Active Learning for Robust and Representative LLM Generation in Safety-Critical Scenarios

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 robust safety measures across a wide range of scenarios is crucial for user-facing systems. While Large Language Models (LLMs) can generate valuable data for safety measures, they often exhibit distributional biases, focusing on common scenarios and neglecting rare but critical cases. This can undermine the effectiveness of safety protocols developed using such data. To address this, we propose a novel framework that integrates active learning with clustering to guide LLM generation, enhancing their representativeness and robustness in safety scenarios. We demonstrate the effectiveness of our approach by constructing a dataset of **5.4K** potential safety violations through an iterative process involving LLM generation and an active learner model's feedback. Our results show that the proposed framework produces a more representative set of safety scenarios without requiring prior knowledge of the underlying data distribution. Additionally, data acquired through our method improves the accuracy and F1 score of both the active learner model as well models outside the scope of active learning process, highlighting its broad applicability.

1 Introduction

LLMs have shown much promise in data generation (Radharapu et al., 2023), which can be leveraged to obtain safety-related data. This data can then be employed to implement safety measures in various models (Radharapu et al., 2023; Sun et al., 2022). However, ensuring that the generated data is both safe and representative poses a key challenge. To address this, we introduce a novel framework that integrates active learning with clustering to guide LLM generation towards a more representative set of texts in safety scenarios.

The challenge of making LLM generations both representative and safe arises from inherent distributional biases in real-world data. These biases often cause LLM-generated content to mirror the imbalances, resulting in an over-representation of common scenarios and an under-representation of rare but critical situations. For instance, in source data for safety-related tasks, self-harm may be less common than medical emergencies. Consequently, generations based on this data, and safety systems built using this data, may not address self-harm effectively. Our proposed framework utilizes iterative feedback from an active learner to guide LLMs to generate safety-critical scenarios with a more uniform distribution so that less common scenarios such as self-harm are not overlooked. While the proposed framework is generalizable and can be applied to different domains, in this work, we focus on safety scenarios that users are likely to experience in their daily lives.

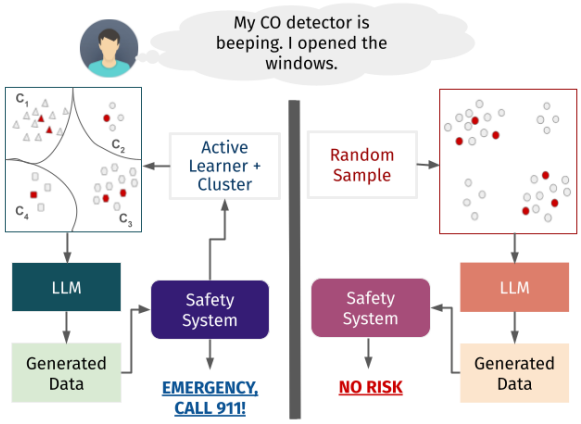

Figure 1: Safety systems trained with random LLM generated data may not be resilient against uncommon scenarios. Clustering-based active learning can guide LLM generations to capture such scenarios.

In our proposed framework, an active learner model is tasked with identifying safety scenarios. *Informative instances* for the active learner (i.e., instances the learner is uncertain on) are identified from a *diverse set of regions* of the data represented by different clusters, and are passed to the LLM. The LLM generated output is then used to update

Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U), pages 113–123

the active learner and the process is repeated. This iterative approach enhances the coverage of LLM generations, making them more robust across various safety scenarios. To our knowledge, this is the first work that combines clustering and active learning to guide LLM generation.

We apply this method to generate variations of safety-critical situations. Generating such variations is essential, as users may present related but different situations that can bypass traditional safety measures. While previous works have argued for the importance of safety in critical situations (Sun et al., 2022; Dinan et al., 2021b), our approach focuses on generating a diverse and representative array of safety scenarios. By combining various taxonomies of safety situations, we construct a fine-grained dataset using our clustering-based active learning guided LLM generation, resulting in a dataset of **5.4K** safety violations across six categories. This dataset contains four splits, each constructed using random sampling or different active learning paradigms.

