An Active Learning Framework for Inclusive Generation by Large Language Models

Sabit Hassan[†] **Anthony Sicilia**[¶] **and Malihe Alikhani**[¶]
[†]School of Computing and Information, University of Pittsburgh, Pittsburgh, PA, USA
[¶]Khoury College of Computer Science, Northeastern University, Boston, MA, USA
sabit.hassan@pitt.edu, {a.sicilia,m.alikhani}@northeastern.edu

Abstract

Ensuring that Large Language Models (LLMs) generate text representative of diverse sub-populations is essential, particularly when key concepts related to under-represented groups are scarce in the training data. We address this challenge with a novel clustering-based active learning framework, enhanced with knowledge distillation. The proposed framework transforms the intermediate outputs of the learner model, enabling effective active learning for generative tasks for the first time. Integration of clustering and knowledge distillation yields more representative models without prior knowledge of underlying data distribution and overbearing human efforts. We validate our approach in practice through case studies in counter-narration and style transfer. We construct two new datasets in tandem with model training, showing a performance improvement of 2%–10% over baseline models. Our results also show more consistent performance across various data subgroups and increased lexical diversity, underscoring our model's resilience to skewness in available data. Further, our results show that the data acquired via our approach improves the performance of secondary models not involved in the learning loop, showcasing practical utility of the framework.

1 Introduction

Despite advancements, Large Language Models (LLMs) have been under scrutiny for exhibiting bias toward under-represented groups (Nozza et al., 2022; Baffour et al., 2023). Standard fine-tuning may not mitigate this (Zhou et al., 2023) as a random sample drawn from skewed data would mirror biases and the fine-tuned model may fail for under-represented groups. To address this, we introduce a novel active learning framework for **generative tasks**, combining clustering and knowledge distillation to yield more inclusive generative models.

Active learning alternates between dataset construction and model training, focusing on the most

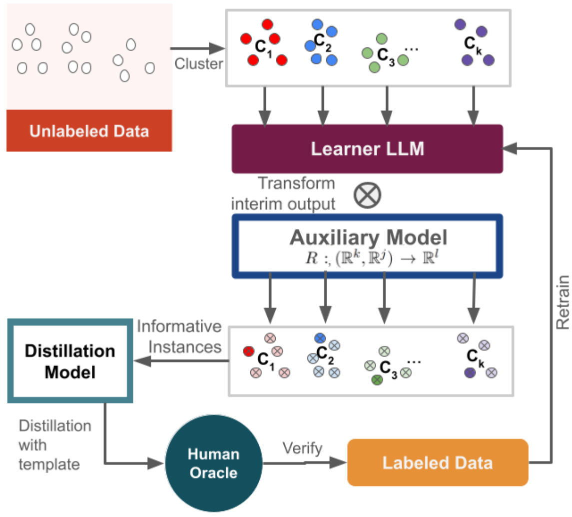

Figure 1: The training loop of our framework uses an auxiliary model to transform the interim output of a learner LLM, and selects informative instances from clustered unlabeled data. A distillation model then generates outputs, verified by humans, to iteratively refine the learner LLM.

informative instances (Settles, 2009). Unlike classification tasks, active learning for generative tasks faces two key challenges: i) traditional informativeness measures like entropy are ineffective due to vast output spaces, and ii) such measures provide only token-level feedback, complicating full-sequence aggregation. To address these, we propose a new informativeness measure (Section 3), computed by an auxiliary model that transforms intermediate output token sequences into a **1D** latent space, conditioned on a regulated attribute. This measure is not dependent on singular tokens and can be applied to large output space of LLMs.

We advocate for clustering-based active learning in our framework, hypothesizing that informative samples from diverse regions would guard against distributional skewness in real-world data, leading to more inclusive generation. As opposed to traditional bias mitigation approaches that often

Proceedings of the 31st International Conference on Computational Linguistics, pages 5403–5414
January 19–24, 2025. ©2025 Association for Computational Linguistics

require extensive post-hoc analysis and rebalancing (Han et al., 2022; Sun et al., 2019), our method would proactively identify underrepresented groups in data without knowing the underlying distribution beforehand. Additionally, we incorporate knowledge distillation with an external LLM in our framework, leveraging its repository of commonsense and expert knowledge (West et al., 2022; Hsieh et al., 2023) to aid the active learner model. The output of the external LLM are verified by human annotators before being passed on to the learner model, reducing necessary human efforts in the active learning paradigm.

