Enabling Classifiers to Make Judgements Explicitly Aligned with Human Values

Yejin Bang[1*] **Tiezheng Yu**[1*] **Andrea Madotto**[2]
Zhaojiang Lin[2] **Mona Diab**[2] **Pascale Fung**[1†]
[1]The Hong Kong University of Science and Technology [2]Meta AI
{yjbang,tyuah}@connect.ust.hk

Abstract

Many NLP classification tasks, such as sexism/racism detection or toxicity detection, are based on human values. Yet, human values can vary under diverse cultural conditions. Therefore, we introduce a framework for value-aligned classification that performs prediction based on explicitly written human values in the command. Along with the task, we propose a practical approach that distills value-aligned knowledge from large-scale language models (LLMs) to construct value-aligned classifiers in two steps. First, we generate value-aligned training data from LLMs by prompt-based few-shot learning. Next, we fine-tune smaller classification models with the generated data for the task. Empirical results show that our VA-MODELs surpass multiple baselines by at least 15.56% on the F1-score, including few-shot learning with OPT-175B and existing text augmentation methods. We suggest that using classifiers with explicit human value input improves both inclusivity & explainability in AI.

1 Introduction

The demand for responsible NLP technology – to make it more robust, inclusive and fair, as well as more explainable and trustworthy – has increased since pre-trained large-scale language models (LLMs) have brought about significant progress in making NLP tasks more efficient and broad-ranging (Brown et al., 2020; Zhang et al., 2022; Chowdhery et al., 2022; Radford et al., 2019; Brown et al., 2020; Petroni et al., 2019; Madotto et al., 2020). Researchers have studied how to align machines with human values as one of the directions to achieve responsible AI technology by teaching machines about moral and social norms (Forbes et al., 2020; Emelin et al., 2020; Jiang et al., 2021), ethics and common human values (Hendrycks et al., 2020) or human preferences (Christiano et al., 2017; Koren, 2008).

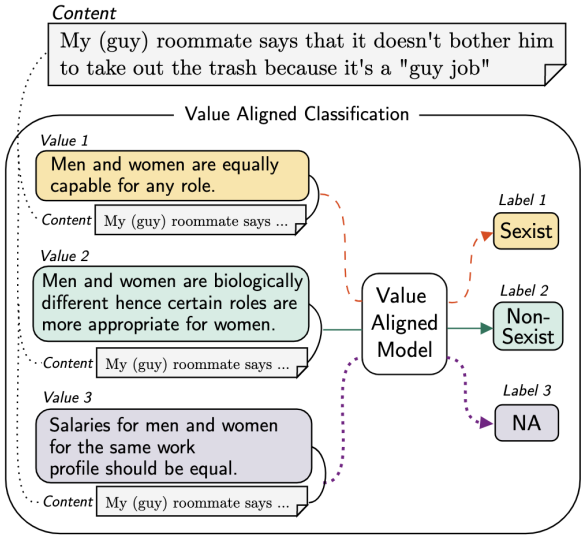


Figure 1: Illustration of proposed value alignment task. Given the same content, VA-MODEL makes variable predictions based on explicitly provided human values.

Value-alignment of AI systems is not a trivial problem as human values are non-consensual by nature (Hanel et al., 2018). Values can be very diverse and most existing works have attempted to align machines with shared human values or average norms, or from a certain cultural perspective with crowd sourced annotations (Jiang et al., 2021). These days, for instance, many societies agree that sexism should be eliminated, and we expect machines to be non-sexist, but different individuals and cultures may perceive sexism differently. As is shown in Figure 1, the same content can be considered to be sexist or non-sexist depending on the values provided to make the judgements.

In this paper, we propose a value-aligned judgement task that separates the value definition process from the development of the models for more inclusive and explainable value-aligned NLP. Our proposed task aims to build a single model to make

[*] Equal contribution.
[†] The author contributed to the original idea as a part of responsible AI project for Meta AI.

dynamic judgements based on explicitly provided human values, requiring the model to understand the value and its corresponding entailment on the given content. The value is provided in the form of instructions, allowing coarse-to-fine customization. We start with value-aligned sexism classification as a proof of concept for the proposed approach, as sexism is one of the most representative examples of varying cultural perspectives.

We also present Value-Aligned Models (VA-MODELS) that leverage value-based knowledge from LLMs. LLMs are trained from vast amounts of human data with embedded human values (Hendrycks et al., 2020). However, they are not controllable and it is difficult to fine-tune such large models with explicit value alignment. Instead, we distill value-based training data from the LLMs using prompt-based data generation with example values, and build VA-MODEL by fine-tuning smaller classification models with the distilled data. Experimental results show that our approach is more stable and accurate than directly applying few-shot learning on LLMs. Moreover, our methodology avoids costly human labeling or crowd-sourcing of values, allowing easier extensions to other value-aligned tasks in different domains. We further investigate model performance using data generated from different scales and types of LLMs, and study the effect of data size for fine-tuning, and analyze the quality of the generated data. Moreover, we study the generalization ability of VA-MODELS by testing its performance on unseen value sets.

Our contributions are as follows: 1) we introduce the value-aligned classification task, where we first define human values externally and then use them at the instruction level in an in-context learning paradigm and construct value-aligned classifiers to make predictions; 2) we propose to leverage prompt-based data generation to distill value-aligned knowledge from LLMs for smaller classification models; 3) experimental results indicate that our approach significantly outperforms strong baselines, including in-context few-shot learning with LLMs and existing text augmentation methods; 4) we systematically study factors that impact prompt-based data generation and highlight research questions and challenges in the value-aligned judgement task through thorough analysis.

2 Related Work

Human Value Alignment One challenge in value alignment is value definition, and there has been a profusion of documents on AI ethical standards (Gabriel, 2020; Dignum, 2017). Jobin et al. (2019) identified 11 clusters of ethical principles among 84 documents, and Fjeld et al. (2020) found eight key themes across 36 of the most influential of them. However, since human values are variable with culture, we anticipate value definition to be dynamic. Meanwhile, the values should be defined externally to the development of the NLP algorithms, like how we adopt definitions of sexism categories based on social studies.