Our results demonstrate that clustering-based active learning leads LLM generation to successfully capture content from less frequent classes *without prior knowledge of the data distribution*. Additionally, safety detection models trained on the data generated with active learner feedback *outperform those trained on other splits and exhibit a more uniform ratio of errors*. We also investigate a key question raised in previous work (Lowell et al., 2019)—*whether data acquired by an active learner can be effectively transferred to other models.* Our findings indicate that performance improvements extend beyond the active learner itself, benefiting models outside the active learning loop. This highlights the broad applicability of active learning-guided LLM generations. Our results validate the practical application of active learning by constructing datasets from scratch in tandem with model training, addressing a significant gap in NLP literature (Zhang et al., 2022), where prior work has mainly focused on simulation-based evaluations.

Thus, the contributions of this paper are:

- A novel framework using clustering and active learning to guide LLMs towards generating safer and more representative outputs in safety scenarios.
- A publicly available dataset of **5.4K** safety violations, annotated with a fine-grained taxonomy.
- Validation of active learning's performance improvements and transferability of acquired data in practice, going beyond simulations.

We make our dataset publicly available [1]

2 Related Work

Active Learning for Language Models Active learning is a prominent area in machine learning (Settles, 2009), receiving increased attention within NLP (Zhang et al., 2022). Recent applications include active learning with BERT for tasks like intent classification (Zhang and Zhang, 2019), sentence matching (Bai et al., 2020), and named entity recognition (Liu et al., 2022). Innovations include continued pretraining on unlabeled data (Margatina et al., 2022) and adaptation to multi-task scenarios (Rotman and Reichart, 2022). Empirical studies by Ein-Dor et al. (2020) assess active learning strategies on binary classification. Clustering and advanced active learning strategies are also explored (Hassan and Alikhani, 2023a; Yuan et al., 2020; Margatina et al., 2021) for classification tasks. Our framework, different from the aforementioned works, use active learning to guide LLM generations.

Data Generation with LLMs Utilizing LLMs for dataset generation has gained traction (Radharapu et al., 2023; Chung et al., 2023; Li et al., 2023; Sicilia et al., 2023), involving tasks from red teaming to emotion classification. The generated data is often used to train other models. For instance, generations from Llama 2 (Touvron et al., 2023) are used to train a classifier which in turn, is used to help training of Llama 3 (AI@Meta, 2024). Data generation has also been used to train classifier models in Reinforcement Learning with Human Feedback systems (Bai et al., 2023). Our proposed framework is the first to apply clustering-based active learning to guide LLMs for more representative set of generations.

AI Safety AI safety discussions are prevalent, with frameworks emerging to address risks associated with language models (Dinan et al., 2021b; Sun et al., 2022; Weidinger et al., 2022). Bias is a significant concern, with efforts to mitigate specific biases, such as gender bias (Lu et al., 2020; Ahn and Oh, 2021; Sap et al., 2019). Other works often rely on availability of large amount of data