We validate our approach with case studies of counter narration (Fanton et al., 2021) and style-transfer of offensive text (Nogueira dos Santos et al., 2018). We choose these tasks as it is crucial to ensure inclusivity of under-represented groups when addressing harmful text. We construct two novel datasets of **1K** counter-narrations and **1K** style transfers, with an emphasis on social acceptability. The datasets contain multiple splits obtained using our framework and baseline methods. This highlights a key difference with prior work as we evaluate the **practical viability of active learning** beyond mainly simulation-based evaluations (Zhang et al., 2022). Practical evaluation of active learning is challenging as it requires developing multiple splits of data from scratch. While this restricts range of datasets, models and settings compared to simulations, our work answers the critical question of whether active learning can be viable in practice. Our results show that the learner LLMs not only perform well with small datasets but also exhibit greater diversity and lower ratio of errors for under-represented groups (Section 5), suggesting a higher degree of inclusivity. We also validate transferability of acquired data to models outside active learning, showcasing genarliazability. Thus, key contributions in this paper are threefold:

- Introducing a novel active learning framework for generative tasks, enhanced with clustering and knowledge distillation.
- Case studies of style-transfer and counter-narration with two publicly available datasets.
- Demonstrating effectiveness of active learning for generative tasks in practice and its transferability to models outside the learning loop.

Our datasets are made publicly available. [1]

[1] https://github.com/sabithsn/generative-AL

2 Related Work

Active Learning in NLP Although active learning has been studied for a multitude of NLP tasks (Zhang et al., 2022), almost all existing works target classification scenarios (Rotman and Reichart, 2022; Ein-Dor et al., 2020). Recent efforts in introducing LLMs with active learning (Hassan et al., 2024; Diao et al., 2023; Margatina et al., 2023), assume a fixed set of outputs —still classification tasks, albeit with LLMs. Active learning for *generative tasks* remains mostly unexplored. Perlitz et al. (2023) show that standard active learning methods for classification do not translate well to generative tasks and call for further investigation. Our work is the first to make significant inroads in active learning for generative tasks. Different from works that propose LLMs as annotators (Xiao et al., 2023), our work is the first to utilize knowledge distillation of LLMs, coupled with clustering, for *generative active learning*. Our work is also one of the first to evaluate active learning in practice as opposed to simulations (Zhang et al., 2022).

Countering Distributional Bias NLP models have been under scrutiny for exhibiting bias (Lu et al., 2020; Ahn and Oh, 2021; Sap et al., 2019) but reducing such biases has been challenging. First, bias can manifest in a multifaceted way: a model can be biased against gender while also exhibiting linguistic bias (Savoldi et al., 2022). Existing approaches often focus on mitigating particular kind of bias such as gender bias (Sun et al., 2019) by training gender neutral embeddings (Zhao et al., 2018) or removing gender direction in embeddding subspace (Liang et al., 2020). These methods however, cannot be generalized to address the different types of bias. Another critical challenge is that true distribution of a data is often unknown. As such, existing approaches are often expensive that includes rebalancing (Han et al., 2022) or re-annotating a large amount of data (Sun et al., 2019). Our approach offers a generalizable efficent method that does not assume availability of additional data or modification of the model itself.

Counter Narration and Style Transfer Counter narration and style transfer have been suggested as alternative approaches (Mathew et al., 2018; Nogueira dos Santos et al., 2018) to filtering-based approaches (Ye et al., 2023; Hassan et al., 2020) for addressing harmful text. Prior work in counter-narration primarily address social-media (Hassan and Alikhani, 2023b; Mathew et al., 2018) or niche-sourced setting (Fanton et al., 2021; Chung et al., 2021). Style-transferring offensive text has been limited to social media context (Nogueira dos Santos et al., 2018; Atwell et al., 2022). The output of both of these tasks may not be deemed appropriate when generated by LLMs. Our datasets are the first to consider social acceptability of the generations.

3 Framework

3.1 Background

Most popular active learning paradigms iteratively add informative samples to training data (Settles, 2009), offering performance boost over randomly choosing samples. In a randomly chosen sample of skewed real-world data, under-represented groups would remain under-sampled. On the other hand, standard active learning can ignore under-represented groups in data if the learner model is confidently incorrect (Hassan et al., 2018). Integrating clustering with active learning can address this challenge and lead to more inclusive models by enforcing informative samples to be chosen from diverse regions in data. In this work, we propose to extend the standard active learning paradigm by applying knowledge distillation to informative samples from each cluster. LLMs like GPT-4 (OpenAI et al., 2023), which already contain vast amounts of information, can serve to distill knowledge to train other models, reducing exposure to harmful content for human annotators.