To teach models value-alignment, the literature has focused on improving the model's reasoning ability relating to human values and morality (Forbes et al., 2020; Emelin et al., 2020; Lourie et al., 2021; Hendrycks et al., 2020). Recently, Solaiman and Dennison (2021) proposed to fine-tune GPT-3 to adapt to a manually crafted values-targeted dataset to arrive at a values-targeted model. However, in their approach, value alignment and definition are intertwined and entangled in an iterative process. We instead separate the value definition and alignment process models about value-aligned judgement with explicit value provision.

Prompt-based Learning Recently, LLMs have shown great performance on prompt-based learning (Brown et al., 2020; Chowdhery et al., 2022), which doesn't require fine-tuning. Instead, the model is directly fed a prompt that includes some examples, and the model can generate results as if it has "learned". Studies on efficient prompt-learning/-construction include Lu et al. (2021); Reynolds and McDonell (2021); Zhao et al. (2021); Schick and Schütze (2020). We consider the literature for prompt-construction in our methodology.

Knowledge Distillation Knowledge distillation is the transfer of knowledge from teacher to student distribution (Hinton et al., 2015). Recent works have attempted to perform distillation from LLMs by prompting for text generation to show that it outperforms existing text augmentation methods (Yoo et al., 2021; Wang et al., 2022). West et al. (2021) retrieves commonsense knowledge symbolically in a text form from GPT-3 for downstream tasks with help of smaller filtering classifiers. We distill *value-specific* knowledge, not all abilities of general language model, from LLMs through value-

aligned training data generation for training smaller value-aligned classifiers. This reduces the cost of human labeling and enables building smaller models specialized for value-aligned judgment task.

3 Value-Aligned Judgement Task

3.1 Task Description

As an effort to align machines with human values, our task focuses on teaching the model that different values can lead to different judgements even given the same content. The task is formulated as follows. A model needs to make a judgement Y_V on content C based on an explicit human value V. In this work, "value" refers to any qualities, standards of behavior, or beliefs that individuals or societies hold, and is expressed in natural language phrases or sentences. The set of values is externally defined by a human user of the system or from existing relevant literature on moral philosophy, and is independent of the development of algorithms. The distinction from the existing value-aligned classification task and conventional classification tasks is that our task expects the model to incorporate *explicitly provided values* along with other inputs for making judgements.

We separate the process of value definition from the development of the value-aligned models so that the models can learn to make dynamic judgements based on external values. For instance, existing sexism classifiers implicitly learn a fixed set of definitions of sexism from labeled data, so the content will be judged based on these static values. Our task requires the model to predict dynamic labels depending on the different explicit values even when the content is the same.

3.2 Value-aligned Sexism Classification

We showcase the value-aligned judgement task with an application to sexism classification. The model needs to judge whether natural language content is sexist or non-sexist based on a given value V. If the value is not applicable or irrelevant, the model needs to predict that it is not applicable (NA). Our rationale for choosing the sexism classification task is that the definition of sexism has changed over time as values have evolved and altered and it still varies across cultures. Thus, we can verify the effect of varying values in a more evident manner in the sexism classification task. Furthermore, the importance of a fine-grained understanding of sexism has been emphasized (Jha and Mamidi, 2017; Sharifirad et al., 2018; Parikh et al., 2019). This aligns with our motivation for explicit value-aligned judgement. Lastly, values related to sexism are complicated, involving religious, cultural, and personal beliefs or values. We thus believe it is a task with enough complexity to act as a case study.

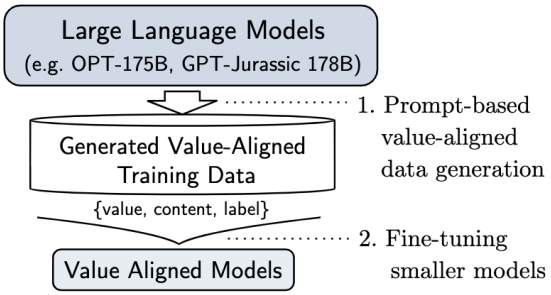


Figure 2: Illustration of the construction of our proposed VA-MODEL. Using LLMs, we first create synthetic value-aligned training data. Then, we transfer the knowledge into smaller models by fine-tuning them on the data, so Value Aligned Models can make value-aligned judgements.

4 Methodology

There is no existing resource for training value-aligned classification models. We therefore propose to leverage LLMs for generating synthetic training data. LLMs have been found to learn significant amounts of inherent knowledge as well as human values during pre-training (Petroni et al., 2019; Hendrycks et al., 2020; West et al., 2021; Roberts et al., 2020). However, the direct usage of LLMs in zero-shot setting for NLP tasks can be unstable and still limited (Wei et al., 2021). The richly embedded knowledge in LLMs nevertheless makes them good resource generators. Therefore, we attempt to build value-aligned models (VA-MODELs) through fine-tuning smaller models on the value-aligned training data generated by LLM(s).

Our proposed method (Figure 2) consists of two steps: 1) prompting human value-aligned contents from LLMs by providing explicit human values and instructions, and 2) fine-tuning smaller LMs on the generated data to teach them about value-aligned judgements. Formally, we build VA-MODEL (parameterized by θ) to maximize the following likelihood:

$$L(\theta) = logP(Y|V, C; \theta). \quad (1)$$

4.1 Value-Aligned Knowledge Distillation: Prompt-based Data Generation

Prompt Construction with Few-shot Examples The prompt construction of in-context few-shot examples affect performance. Thus we refer to the existing literature on different prompt-techniques (Reynolds and McDonell, 2021; Zhao et al., 2021; Yoo et al., 2021). For the few-shot examples, we create a pool of 10 human-labeled samples (value, content, and value-aligned labels) for each value. According to Lu et al. (2021), the order of the few-shot samples in the prompt affects the in-context learning for LLMs. Therefore, we randomly select and order five samples out of the pool.

To select the most appropriate prompt for generating value-aligned synthesized data, we test five candidate prompt templates with reference to literature. All prompt templates consist of a label, a value, and value-aligned content examples. The best-performing prompt template is selected based on testing with a smaller size of the samples. The prompt templates and their performance are available in Appendix A.

Generation We feed the prompts to LLMs to generate value-aligned synthetic training samples. Our method is model agnostic in that any LLMs can be adopted for this step. Recently, LLMs have scaled to more than 500 billion parameters (Chowdhery et al., 2022; Smith et al., 2022), and some models with more than 100 billion parameters are available publicly, such as Jurassic-1 Jumbo (GPT-Jurassic Lieber et al. (2021)), Open Pre-trained Transformer (OPT Zhang et al. (2022)), and GPT-3. In this paper, we choose OPT-175B for the main experiment and provide an analysis of the effects of the size and types of LLMs.