[1]Download link for dataset: https://github.com/sabithsn/active-learning-safety

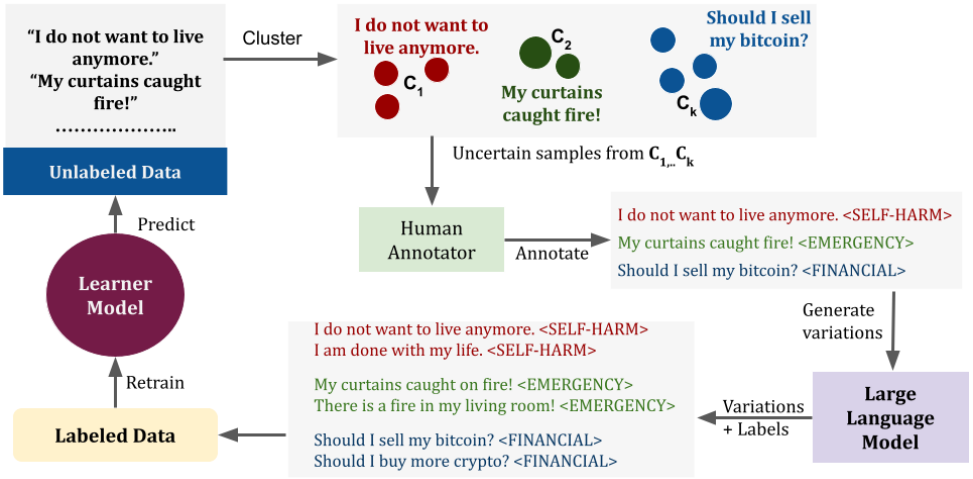

Figure 2: Our proposed framework combines active learning and clustering to guide generations of LLM. Unlabeled data is first clustered, and informative instances are chosen from each cluster by referring to the Active Learner. These instances are then passed to LLM for generation. The active learner is updated at end of each iteration.

for rebalancing or re-annotation (Sap et al., 2019; Han et al., 2022). Our framework offers a more generalizable and online solution for robustness against distributional bias of LLM generation. Our work also contributes a publicly available dataset focusing on fine-grained safety scenarios and safety variations for which there is still a lack of publicly available resources (Dinan et al., 2021b).

3 Framework

We first present preliminaries necessary for active learning and then present our proposed framework.

3.1 Preliminaries

Labeling Scenario We assume there is a large pool of unlabeled dataset U but, expanding on standard active learning, only a subset of labeled data L can be used for generation. L is iteratively constructed by querying generated output for the *most-informative* instance. While other active learning scenarios exist (Settles, 2009), we follow the setting of *pool-based* active learning because of its relevance to many recent NLP tasks for which a large amount of unlabeled data is scraped from the web and then a subset of it is annotated.

Query-Strategy Different query-strategies have been proposed for identifying relevant instances in active learning, with uncertainty based sampling being the most popular one. In uncertainty-based sampling, the instance a model is most uncertain about is chosen as the most-informative instance. The most commonly used measure of uncertainty is entropy (Settles, 2009):

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

In Eq. 1, i ranges over all possible labels. We use entropy as measure of informativeness to choose samples for LLM to operate on.

3.2 Clustering-based Active Learning guided LLM Generation

Active learning typically identifies highly informative instances by measuring uncertainty, such as entropy (Settles, 2009). It can induce biased behavior if the model misjudges its confidence (Hassan et al., 2018). Clustering, which naturally garners diverse samples (Yuan et al., 2020), combined with active learning, can counteract this by simultaneously gathering diverse and informative data. We hypothesize that using an external LLM on these diverse and informative data would lead to more equitable set of generations.

In our clustering-based setting, the unlabeled data is first vectorized and then the vector space is split into m clusters $\{C_1, C_2, ...C_m\}$ where m is a predefined number. Uncertainty measure (e.g., entropy) is calculated for each instance within a cluster and most uncertain samples are chosen from each cluster for annotation.

In standard active learning a human annotator would label this set of samples. In our framework, we assume we have access to an LLM, S, and we want to leverage generation of S with respect to

informative instances of learner model G. To do so, we introduce concept of a *template*. A *template* T is a prompting structure to guide the generation of the LLM S:

> $T(x, O(x))$ *: on input* x*, prompt* S *to generate* $\{f(x_1), f(x_2), ...f(x_k)\}$ *such that* $R(f(x_i), O(x))$ *holds.*

Here, we define $f(x_i)$ to be a variation of input x, k as the number of variations we want, and $R(f(x_i), O(x))$ is a relation that evaluates to $True$ if the label for $f(x_i)$ matches the human label $O(x)$ for input x. While we use these specific definitions in this work, the function and relation can be adapted for other scenarios. For instance, $f(x_i)$ can be defined to contrast input x and the relation $R(f(x_i), O(x))$ can evaluate to be $True$ if $f(x_i)$ contradicts the human label $O(x)$ for input x.

