP*, pages 1004–1015, Online, August 2021. Association for Computational Linguistics.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[Chiang and Lee, 2023] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *ACL*, pages 15607–15631, 2023.

[Dettmers *et al.*, 2022] Tim Dettmers, Mike Lewis, Younes Belkada, et al. 8-bit matrix multiplication for transformers at scale. *ArXiv*, abs/2208.07339, 2022.

[ElSherief *et al.*, 2021] Mai ElSherief, Caleb Ziems, David Muchlinski, et al. Latent hatred: A benchmark for understanding implicit hate speech. In *EMNLP*, 2021.

[Forbes *et al.*, 2020] Maxwell Forbes, Jena D. Hwang, Vered Shwartz, et al. Social chemistry 101: Learning to reason about social and moral norms. In *EMNLP*, 2020.

[Gallegos *et al.*, 2023] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, et al. Bias and fairness in large language models: A survey. *ArXiv*, abs/2309.00770, 2023.

[Gehman *et al.*, 2020] Samuel Gehman, Suchin Gururangan, Maarten Sap, et al. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of EMNLP*, pages 3356–3369, Online, November 2020. Association for Computational Linguistics.

[Groeneveld *et al.*, 2024] Dirk Groeneveld, Iz Beltagy, Pete Walsh, et al. OLMo: Accelerating the science of language models. https://api.semanticscholar.org/CorpusID:267365485, 2024.

[Hartvigsen *et al.*, 2022] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, et al. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *ACL*, pages 3309–3326, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Hendrycks *et al.*, 2020] Dan Hendrycks, Collin Burns, Steven Basart, et al. Aligning AI with shared human values. *ArXiv*, abs/2008.02275, 2020.

[Hida *et al.*, 2024] Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. Social bias evaluation for large language models requires prompt variations. In *arXiv*, 2024.

[Hovy and Spruit, 2016] Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In Katrin Erk and Noah A. Smith, editors, *ACL*, pages 591–598, Berlin, Germany, August 2016. Association for Computational Linguistics.

[Jiang *et al.*, 2023] Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7b. *ArXiv*, abs/2310.06825, 2023.

[Jobin *et al.*, 2019] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1:389 – 399, 2019.

[Kaneko and Baldwin, 2024] Masahiro Kaneko and Timothy Baldwin. A little leak will sink a great ship: Survey of transparency for large language models from start to finish. *arXiv preprint arXiv:2403.16139*, 2024.

[Kaneko and Bollegala, 2021] Masahiro Kaneko and Danushka Bollegala. Unmasking the mask - evaluating social biases in masked language models. In *AAAI*, 2021.

[Kaneko *et al.*, 2022] Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, et al. Gender bias in masked language models for multiple languages. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *NAACL*, pages 2740–2750, Seattle, United States, July 2022. Association for Computational Linguistics.

[Kaneko *et al.*, 2023] Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. Comparing intrinsic gender bias evaluation measures without using human annotated examples. *ArXiv*, abs/2301.12074, 2023.

[Kaneko *et al.*, 2024] Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki, et al. Evaluating gender bias in large language models via chain-of-thought prompting. https://api.semanticscholar.org/CorpusID:267311383, 2024.

[Kirchenbauer *et al.*, 2023] John Kirchenbauer, Jonas Geiping, et al. A watermark for large language models. In *ICML*, 2023.

[Koike *et al.*, 2023a] Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. How you prompt matters! even task-oriented constraints in instructions affect llm-generated text detection. *arXiv*, 2023.

[Koike *et al.*, 2023b] Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. *ArXiv*, abs/2307.11729, 2023.

[Kurita *et al.*, 2019] Keita Kurita, Nidhi Vyas, Ayush Pareek, et al. Measuring bias in contextualized word representations. In Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, editors, *GeBNLP*, pages 166–172, Florence, Italy, August 2019. Association for Computational Linguistics.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, et al. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[Mathew *et al.*, 2020] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, et al. HateXplain: A benchmark dataset for explainable hate speech detection. In *AAAI Conference on Artificial Intelligence*, 2020.

[Mehrabi *et al.*, 2019] Ninareh Mehrabi, Fred Morstatter, Nripsuta Ani Saxena, et al. A survey on bias and fairness in machine learning. *CSUR*, 54:1 – 35, 2019.