Generated Content Extraction & Processing The generated content is generated in succession after the prompt as natural text, and extracted through pattern matching. We gather all extracted content to construct a synthetic training set for teaching the smaller models in the next step, and process the generated data as follows. Firstly, we keep only unique samples by dropping all duplicates. Then, we remove exact copies of the few-shot examples used in the prompts. Finally, any content less than three words is filtered out as it is less informative.

4.2 Fine-tuning Smaller Models – Value-Aligned Models

In the next phase, we build classifiers by fine-tuning relatively smaller transformer-based models (e.g., ALBERT-base, RoBERTa-base, BART-base) with the generated training data to enable them to make value-aligned judgements. We add a linear layer on top of the pooled output of the smaller models to construct our proposed VA-MODEL. In order to make the model intake both values and content in the learning phase, the input text is formatted into "`value [sep] content [sep]`" and the output is a value-aligned judgement.

The classifiers need to predict different labels according to explicitly provided values given the same content. Recalling the example of value-aligned sexism classification in Figure 1, the same content can be considered to be sexist, non-sexist or NA depending on the considered values.

5 Experiments

In this paper, we conduct value-aligned sexism classification. Models are expected to label content with label choices sexist, non-sexist, NA *depending on* explicitly provided values.

5.1 Dataset

We borrow multi-label sexism categorization data (multi-sexism) (Parikh et al., 2019), which offers fine-grained sexism categorization for sexist content. Example categories include, but are not limited to, *Role-stereotyping*, *Pay gap*, and *Mansplaining*. We select 10 items of content per category to have a small set of human-labeled data for the prompt-construction in our methodology and baselines. The rest of the data are used as the test set.

Based on the description of each category, we manually compose two opposing values – one making the content sexist (value) and another making the content non-sexist (counter-value). For instance, any *Role Stereotyping* contents will be considered to be sexist based on the value "Men and women are equally capable for any role," but can also be considered to be non-sexist with the different value "Men and women are biologically different; hence certain roles are more appropriate for women." In total, we consider 19 categories of sexism and two corresponding values (value, counter-value) for each category, translated into 38 (19×2) human values.

Test set We use the original multi-label sexism content (human-labeled, non-synthetic) for creating a test set for the value-aligned judgement task, excluding that used for prompt-construction in the training data generation. Originally, each item of content is labelled with one/multiple sexism categories. For our task setup, we translate the data into the form of triplet {content, value, label}, and we assign value-dependent labels to each sample. For instance, if content C was originally labelled as *Role-stereotyping* (*RS*), we convert into three testing samples, {C, value$_{RS}$, Sexist}, {C, counter-value$_{RS}$, Non-Sexist}, and {C, random value/counter-value, NA}. Note that values for NA labels are totally unrelated to the content category. In this way, we can inspect the model's performance in making a value judgement on the same content with different values. In total, there are 17,720 test samples, with a label ratio of 1:1:1.

5.2 Models

5.2.1 VA-MODELS (Ours)

Generating value-aligned training data Using the method explained in Section 4.1, we get 100 content pieces from each of the value and counter-value prompts. In sum, there are 200 unique pieces of content per category.[1] Then, all content per category is paired with a value and counter value and corresponding labels {content, *value*, 'Sexist'} and {content, *counter-value*, 'Non-Sexist'}. So, each content item has a duplicate but is paired with different values and value-aligned judgements. To prevent the model from only learning two value and label associations, we synthetically make the class 'NA' by assigning irrelevant values/counter-values to the content (e.g., assigning the value of *Pay Gap* to a content of *Role Stereotyping* so the label is 'NA'). In total, there are 10,722 samples, including the prompt construction samples. We split them into training and validation sets with a ratio of 4:1.

Building VA-MODELS We finetune smaller models with the generated value-aligned training data. We build VA-MODELS to incorporate explicit human values to make judgements for value-aligned sexism classification following Section 4.2.

For the smaller models, we take base versions of ALBERT (12M params.) (Lan et al., 2019), RoBERTa (125M params.) (Liu et al., 2019) and BART (110M params.) (Lewis et al., 2019) to construct VA-ALBERT, VA-ROBERTA and VA-BART, respectively. RoBERTa has been proved to be robust in various NLP tasks and BART shows comparable performance to RoBERTa on GLUE tasks.

5.2.2 Baselines

To examine our proposed approach, we compare it with multiple baselines, including a random baseline, prompt-based few-shot learning with OPT-175B, and fine-tuning transformer-based models. For the fine-tuning setting, we fine-tune on different data setups – only with human-labeled data (without generated data) and with semantically augmented data.

Random Baseline We randomly select the predicted label for each test sample with the same label probability distribution as in the training data.

OPT-175B (few-shot) This baseline uses OPT-175B with a prompt-based few-shot learning for *label prediction*. [2] We provide 20 few-shot samples in the context.

Human-Labeled (HL)-Models We only use the small subset of human-labeled samples as training data to fine-tune smaller transformer-based LMs with a linear layer trained on top. We choose the base versions of ALBERT, RoBERTa and BART as the backbone models for a fair comparison with our VA-MODELS.

Nlpaug-Models Nlpaug (Ma, 2019) is semantic augmentation method using BERT-base embedding. We conduct augmentation with prompt construction examples by insertion and substitution. For each examples, we make 10 augmented samples (five insertions and five substitutions). Then, we fine-tune the base versions of ALBERT, BART and RoBERTa on the semantically augmented data and prompt-construction examples so we can evaluate the effectiveness of the prompt-based augmentation in our method.

5.3 Experimental setup

Evaluation metric We evaluate our experiments with both F1 score and accuracy. For the main results, we report all accuracy, weighted F1-score (W-F1), precision and recall.

[1] Reflecting the original ratio of multi-sexism, we keep the original number of samples if there are less than 100.

[2] We use prompt-based few-shot learning with OPT-175B for generating *value-aligned content* in our methodology while the baseline used it for directly predicting *label*. Refer to Appendix B for details.