This approach, while promising, faces the critical challenge of identifying informative samples. As classification models typically deal with a manageable number of classes, using probabilities in measures such as entropy (Eq. 1) can gauge informativeness. For generative tasks however, the choices at each timestep match the extensive vocabulary size, typically on the order of 10^4 (Perlitz et al., 2023), making probability-based measures less reliable. Further, the vocabulary choices are often interchangeable and dependent on previous tokens, which are not taken into account by measures such as entropy. Thus, we propose a novel method to implement active learning for generative tasks next.

3.2 Active Learning for Generative Tasks

Preliminaries We assume there is a large pool of unlabeled dataset U but only a small subset of labeled training data L can be obtained. L is iteratively constructed by querying label for the *most-informative* instance with respect to a learner model G. While other active learning scenarios exist (Settles, 2009), we focus on pool-based active learning because for many NLP tasks a large amount of unlabeled text data can be obtained from the web. The most commonly used measure of informativeness is the uncertainty metric entropy and the instance with highest entropy is defined as:

$$x_E^* = \underset{x}{argmax} \; -\sum_i P_\theta(y_i|x) log P_\theta(y_i|x) \quad (1)$$

In Eq. 1, i ranges over all possible outputs. For an instance x, the probability of class y_i is denoted as $P_\theta(y_i|x)$. Entropy is higher when probabilities are evenly distributed, indicating greater model uncertainty in class selection. However, as discussed earlier, this approach does not translate well to generative tasks. Thus, we propose a new measure for informativeness next.

Infromativeness Measure for Generative Tasks To formulate a new measure of informativeness, we introduce the concept of a *regulated attribute* H. We define a regulated attribute as a property an LLM is expected to preserve during generation. For instance, when style-transferring offensive text or generating counter-narration, we expect the text to be inoffensive post generation. Thus, inoffensiveness would be the regulated attribute of the task. While we focus on a single regulated attribute in this work due to the nature of the tasks, it can be a weighted combination of multiple attributes.

We propose training an auxiliary model R, to estimate the value of regulated attribute on interim output of the learner model. We can imagine the interim output to be in a **2-D** space, where each token is a vector. R maps the **2-D** interim output $G(X_i)$ to a **1-D** latent space conditioned on regulated attribute H. Then, a softmax over this **1-D** space would inform how well the learner model is adhering to regulated attribute on an instance. While there are no restrictions on R's internal structure (e.g., it can be another pretrained transformer), it is required that the final output layer is a linear layer so that softmax can be applied. Formally, we define the informativeness of a sample as follows:

$$E_i = Softmax(R(G(x_i), H)) \quad (2)$$

In the equation, $G(x_i)$ is the LLM generated output on input x_i. R is an auxiliary model with a linear output layer that transforms $G(x_i)$ with respect to regulated attribute H. Since $R(G(x_i), H)$ is a vector with a single dimension, we can apply softmax over this vector. We replace the entropy portion of Eq. 1 with this measure to obtain most important sample for generative tasks:

$$x_E^* = \underset{x}{argmax} \; - Softmax(R(G(x), H)) \quad (3)$$

Clustering-based and Knowledge Distillation While the proposed measure is applicable to standard active learning paradigm (Section 5), we advocate for using it in conjunction with clustering as outlined in Section 3.1. The unlabeled data is vectorized and split into m clusters $\{C_1, C_2, ...C_m\}$ and most informative samples according to our measure (Eq. 3) are chosen from each cluster.

Next, we apply knowledge distillation on these samples. We assume we have access to a large language model, S, and we want to leverage knowledge distillation of S to assist training of learner model G. In our approach, we apply knowledge distillation with the concept of a *template* T:

T(x) : on input x, prompt S to generate $f(x)$ while respecting instruction $I(x)$.

$f(x)$ is the target task, e.g., "style transfer input text from offensive to inoffensive". And I is instruction specific to the task. I can be in different forms depending on the task. For instance, $I(x)$ can be "ensure generated output respects social acceptability". Instead of standard human labeling step, we pass this template to S on instances corresponding to Eq.3 for each cluster C_i. The generated content is then verified by a human and added to the training data. The learner model is retrained and this process is repeated until resources runs out. We consolidate our approach in algorithm 1.