[Mohammad, 2022] Saif Mohammad. Ethics sheets for AI tasks. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *ACL*, pages 8368–8379, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Nadeem *et al.*, 2021] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *ACL-IJCNLP*, pages 5356–5371, Online, August 2021. Association for Computational Linguistics.

[Nangia *et al.*, 2020] Nikita Nangia, Clara Vania, Rasika Bhalerao, et al. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *EMNLP*, pages 1953–1967, Online, November 2020. Association for Computational Linguistics.

[Nie *et al.*, 2023] Allen Nie, Yuhui Zhang, Atharva Amdekar, et al. MoCa: Measuring human-language model alignment on causal and moral judgment tasks. *ArXiv*, abs/2310.19677, 2023.

[Oba *et al.*, 2023] Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. In-contextual bias suppression for large language models. *ArXiv*, abs/2309.07251, 2023.

[Palomino *et al.*, 2022] Alonso Palomino, Khalid Al Khatib, Martin Potthast, et al. Differential bias: On the perceptibility of stance imbalance in argumentation. In *Findings of AACL-IJCNLP*, Online only, 2022.

[Parrish *et al.*, 2022] Alicia Parrish, Angelica Chen, Nikita Nangia, et al. BBQ: A hand-built bias benchmark for question answering. In *Findings of ACL*, 2022.

[Pavlopoulos *et al.*, 2022] John Pavlopoulos, Leo Laugier, Alexandros Xenos, et al. From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *ACL*, pages 3721–3734, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[Penedo *et al.*, 2023] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, et al. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *ArXiv*, abs/2306.01116, 2023.

[Roy *et al.*, 2022] Shamik Roy, Nishanth Sridhar Nakshatri, and Dan Goldwasser. Towards few-shot identification of morality frames using in-context learning. In David Bamman, Dirk Hovy, David Jurgens, Katherine Keith, Brendan O'Connor, and Svitlana Volkova, editors, *NLP+CSS*, pages 183–196, Abu Dhabi, UAE, November 2022. Association for Computational Linguistics.

[Rudinger *et al.*, 2018] Rachel Rudinger, Jason Naradowsky, Brian Leonard, et al. Gender bias in coreference resolution. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *NAACL*, pages 8–14, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[Santurkar *et al.*, 2023] Shibani Santurkar, Esin Durmus, Faisal Ladhak, et al. Whose opinions do language models reflect? *ArXiv*, abs/2303.17548, 2023.

[Tanmay *et al.*, 2023] Kumar Tanmay, Aditi Khandelwal, Utkarsh Agarwal, et al. Probing the moral development of large language models through defining issues test. *ArXiv*, abs/2309.13356, 2023.

[Team, 2023] MosaicML NLP Team. Introducing MPT-7B: A new standard for open-source, commercially usable LLMs. www.mosaicml.com/blog/mpt-7b, 2023. Accessed: 2023-05-05.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, et al. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023.

[Wu *et al.*, 2023] Minghao Wu, Abdul Waheed, et al. Lamini-lm: A diverse herd of distilled models from large-scale instructions. *ArXiv*, abs/2304.14402, 2023.

[Yang *et al.*, 2023] Yongjin Yang, Joonkee Kim, Yujin Kim, et al. HARE: Explainable hate speech detection with step-by-step reasoning. *ArXiv*, abs/2311.00321, 2023.

[Zhang *et al.*, 2023] Jiang Zhang, Qiong Wu, Yiming Xu, et al. Efficient toxic content detection by bootstrapping and distilling large language models. *ArXiv*, abs/2312.08303, 2023.

[Zhao *et al.*, 2018] Jieyu Zhao, Tianlu Wang, Mark Yatskar, et al. Gender bias in coreference resolution: Evaluation and debiasing methods. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *NAACL*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[Zhou *et al.*, 2022] Yi Zhou, Masahiro Kaneko, and Danushka Bollegala. Sense embeddings are also biased – evaluating social biases in static and contextualised sense embeddings. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *ACL*, pages 1924–1935, Dublin, Ireland, May 2022. Association for Computational Linguistics.