Model	*Accuracy*	*W-F1*
Random Baseline	$33.53_{0.41}$	$33.53_{0.41}$
OPT-175B (few-shot)	$55.18_{7.75}$	$54.78_{7.20}$
HL-ALBERT	$58.70_{4.43}$	$51.67_{3.96}$
HL-RoBERTa	$64.53_{2.54}$	$55.23_{1.91}$
HL-BART	$63.23_{1.87}$	$54.93_{1.47}$
Nlpaug-ALBERT	$62.87_{2.13}$	$58.80_{3.44}$
Nlpaug-RoBERTa	$61.52_{3.03}$	$58.67_{2.89}$
Nlpaug-BART	$59.03_{1.38}$	$58.49_{1.60}$
VA-ALBERT	$70.10_{1.65}$	$70.75_{1.48}$
VA-RoBERTa	$73.24_{0.39}$	$73.82_{0.32}$
VA-BART	$\mathbf{74.07}_{0.82}$	$\mathbf{74.36}_{0.60}$

Table 1: Evaluation results of baselines and our proposed VA-MODELs, on the value-alignment task. We use 200 value-aligned training data samples generated from the LLMs per category to fine-tune VA-MODEL. Experiments are ran with five random seeds and results are reported in $mean_{std}$ format. All our VA-MODEL performances are statistically significant (t-test with p-value < 0.05). Scores are all in percentage (%).

Implementation Details For generating value-aligned training data, we conduct the main experiment with OPT-175B model with top-p 0.7 and temperature 1. For our VA-MODELs and HL-Models baselines, we use pre-trained transformer-based LMs available through the HuggingFace API. Further implementation details such as hyperparameters are given in Appendix B.

6 Results and Analysis

6.1 Main Results

Effectiveness of our method Table 1 shows the performance of the models on the value-aligned sexism classification task. Our models achieve better scores on W-F1 and accuracy than the baselines by large gaps ($15.56 \sim 40.83\%$ gain in W-F1), which signifies the robustness and superiority of our approach. Our VA-ALBERT also surpasses all baselines, including those back-boned with bigger models (e.g., Nlp-RoBERTa, HL-RoBERTa). This highlights the effectiveness of the value-aligned knowledge distillation with LLMs.

We observe that the OPT-175B few-shot learning approach performs better than random label assignments on the test set and HL-ALBERT, but still performs worse than or as comparable as the other baselines. This indicates that LLMs with prompt-based few-shot learning can understand the value-aligned classification task to some extent, but the performance is still low. HL-Models surpass OPT-175B (few-shot) under all evaluation metrics except HL-ALBERT in W-F1 score, showing that the models can capture our task with limited human-labeled data due to the effectiveness of fine-tuning. Nlpaug is one of the conventional data augmentation approaches and we augment the same amount of data as VA-MODEL. In comparison with HL-Models, Nlpaug-models show higher W-F1 scores with small drops in accuracy.

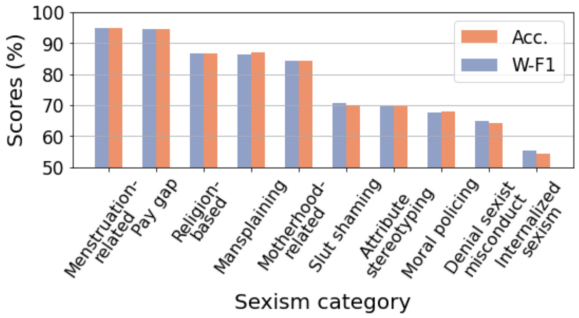


Figure 3: Evaluation results of VA-BART per sexism category on the test set. Only the top and bottom five categories (based on W-F1) are displayed. The performance for the nine categories in the middle are $\sim 80\%$ for Acc. and W-F1. The full results for the 19 categories are available in Appendix D.

Overall, the experimental results support our proposed approach for the value-aligned judgement task. OPT-175B (few-shot) shows much lower and unstable performance than VA-MODELs although the value-aligned training data of VA-MODELs is generated from OPT-175B. For the prompt-based few-shot approach, especially when the task setup is complicated like value-aligned classification, the model cannot easily overfit the task by giving several prompts, leading to a higher chance to predict random labels. Instead, we used a knowledge distillation approach through training data generation, which is a simpler task for the model as the main objective of the general language model is text generation. Moreover, utilizing the LLMs for generating knowledge distilled data is more effective than simple semantic text augmentation (e.g., Nlpaug).

Per-Category performance Figure 3 presents the per-category evaluation scores of VA-BART. The results vary significantly between categories, indicating the complexity of our proposed task. The results for both *Menstruation-related Discrimination* and *Pay Gap* achieve scores higher than 90%, while the results for *Internalized Sexism* are rela-

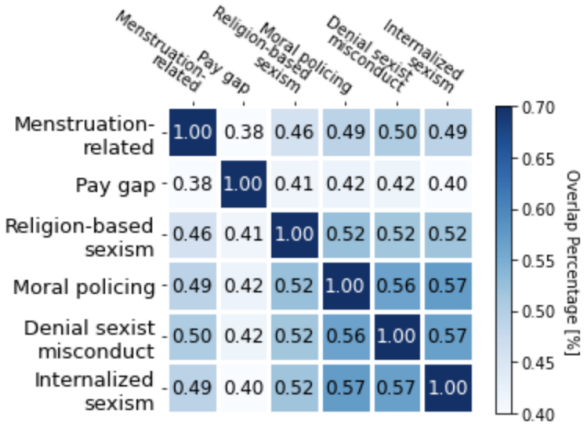


Figure 4: Vocabulary overlaps (%) of the generated data among sexism categories. Only top-3 and bottom-3 categories are displayed in descending order of W-F1 (top to bottom; left-to-right). Full set is in Appendix C.

tively low. We conjecture reasons for the high performance of certain categories are varying quality of generated training data per categories and more distinguishable features than other. We investigate this point further in Section 6.2.

6.2 Quality Analysis for Generated Value-Aligned Training Data

Distinction between generated data & test set The vocabulary overlap between all generated data (training set for VA-MODELs) and test set data is 51.79%. Moreover, we check how many of generated data samples that share more than 80% of vocabulary with at least one of the test data samples, finding that *only* 0.01% of generated data samples reach the threshold (80%). Therefore, the data generated from OPT-175B for training VA-MODELs is clearly distinct from the test set.