Algorithm 1 Knowledge Distilled Clustering-Based Active Learning for Generative Tasks

```
U, L ← unlabeled data, labeled data
S ← LLM for distillation
G ← bootstrapped model
R ← auxiliary model
H ← regulated attribute
B ← labeling budget
N ← annotation batch size
m ← initial number of clusters
V ← vectorize U
Cluster V into {C_1, C_2, ... C_m}
while B ≥ 0 do
    for i=0,1,...m do
        for j=0,1,...|C_i| do
            E_ij ← R(G(X_ij), H)
        end for
        x*_i ← argmax_j(E_ij)
        T*_i ← generation template T for x*_i
        {(x*_ik, y*_ik)} ← Distill S with T_i(x_i)
        Add {(x*_ik, y*_i)} to L
    end for
    G ← retrain on L
    B = B − N
end while
```

4 Case Studies

4.1 Task Definitions

Counter-Narration When faced with offensive or sensitive content, an LLM may opt to not respond, potentially allowing harmful ideologies to persist. Counter-narration (Fanton et al., 2021), also known as counterspeech (Mathew et al., 2018), has been proposed as an alternative to address this issue. (Benesch, 2014) suggest that counterspeech can be more effective long-term in addressing offensive language. However, these approaches are often discussed within the context of online hate. From an LLM's safety perspective, overly aggressive responses may alienate users. Therefore, the scope of counter-narration needs to be adjusted to be appropriate for an LLM response in a socially acceptable manner. For social acceptability, an LLM needs to generate counters that are polite and respectful (Brown and Levinson, 1987), focusing on non-imposition and the preservation of freedom and desire. We propose utilizing the template to prompt distillation model to adhere to social acceptability and craft more appropriate counter-narration for the learner model. Examples of this approach is illustrated in Table 1 alongside examples from existing works.

Style-Transfer Simply censoring model output when it generates offensive language can detract from user experience. Style transfer —the task of rephrasing text to retain specific stylistic properties without altering the underlying intent (Prabhumoye et al., 2018), has been suggested as an alternative method to mitigate offensiveness (Nogueira dos Santos et al., 2018; Atwell et al., 2022). However, existing models primarily address social media con-

Sensitive Content	Prior Counter-Narration	Proposed Counter-Narration
We should stop immigrants.	If you really wanted to stop the migration, you would destroy our economy, by stopping the flow of people.. (Fanton et al., 2021).	immigrants often contribute significantly to our society and economy through their diverse skills and cultures.
Offensive Text	**Prior Style-Transfer**	**Proposed Style-Transfer**
you're just too dumb to see you're wrong	You just can't see you're wrong. (Atwell et al., 2022)	it seems we have a difference of opinion on this issue

Table 1: Existing counter-narrative/style-transfer may not be appropriate for LLMs as they can be aggressive. Our proposed approach do not simply counter offensive text or paraphrase them but are also more respectful and polite.

tent and may not ensure a non-aggressive stance essential for LLM safety. We propose adapting style-transfer to not only remove offensiveness but also to ensure it conforms to social expectations that LLMs should not be aggressive toward users. Like counter-narration, we advocate for integrating this social acceptability into the style-transfer process. Table 1 demonstrates how our proposed style-transfer differs from prior work.

4.2 Datasets

We contribute two datasets constructed using our framework for counter narration and style-transferring offensive text into inoffensive ones. To construct the datasets, we start with an unlabeled pool of data. From this unlabeled pool of data, instances are chosen either randomly (for baseline) or informative instances added to the train data iteratively following our proposed active learning paradigm. Number of instances in train data is kept small as the goal of active learning is to train models with limited resources. We construct three training splits for each task:

1. **Standard:** Samples are chosen randomly from unlabeled data for fine-tuning.

2. **TopN-AL:** N most informative samples are added to training data without clustering.

3. **Cluster-AL:** The unlabeled data is clustered into K regions. N/K most informative instances are chosen from each region.

Standard fine-tuning with randomly chosen samples serves as baseline. TopN-AL relies solely on informativeness to obtain fine-tuning data but does not utilize clustering. Cluster-AL is the paradigm of our framework, outlined in Algorithm 1. All three approaches have access to GPT-4 as distillation model. To evaluate these different paradigms, we choose a randomly chosen subset of the unlabeled data as the test set. The test set is kept fixed with all splits for consistent results.

Counter-Narration Dataset We start by creating unlabeled pool of data from the source text of (Hassan and Alikhani, 2023b) and (Fanton et al., 2021). Hassan and Alikhani (2023b) contains 250 offensive texts and Fanton et al. (2021) contains 5K hatespeech targeting women, LGBTQ+, muslims, migrants, disabled and jews. While both of the original datasets contain counter-speech data, we do not use them and create new set of counter-narrations according to Section 4.1 as counters in (Hassan and Alikhani, 2023b; Fanton et al., 2021) may not be socially acceptable for use in context of LLM safety (Table 1).

Utilizing the template with GPT-4 (OpenAI et al., 2023), we generate counter-narrations of 400 random samples to create the test set. To build the train splits, we use 100 random instances for bootstrapping the learner model. 100 more are added according to definitions of Standard, TopN-AL and Cluster-AL splits. The total count of this dataset is 3X200 + 400 = 1000 pairs.