Diversity of Data We calculate the vocabulary overlaps for each sexism category of the generated data in Figure 4. We observe that the vocabulary overlaps are generally small, which illustrates that OPT-175B can generate diverse data for different values (e.g., sexism categories) provided in prompts. We can observe the trend that the overlaps among high performing categories are small, especially *Pay gap* and *Menstruation related*, which make data sample distinguishable to others. In contrast, low performing categories, overlaps are relatively higher.

Human evaluation LLMs are powerful few-shot learners, yet they are not perfect. Thus, we conducted a human evaluation on generated text from LLM with crowd-sourcing (Amazon Mechanical Turk) with 190 samples, 10 samples from each category. The workers were asked to judge the generated text given the value used for prompting the data. We gathered 3 annotations per sample and took the majority label for a final judgment. We assessed whether the generated text is value-aligned by checking if it gets the same annotation from the label used for generating the text. 63.68% of the tested generated texts are of good quality for value-aligned classification, meaning it aligns with the label given the value.

Model	*Accuracy*	*W-F1*
VA-ALBERT	$70.10_{1.65}$	$70.75_{1.48}$
w/o human labeled data	$70.79_{1.40}$	$71.29_{1.33}$
VA-RoBERTa	$73.24_{0.39}$	$73.82_{0.32}$
w/o human labeled data	$72.90_{2.06}$	$73.19_{1.68}$
VA-BART	$\mathbf{74.07}_{0.82}$	$\mathbf{74.36}_{0.60}$
w/o human labeled data	$72.30_{1.24}$	$72.71_{0.90}$

Table 2: Effectiveness of generated data. We remove human-labelled data from the training set and **only** use synthetic samples generated from LLM for training (w/o human-labeled data). The minimal drops in performance show the effectiveness of value-aligned training data generated from LLMs for the value alignment task.

Effectiveness of generated training data To investigate the standalone effectiveness of the generated training data (value-aligned knowledge distillation), we study the performance of VA-MODELs when they are trained *without* any of human-labeled data but **only** with generated data (Table 2). Minor performance degradations in both VA-BART and VA-RoBERTa are investigated, -1.65% and -0.63% W-F1 respectively. However, these values are still above those of the baselines. Interestingly, VA-ALBERT showed a minimal performance gain on both accuracy and W-F1. This indicates that the value alignment knowledge distilled from LLMs is the main contributor for VA-MODEL to understand the task.

6.3 Generalization Ability on Unseen Values

To understand capacity of models to generalize value-aligned judgement over unseen values, we conduct an experiment in which three randomly selected sexism categories are separated from the training process (i.e., models have never seen values related to the three categories in the training phase and are evaluated on test set only composed of those unseen values) and the results are pre-

Model	*Accuracy*	*W-F1*
OPT-175B (fewshot)	32.97	30.23
HL-ALBERT	$40.25_{6.05}$	$37.52_{7.68}$
HL-RoBERTa	$47.79_{5.65}$	$45.35_{6.09}$
HL-BART	$46.09_{3.42}$	$45.97_{3.68}$
Nlpaug-ALBERT	$48.10_{11.0}$	$40.62_{8.05}$
Nlpaug-RoBERTa	$40.14_{2.00}$	$30.64_{3.56}$
Nlpaug-BART	$47.76_{3.89}$	$42.40_{5.45}$
VA-ALBERT	$55.15_{6.83}$	$53.14_{9.05}$
VA-RoBERTa	$\mathbf{58.13}_{5.33}$	$\mathbf{56.60}_{6.56}$
VA-BART	$57.98_{5.12}$	$55.94_{5.48}$

Table 3: Performances of VA-MODELs and baselines on **unseen** values in value-aligned sexism classification. In the training phase, models did not see any of the values in the test set.

sented in Table 3. Overall, there are drops in performance compared to the main experiment (Table 1), while all of our VA-MODELs continue to outperform all baselines. The baselines experience larger drops (maximum 43.18% for Nlpaug-RoBERTa) than the VA-MODELs (17.22% for VA-RoBERTa). Considering the model was never taught or received any direct supervision on the test values, it is expected behavior as other generalization problem. We leave how to improve the models' generalization ability in value-aligned judgement task for future work.

6.4 Ablation Studies

LLMs capacity for prompting We first investigate how the size of LLMs affects the capacity for generating value-aligned training data by evaluating the final performance of VA-MODEL trained on data from varying sizes of LLMs. Unsurprisingly, as is shown in Table 4, we can continually boost the model's performance when the LLMs size increase. We also train VA-BART with the data prompted from GPT-Jurassic. Results for GPT-Jurassic 6B are slightly higher than those of OPT-6.7B, although the model size is smaller. However, when the LLMs become extremely large, GPT-Jurassic 178B performs similar to OPT-175B with only 0.12% difference. Since similar model sizes show similar performance with minimal differences, the types of LLMs do not have much effect on the generated data quality for our task.

Effect of the size of generated training data To investigate whether increasing the the number of generated data can gain further improvements, we fine-tune VA-MODELs with different training data size. In Figure 5, we show that the W-F1 score does not show any gain when the size exceeds 200 samples except for VA-ALBERT. As we analysed in Section 6.2, the generated data has noise. We conjecture that when using more generated data, the additional data will not only bring more value alignment knowledge, but also add more noise to the training set. Therefore, when the degradation in model performance caused by the noisy data is greater than the improvement in model performance from the additional knowledge, the overall results decrease.

Models	*Accuracy*	*W-F1*
VA-BART(OPT-1.3B)	$65.72_{2.03}$	$66.50_{2.15}$
VA-BART(OPT-6.7B)	$65.69_{2.28}$	$66.44_{2.45}$
VA-BART(OPT-175B)	$\mathbf{74.07}_{0.82}$	$\mathbf{74.36}_{0.60}$
VA-BART(GPT-Jurassic 6B)	$69.09_{1.59}$	$69.89_{1.49}$
VA-BART(GPT-Jurassic 17B)	$71.03_{1.01}$	$71.68_{0.90}$
VA-BART(GPT-Jurassic 178B)	$74.04_{1.16}$	$74.24_{0.91}$

Table 4: Effect of size and types of LLMs on value-aligned training data generation. We prompted OPT and GPT-Jurassic ranging 1.3B $\sim$ 178B. The bigger the model, the better the final performance in the value alignment task. All VA-BART variations are fine-tuned with the same number of training samples.

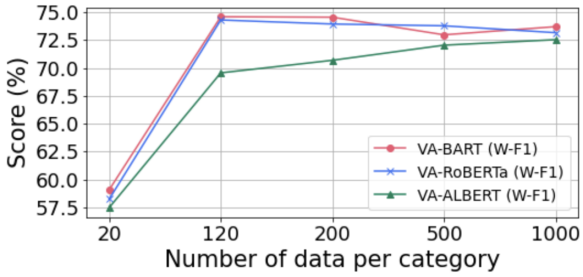


Figure 5: Evaluation results (W-F1) of VA-MODELs over different size of generated training data.

7 Discussion and Broader Impacts

In the present section, we provide a discussion on the broader impacts of our proposed framework, including its potential benefits and risks of misuse.

Benefits Our proposal entails the decoupling of the human value definition from the value-alignment task, with the former being defined through collaboration among society, ethicists, social scientists, and other relevant parties. Meanwhile, the mechanics of value alignment should remain independent of the initial value definition to

prevent engineers from directly defining values in the training data or code. This separation of tasks brings numerous benefits, such as the development of localized values that align with the values espoused by users and developers, accounting for cultural and historical differences. Additionally, it enables accountability for whoever provides the value definition, whether from human experts or digital sources.

Potential Risks The deployment of such models in real-world applications raises concerns regarding potential risks. One such concern is the lack of a base guardrail for determining appropriate values for alignment, which could lead to nefarious use of the technology. To mitigate this risk, future work could consider using a framework based on universal human rights as a value system.

While the classifier has not yet achieved deployment-level performance, it is still important to consider potential scenarios involving risk. In particular, two scenarios can be envisioned in terms of the users of the system. Firstly, the general public may provide their own values, some of which may be harmful. If such values are used to make judgments, it could potentially cause harm in human-computer interaction (Weidinger et al., 2021). To address this, a safety layer could be implemented to assess the appropriateness of input values, ensuring that only non-harmful values are used like the method of (Xu et al., 2020). Secondly, the classifier may be utilized by a system releaser, such as an organization or company, for classifying content based on certain values. In this case, the values used should be carefully determined in consultation with society, ethicists, social scientists, and compliance teams, rather than solely by engineers. By involving various stakeholders in the decision-making process, potential risks can be more effectively mitigated.

8 Conclusion and Future Work

In this paper, we propose a task that focuses on teaching a model human value alignment knowledge. We also introduce value-aligned models (VA-MODEL) that generate value-aligned training data from LLMs by prompt-based data generation and fine-tune smaller classification models with the value-aligned generated training data. Experimental results show that VA-MODEL generally outperforms strong baselines. Further analysis illustrates that the generated data from larger LLMs helps increase the performance, and more generated data can cause performance reduction when the data size is too large. Finally, we highlights several research challenges for future work: improvements in 1) the robustness of the model on diverse values, 2) the models' generalization ability for our value-aligned judgement task, 3) higher quality generated data with more human curation.

Limitations

Our methodology is currently tested with only English. We conjecture that the methodology should be applicable to other languages, but may be limited by the capacity of LLMs in those specific languages. It is possible that value-aligned knowledge distillation may be more difficult with languages from countries and regions that do not have a complete set of human value definitions. Thus, exploring the value-aligned task in different languages other than English is a promising research direction.

Our main experimental results are based on a 175B parameter model, which requires large GPU resources or access through an API. This may hinder other researchers from reproducing experimental results. Additionally, we explored different sizes of LLM including 1B and 6B models, which do not require large GPU resources, and showed they can achieve comparable results. We hope they can be possible alternative options for researchers who may not have access to 100B+ models. Although sexism is a suitable case study for us to investigate the feasibility of the value alignment task as we have shown throughout this work, it is still one domain. Further expansion to different domains or value-aligned classification tasks such as the detection of racism, toxicity, other than sexism, are needed.

Ethics Statement

Recently, many works study the ethical consideration of applying language models to applications. For example, Kumar et al. (2022) overviewed the strategies for detecting and ameliorating different kinds of risks/harms of language generators. Mohammad (2022) proposed an ethics sheets for AI tasks. In this work, we conduct experiments with some values that are unconventional and counter to the current contemporary society. However, we are not suggesting tolerance for sexist behaviors or beliefs. Instead, we explain the existence of differ-

ent perspectives in the discussion of sexism across cultures or religions. Our value-aligned sexism classification task is a case study of this decoupled process.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Virginia Dignum. 2017. Responsible artificial intelligence: designing ai for human values. Daffodil International University.

Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. 2020. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *arXiv preprint arXiv:2012.15738*.

Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Berkman Klein Center Research Publication*.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Iason Gabriel. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437.

Paul HP Hanel, Gregory R Maio, Ana KS Soares, Katia C Vione, Gabriel L de Holanda Coelho, Valdiney V Gouveia, Appasaheb C Patil, Shanmukh V Kamble, and Antony SR Manstead. 2018. Cross-cultural differences and similarities in human value instantiation. *Frontiers in Psychology*, 9:849.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*.

Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

Yehuda Koren. 2008. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 426–434, New York, NY, USA. Association for Computing Machinery.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2022. Language generation models can cause harm: So what can we do about it? an actionable survey. *arXiv preprint arXiv:2210.07700*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13470–13479.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Andrea Madotto, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Language models as few-shot learner for task-oriented dialogue systems. *arXiv preprint arXiv:2008.06239*.

Saif Mohammad. 2022. Ethics sheets for ai tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8368–8379.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1642–1652, Hong Kong, China. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Timo Schick and Hinrich Schütze. 2020. Few-shot text generation with pattern-exploiting training. *arXiv preprint arXiv:2012.11926*.

Sima Sharifirad, Borna Jafarpour, and Stan Matwin. 2018. Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 107–114.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. In *Advances in Neural Information Processing Systems*, volume 34, pages 5861–5873. Curran Associates, Inc.

Boxin Wang, Wei Ping, Chaowei Xiao, Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Bo Li, Anima Anandkumar, and Bryan Catanzaro. 2022. Exploring the limits of domain-adaptive training for detoxifying large-scale language models. *arXiv preprint arXiv:2202.04173*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

A Prompt Template Choice for Value-aligned Data Generation

Final Prompt For prompting the value-aligned training samples, we selected the final prompt template out of a set-of five prompt templates based on experimental results with the OPT-175B model. The prompt template consists of an instruction header and an enumeration of content as follows: `Generate {label} content that is relevant to the Value. Value:{value}\n`.