Style-Transfer Dataset Similar to counter-narration, we start with a set of offensive texts as our unlabeled pool of data. For this, we use offensive-text portion of APPDIA (Atwell et al., 2022) and a subset of the OLID dataset (Zampieri et al., 2019). The APPDIA portion of the data contains 2K offensive comments from Reddit and the OLID portion of the data contains 2K offensive posts from Twitter. Similar to the counter-narration scenario, Atwell et al. (2022) do contain style-transferred counterparts but we do not use them as they may not be socially acceptable for LLMs.

The construction of the training and test sets follow the same process as counter narration, yielding a total of 3X200 + 400 = 1000 pairs in the dataset.

Dataset Validation Two graduate student annotators were hired to verify the counter-narrations and style-transfers distilled from GPT-4. The annotators were provided with definitions of counter-narration and style-transfer from (Hassan and Alikhani, 2023b) and (Atwell et al., 2022) respectively, with the additional constraint that the generations should be socially acceptable (i.e., not aggressive). The distilled generations were found to be reliable in $> 98\%$ cases. The annotators were hired and paid according to our institution's guidelines.

5 Experiments

5.1 Experiment Setup

In our experiment setup, we aim to evaluate active learning in practice rather than simulation. To this end, learner models are trained over five iterations of active learning while constructing the train splits, before evaluation on the fixed test. The *transfer models* reside outside the active learning setting. The success of our framework would be indicated by not only performance on the two tasks, but also by being inclusive of different groups in data.

5.2 Models

Learner Model: We use FLAN-T5-base (Chung et al., 2022) as the learner model. FLAN-T5 is an instruction-tuned model with 220 million parameters. We choose FLAN-T5 as learner model as instruction-tuning makes it a capable LLM while its relative smaller size makes it easily deployable. The model is fine-tuned at each iteration of active learning for 10 epochs with learning rate of 3e-5.

Distillation Model: We choose GPT-4 (OpenAI et al., 2023) for knowledge distillation due its vast knowledge base and advanced generation ability.

Auxiliary Model: In this work, we use lightweight transformers with attached linear layers as auxiliary models. For counter-narration, we use a DistillBERT (Sanh et al., 2020), presented in Kim et al. (2023), trained to determine if a bot-response is safe in with respect to a human prompt.

For style-transfer, we are only interested in the offensiveness of text post-style transfer, not the input text. We fine-tune a bert-base-cased (Devlin et al., 2019) on the Jigsaw dataset [2], achieving 93% macro-averaged F1 score on the Jigsaw test set.

Clustering: For vectorizing the data, we use sentence transformer MiniLM-V2 (Wang et al., 2020). The vectorized data is clustered using KMeans with default parameters of scikit-learn [3]. Number of clusters is set to 10, similar to prior work in classification domain (Hassan and Alikhani, 2023a).

Transfer Models: Data acquired by the learner model is used to fine-tune a DialoGPT-large (Zhang et al., 2020) and a Mistral-7B model (Jiang et al., 2023). While Flan-T5 is an encoder-decoder model, DialoGPT is an earlier decoder-only model with 774 million parameters. Mistral-7B on the other hand, is a modern and larger model with 7 billion parameters that utilizes group query attention. These models are chosen to be significantly different from the learner FLAN-T5, and span different generations of LLMs. Transferability across generations would validate ubiquity of our approach.

5.3 Counter-Narration Results

Evaluation Since there is no clear automated way to evaluate counter-narration (Hassan and Alikhani, 2023b), we conducted human evaluation of 3600 model outputs (400 per model and data split). Table 2 presents examples of generated counter-narration. We randomized the outputs from the baselines and our proposed model, and tasked two graduate student annotators from earlier (Section 4.2) with determining if the generated counter-narration effectively counters the offensive text. A counter-narration is deemed inaccurate if it concurs with the offensive text, strays off-topic, or is incoherent. We also calculate the error-ratio-variance with respect to the different target groups present in the original data (e.g., MUSLIM, MIGRANT). This metric is computed by first calculating the error percentage for each group and then the variance across these percentages. A lower variance would suggest that the model is more inclusive by not failing disproportionately for certain groups.

Clustering-based AL yields more effective counters. From Table 3 we can observe that our proposed approach, Cluster-AL, outperforms the baseline active learning (without clustering) by **5.5%** and random sampling by **10.8%** for the learner FLAN-T5. while having access to the same amount of fine-tuning data.