Here the label and value in brackets is replaced with the target label, and we provide five content examples with format `Content:{content}`, concatenated to the previous prompt. For these examples, We manually create the ten most representative examples that align with the described value and randomly select five of them for each prompt. Then, the model is encouraged to generate content relevant to the provided value and label it with the prompt `Content:`.

Tried Prompt Templates We tried five prompt templates, including the final prompt template as follows:

1. `Generate {label} content that is relevant to the Value. Value:{value}\n`.
2. `“Each item in the following list contains a value and the respective "{label}" content according to the value.Value:{value} Content:{content}”`
3. `“value="{value}"\n label="{label}"\n content={content}”`
4. `“Value:{value} Label:{label} Content:”`
5. `“Generate label content that is relevant to the Value.\nValue:{value} Content:{content}”`

We mainly investigated the effectiveness of the different prompt templates with the OPT-175B model as we conducted the main experiment with it. We also did investigation with the GPT-Jurassic 6B model. Interestingly, the GPT-Jurassic models showed better performance with data prompted with prompt template #2, which was different from OPT-175B. This may have resulted from the different training objectives and pre-training resources

of the models. Although the overall structure of our methodology is model agnostic, there should be some exploration made on prompt template construction dependent on models.

The experimental results are shown in Table 5.

VA-RoBERTa w/ prompt type	*Acc.*	*Prec.*	*Rec.*	*W-F1*
1	**73.91%**	73.91%	75.14%	**74.31%**
2	72.71%	72.71%	75.22%	73.34%
3	71.25%	71.25%	75.48%	72.06%
4	69.75%	69.75%	74.12%	70.60%
5	72.07%	72.07%	73.82%	72.61%

Table 5: Evaluation results of VA-RoBERTa trained on OPT-175B generated data with different prompt types. We prompted 120 data samples per categories.

B Experimental Details

Hyperparameters For hyper-parameters, we perform a grid search to find the best performing set of parameters among the learning rates {1e-5, 5e-5} and batch sizes {32, 64}.

Training Details For each model we train for a maximum of 30 epochs with early-stopping with patience of 5. Each experiment is conducted on an Nvidia RTX 3090 device, and each epoch takes around 2–10 minutes depending on the number of training samples.

Random Seeds We ran each of experiments five times with different random seeds and reported the mean and standard deviation in a format of $mean_{std}$.

OPT-175B (few-shot) Baseline Prompt For each test sample, we construct a prompt with the task instruction and several examples as shown below:

"`Predict a Label for the Content based on the given Value:` **`Value`**`. Content:` **`Content`** `Label:` **`Label`** `\n Predict a Label for the Content based on the given Value:` **`Value`**`. Content:` **`Content`** `Label:`"

In the prompt, the bold words will be replaced by the actual data. The first sentence is the few-shot example and we repeat it N times by randomly selecting five samples for each label category. The second sentence is the test sample, and the model will generate the corresponding label in the text. During generation, we set top-p 0.9 and generate labels five times. Finally, we calculate the average scores among the results.

C Vocabulary overlaps of generated training data among sexism categories

Figure 6 presents the vocabulary overlaps of the value-aligned training data generated from OPT-175 among the sexism categories. We calculate the vocabulary overlaps for each sexism category of the generated data.

D Per Category Results

Figure 7 presents the evaluation results of VA-BART for each sexism category on the test set.

E Human Evaluation Details

We gathered annotations from the crowd-sourcing platform Amazon Mechanical Turk (AMT). We controlled the quality by only taking annotations from AMT-qualified workers with a high acceptance rate (>95%). We also set up only adults can access our task as it contains potentially offensive content and also added a warning regarding it. We paid 0.3 US dollars per task. The screenshots of actual tasks and instructions posted on AMT are shown in Figure 8.

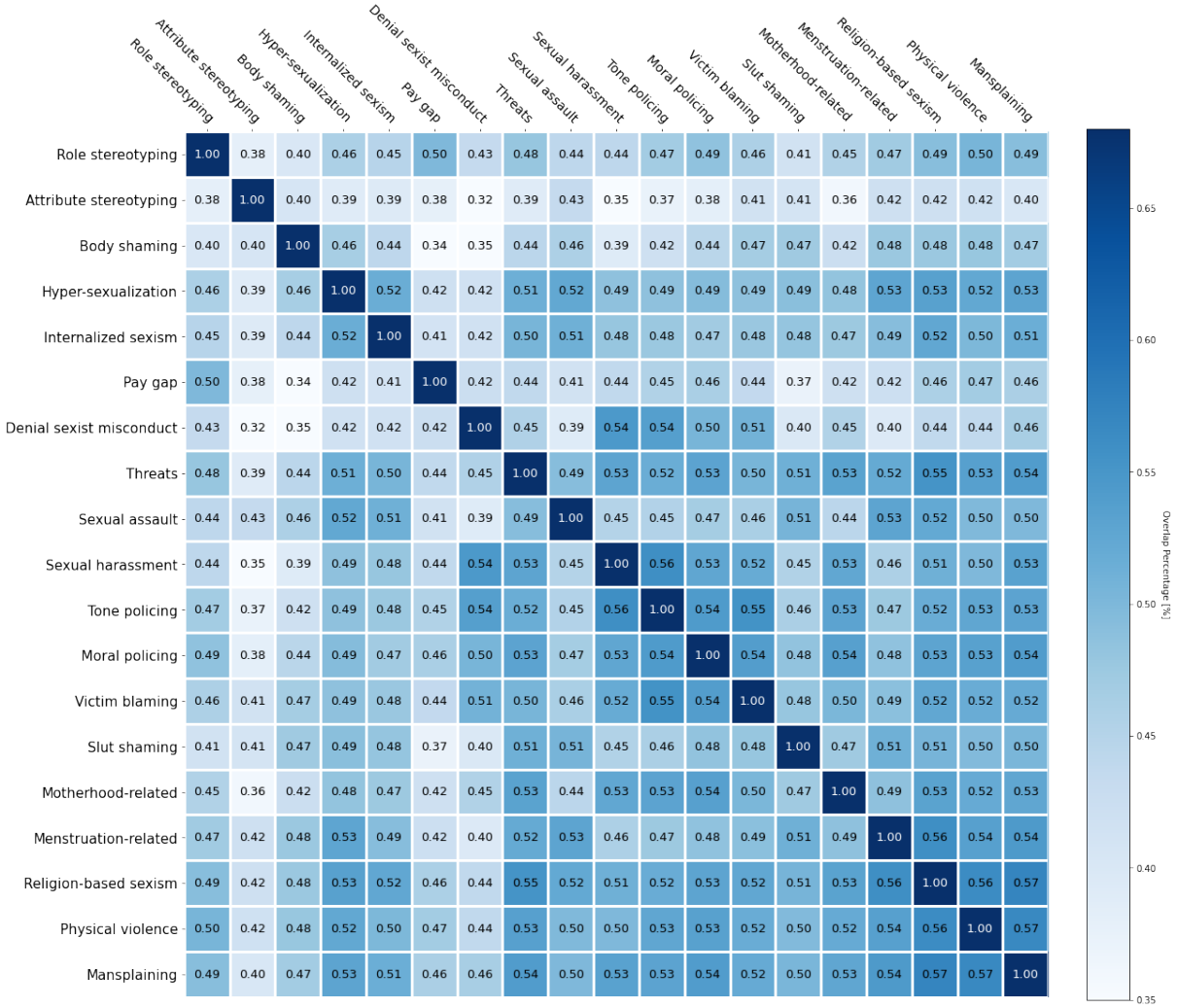


Figure 6: Vocabulary overlaps of the value-aligned training data generated from OPT-175 among sexism categories. Categories are in descending order of W-F1 (top to bottom; left-to-right).

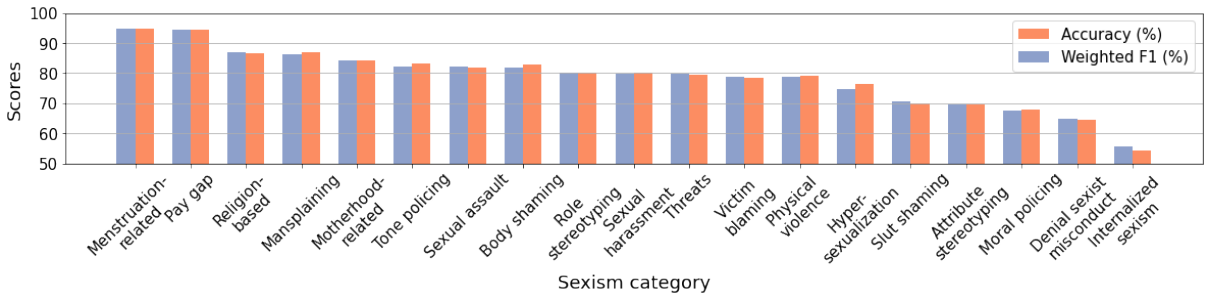
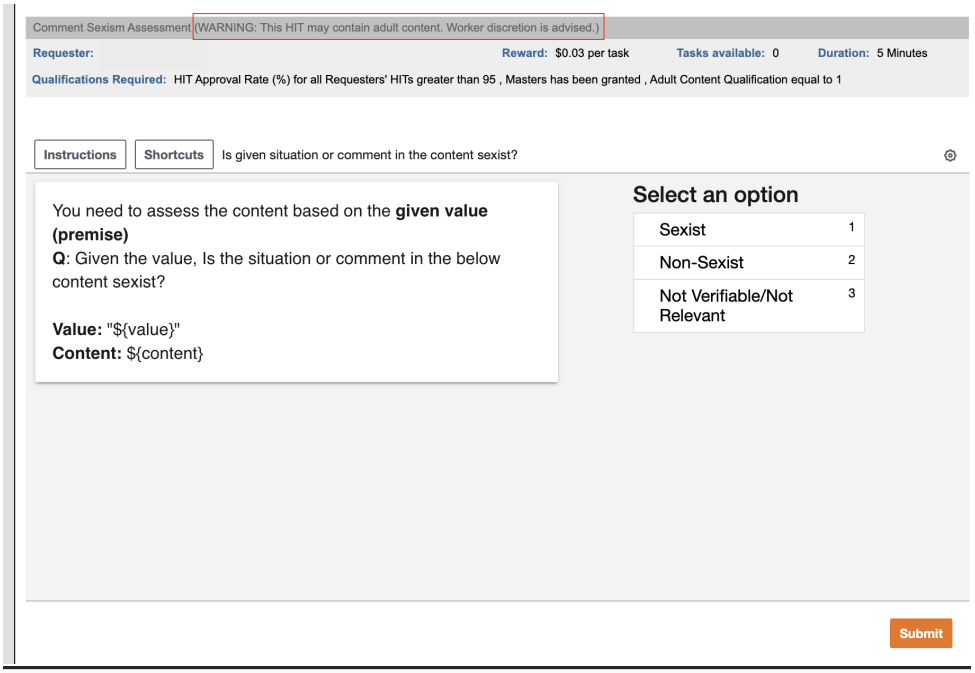


Figure 7: Evaluation results of VA-BARTfor each sexism category on the test set.



Sexist Comment Assessment Instructions

You need to assess the content based on the **given value (premise)** instead of your belief. You can assume that "I" in sentences is a woman most of the time. Example is as below:

Example 1

Value: Women should humbly consider the advice of others, whether it is appropriate or not.

Content: I'm a teacher. A male colleague suggested that I should be more assertive to get what I want. I'm already the head of a department.

Label: **Not-Sexist**

Example 2

Value: Giving unsolicited advice or explanation to women for something that they actually know well or disapprove is not acceptable.

Content: I'm a teacher. A male colleague suggested that I should be more assertive to get what I want. I'm already the head of a department.

Label: **Sexist**

Possible Labels

Sexist: Given the value, the comment or situation is sexist.

Non Sexist: Given the value, the comment or situation is not sexist.

Not Justifiable: With the value, the comment or situation cannot be assessed

Figure 8: Human evaluation for generated text Setup. The first screenshot shows the task for the annotator and the second screenshot shows a detailed explanation and instructions for the annotator to complete the task. The evaluation was done by annotators from Amazon Mechanical Turk.