Clustering-based AL is more inclusive. We can also observe a substantial reduction in error-ratio-variance; from **0.01** in standard fine-tuning with random sampling to **0.0049** for Cluster-AL, which

[2] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-\challenge

[3] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Offensive Text 'All children from the age of 4 will have mandatory lessons on same sex relationships.' – This is not about Education, this is about willingly confusing and indoctrinating kids...
Random: everyone has the right to express their identity.... we should foster understanding and understanding.
TopN-AL: it's important to remember that every child, regardless of their sexual orientation deserves respect...
Cluster-AL: everyone has the right to learn about same sex relationships, which also includes children. promoting diversity enhances our society.

Offensive Text Children with Down syndrome are retarded, so more likely to abuse drugs and alcohol.
Random: actually, down syndrome doesn't identify any specific group, like children, who are more likely to abuse drugs...
TopN-AL: down syndrome is not a genetic disorder, it's a medical condition that affects individuals with...
Cluster-AL: people with down syndrome, like anyone, have diverse abilities and temperaments. they deserve respect...

Table 2: Examples of counter-narrations generated by different approaches for the learner model Flan-T5. We can observe that randomly choosing samples may result in more failed counter-narrations. Our generations are also more socially acceptable compared to prior work in Table 1.

Approach	**FLAN-T5#**		**DialoGPT***		**Mistral 7B***	
	CS-Score ↑	**↓ Error Var.**	**CS-Score ↑**	**↓ Error Var.**	**CS-Score ↑**	**↓ Error Var.**
Standard	65.0	0.0101	55.5	0.0238	85.8	0.0130
TopN-AL	70.5	0.0061	60.3	0.0230	86.5	0.0073
Cluster-AL	**75.8**	**0.0049**	**66.8**	**0.0155**	**89.3**	**0.0018**

Table 3: CS-Score is the precentange of generated text evaluated to be proper counters. Error Var. denotes variance of error ratios across different targets of hatespeech. A lower value would indicate the errors are less skewed. Our active learning based approach shows efficacy over standard fine-tuning. Clustering-based active learning yields the best results. # marks the learner model and * shows transferability to other models.

is also lower than baseline active learning **(0.0061)**. Lower error-ratio-variance coupled with enhanced performance suggests that our method not only produces more accurate counter narrations but also performs more consistently across various groups. This consistency is illustrated in Figure 2, which displays the error ratio for each class alongside the class's ratio in the original dataset. The initial data pool had few instances targeting Persons of Color (POC) and Disabilities, leading to high error rates with standard fine-tuning with random samples. However, our method reduces error rates for these underrepresented groups, likely due to the clustering step ensuring that active learning selects informative instances from diverse data regions. In skewed data regions (e.g., POC), where random sampling would likely overlook valuable samples, our framework effectively includes them.

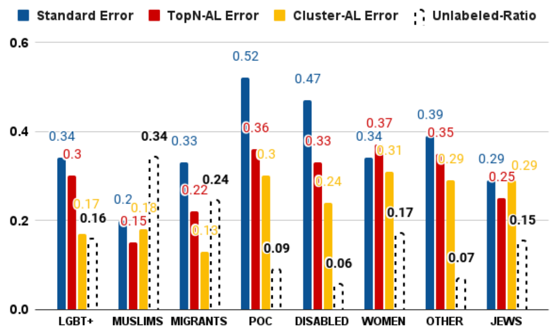

Figure 2: Error ratio of resulting models, along with original data distribution (dashed line). Our clustering-based active learning approach is robust against data distribution skewness.

5.4 Style-Transfer Results

Evaluation Unlike counter-narration, the offensive counterparts in the original datasets used for style-transfer unlabeled data (Atwell et al., 2022; Zampieri et al., 2019) lack tags for offensiveness classes. Therefore, we adopt the Measure of Textual Lexical Diversity (MTLD) (McCarthy and Jarvis, 2010) to assess diversity. Although MTLD doesn't directly measure bias, a higher MTLD, if other metrics are stable, would suggest that more diverse samples are selected. We also use a BERT-based classifier fine-tuned on the OLID dataset (Zampieri et al., 2019) (F1 score 89%) to compute automated SafeScore —the percentage of style-transferred text predicted to be inoffensive.

Clustering-based AL yields less offensive style-transfers. Table 5 shows that our approach reduces offensiveness more effectively. For the learner model FLAN-T5, SafeScore increased from **94.9%** to **97%**. Additionally, the MTLD value significantly increased from **60.89** to **99.81**, indicating that our method acquired more diverse samples compared to standard fine-tuning with random

Offensive Text *what the f****	
Random: what the f***?	
TopN-AL: i'm sorry, but i can't assist with that	
Cluster-AL: it seems like there might be some misunderstanding here. let's try to clarify this together.	
Offensive Text *And when they met he looked at her and said I'm in charge in you're my b*****	
Random: when they met, he expressed he's in charge of you're my b****.	
TopN-AL: when they met he stated he was in charge in her position and he tried to control the situation	
Cluster-AL: when they came across each other, he looked at her and said he would take the lead	

Table 4: Examples of style-transfer by the different approaches for the model FLAN-T5. Random sampling may fail to remove offensiveness from complicated samples whereas our proposed approach can do so effectively. Our generations are also more respectful compared to prior work in Table 1.

Approach	FLAN-T5[#]		DialoGPT*		Mistral 7B*	
	Safe-Score ↑	MTLD ↑	Safe-Score ↑	MTLD ↑	Safe-Score ↑	MTLD ↑
Standard	94.9	60.89	97.0	80.03	**99.5**	124.0
TopN-AL	96.5	92.88	96.7	77.54	99.0	122.8
Cluster-AL	**97.0**	**99.81**	**98.0**	**82.44**	**99.5**	**124.8**

Table 5: SafeScore is the percentage of style-transferred text predicted to be inoffensive. MTLD measures lexical diversity in generated output. Our proposed approach has more substantial improvement for smaller models FLAN-T5 and DialoGPT. [#] marks the learner model and [*] shows transferability to other models.

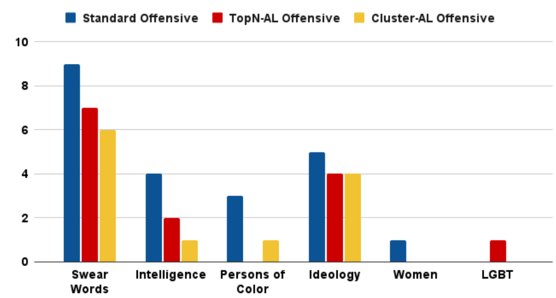

Figure 3: Offensive counts after style transfer. Random sampling leads to higher error rates overall, especially for target groups like persons of color. Cluster-AL achieves the lowest offensiveness overall and for most individual groups.

sampling. While TopN-AL surpasses the standard approach, Cluster-AL exceeds both.

Clustering-based AL is more robust across data types. Our analysis of instances deemed offensive post style-transfer in Figure 3 reveals that random sampling often fails for under-represented groups like Persons of Color, and overlooks use of uncommon swear words. Both TopN and Cluster-AL approaches reduce offensiveness more effectively, with Cluster-AL showing greater robustness across most data types.

5.5 Transferability of Active Learning

Table 3 and 5 demonstrate the transferability of active learning. For counter-narration, Cluster-AL shows substantial gains, improving efficacy by **10%** over standard fine-tuning for DialoGPT. Although less pronounced, a similar pattern is observed for Mistral 7B, where efficacy increases to **89.3%** from **85.8%**, and the error ratio variance also decreases significantly for both models (from **0.0238** to **0.0155** and from **0.0130** to **0.0018**), suggesting that the models are more inclusive of different subgroups in data.

The transferability for style-transfer is less pronounced. While we do see a higher MTLD score for DialoGPT and Mistral 7B, the performance improvement is modest. This can be attributed to the fact that the models were already achieving very high SafeScore. This suggest that the data acquired by active learning is more useful to other models when the task is more challenging.

6 Conclusion and Future Work

We presented a novel framework for active learning, making significant inroads toward making active learning viable for generative tasks. By integrating an auxiliary model that transforms interim output, along with clustering and knowledge distillation, our framework effectively generates counter-narrations and style-transfers. The results show that our approach prevents high error rates for under-represented groups and achieves greater lexical diversity. Our approach achieves this without having prior knowledge of the data distribution, highlighting the framework's potential as a method for more inclusive LLM generation. Creation of two novel datasets and their use across different models validate the practical viability.

Our findings pave the way for further research in expanding active learning to different generative tasks. Future applications could include dialogue systems (Sicilia et al., 2023), digital health interventions (Wang et al., 2023) and multimodal generative tasks such as signed language generation (Inan et al., 2022). Future research could explore different ways of integrating knowledge distillation and clustering and expand on our use of auxiliary models to integrate more complex models. The proposed framework could also be adapted to address specific biases, like behavioral or gender biases.

Limitations

We presented a novel framework for active learning for generation and applied the framework in practice by constructing two datasets with multiple splits and training models simultaneously. When applying active learning in practice, it is necessary to construct the dataset splits from scratch. Thus, a larger number of simulated experiments and datasets, as seen in prior work, is not feasible in practice. We hope our work can initiate a trend of evaluating active learning in practice so that active learning for generative tasks is adopted broadly in practical applications.

Additionally, as seen from our results, our proposed approach can draw out under-represented groups from data without knowing underlying distribution. While this improves representation of minority groups, it cannot eliminate the problem of representation completely. Thus, it is still important to monitor behavior of these models before deploying.

Ethical Considerations

We presented an annotation efficient approach for drawing out under-represented groups from data. While this comes with better representativeness of minority groups and inclusivity in behavior of generative models, it is important to use the approach responsibly and not change the algorithm to exacerbate biases *against* minority groups.

Since our approach also comes with the benefit of lower involvement of human annotators due to integration of knowledge distillation with clustering-based active learning, we advocate for considering reallocation of saved annotation resources. The saved resources could be used for continued evaluation and training purposes rather than simply reducing human involvement.

References

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. APPDIA: A discourse-aware transformer-based style transfer model for offensive social media conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6063–6074, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Perpetual Baffour, Tor Saxberg, and Scott Crossley. 2023. Analyzing bias in large language model solutions for assisted writing feedback tools: Lessons from the feedback prize competition series. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 242–246, Toronto, Canada. Association for Computational Linguistics.

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. *Preprint*, arXiv:2302.12246.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2022. Balancing out bias: Achieving fairness through balanced training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11335–11350, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sabit Hassan and Malihe Alikhani. 2023a. D-CALM: A dynamic clustering-based active learning approach for mitigating bias. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5540–5553, Toronto, Canada. Association for Computational Linguistics.

Sabit Hassan and Malihe Alikhani. 2023b. Discgen: A framework for discourse-informed counterspeech generation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 420–429, Nusa Dua, Bali. Association for Computational Linguistics.

Sabit Hassan, Younes Samih, Hamdy Mubarak, and Ahmed Abdelali. 2020. Alt at semeval-2020 task 12: Arabic and english offensive language identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1891–1897.

Sabit Hassan, Shaden Shaar, Bhiksha Raj, and Saquib Razak. 2018. Interactive evaluation of classifiers under limited resources. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 173–180.

Sabit Hassan, Anthony Sicilia, and Malihe Alikhani. 2024. Active learning for robust and representative llm generation in safety-critical scenarios. *Preprint*, arXiv:2410.11114.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Mert Inan, Yang Zhong, Sabit Hassan, Lorna Quandt, and Malihe Alikhani. 2022. Modeling intensification for sign language generation: A computational approach. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2897–2911, Dublin, Ireland. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jinhwa Kim, Ali Derakhshan, and Ian G. Harris. 2023. Robust safety classifier for large language models: Adversarial prompt shield. *Preprint*, arXiv:2311.00172.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*.

Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active learning principles for in-context learning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5011–5034, Singapore. Association for Computational Linguistics.

Binny Mathew, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2018. Thou shalt not hate: Countering online hate speech. In *International Conference on Web and Social Media*.

Philip M. McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42:381–392.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia. Association for Computational Linguistics.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Pipelines for social bias testing of large language models. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 68–74, virtual+Dublin. Association for Computational Linguistics.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Yotam Perlitz, Ariel Gera, Michal Shmueli-Scheuer, Dafna Sheinwald, Noam Slonim, and Liat Ein-Dor. 2023. Active learning for natural language generation. *Preprint*, arXiv:2305.15040.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Guy Rotman and Roi Reichart. 2022. Multi-task active learning for pre-trained transformer-based models. *Transactions of the Association for Computational Linguistics*, 10:1209–1228.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022. Under the morphosyntactic lens: A multifaceted evaluation of gender bias in speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1807–1824, Dublin, Ireland. Association for Computational Linguistics.

Burr Settles. 2009. Active learning literature survey.

Anthony Sicilia, Yuya Asano, Katherine Atwell, Qi Cheng, Dipunj Gupta, Sabit Hassan, Mert Inan, Jennifer Nwogu, Paras Sharma, and Malihe Alikhani. 2023. Isabel: An inclusive and collaborative task-oriented dialogue system. *Alexa Prize TaskBot Challenge 2 Proceedings*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Yan Wang, Heidi Donovan, Sabit Hassan, and Malihe Alikhani. 2023. MedNgage: A dataset for understanding engagement in patient-nurse conversations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4613–4630, Toronto, Canada. Association for Computational Linguistics.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Ruixuan Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen, and Haobo Wang. 2023. FreeAL: Towards human-free active learning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14520–14535, Singapore. Association for Computational Linguistics.

Meng Ye, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran, and Malihe Alikhani. 2023. Multilingual content moderation: A case study on Reddit. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3828–3844, Dubrovnik, Croatia. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *NAACL*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241, Toronto, Canada. Association for Computational Linguistics.