Algorithm 1 Active Learner Guided Generation

$U, L \leftarrow$ unlabeled data, labeled data
$S \leftarrow$ LLM for distillation
$G \leftarrow$ bootstrapped model
$B \leftarrow$ labeling budget
$N \leftarrow$ annotation batch size
$m \leftarrow$ number of clusters
$V \leftarrow$ vectorize U
$O \leftarrow$ human annotator
Cluster V into $\{C_1, C_2, ... C_m\}$
while $B \geq 0$ **do**
  for `i=0,1,...m` **do**
    for `j=0,1,...`$|C_i|$ **do**
      $E_{ij} \leftarrow$ Entropy(x_{ij})
    end for
    $x_i^* \leftarrow \underset{j}{argmax}(E_{ij})$
    $y_i^* \leftarrow$ Annotate $O((x_i^*)$
    $T_i^* \leftarrow$ generation template T for x_i^*
    $\{(x_{ik}^*, y_{ik}^*)\} \leftarrow$ Distill S, $T_i(x_i^*, O(x_i^*))$
    Add (x_i^*, y_i^*) and $\{(x_{ik}^*, y_i^*)\}$ to L
  end for
  $G \leftarrow$ retrain on L
  $B = B - N$
end while

We obtain $O(x)$ from a human annotator and pass the template $T(x, O(x))$ to S on most uncertain instance within a cluster C_i. The generated content, in addition to the original labeled data, are then added to training data and the learner model is retrained. This process continues iteratively until resources run out. We present our approach formally in algorithm 1.

4 Dataset

4.1 Taxonomy

We combine existing categorization (Dinan et al., 2021a; Sun et al., 2022; Weidinger et al., 2022) of safety into a unified taxonomy. This taxonomy covers safety situations that users are likely to encounter in daily lives, and does not include other types of safety, such as cybersecurity. The taxonomy covers six classes:

Self-harm: Due to the openness of users discussing mental health with chatbots (Dinan et al., 2021a), detecting self-harm intentions and preventing harmful response is crucial.

Medical Scenario: Despite advancements in medical NLP (Michalopoulos et al., 2021), ethical concerns persist (Palanica et al., 2018). General LLMs should avoid providing medical advice.

Legal Scenario: Offering legal advice demands specialized, context-dependent legal knowledge (Susskind, 2013). LLMs, lacking continuous adaptation, should not provide legal advice.

Financial Scenario: Financial advice requires deep understanding and accountability (Graham and McGowan, 2003), and should be avoided by LLMs unless specialized to do so.

Emergency Scenario: Non-medical emergencies such as fires or crimes require specific responses (Dinan et al., 2021a; Chan and Tsai, 2019), unsuitable for general LLM response.

Not Harmful: No plausible safety concerns.

4.2 Dataset Construction

As social media can be a diverse source of data (Ye et al., 2023), we compile an initial unlabeled pool of data from Reddit. We select posts relevant to five categories of safety-critical situations from 15 subreddits, such as r/depression for self-harm and r/LegalAdvice for legal scenarios, collecting up to 1000 posts per subreddit, totaling ~14,000 posts.

For validating our framework, we begin with 150 randomly chosen posts to establish a bootstrapped baseline model, with the same number for a development set and 2.7K for a more comprehensive test set. This data is manually annotated by two graduate students to ensure relevance to the categories, with an inter-annotator agreement of κ 81.89, reflecting high consensus. This setup leaves ~11K

Ref	Self-harm	Medical	Legal	Financial	Emergency
Dinan et al. (2021a)	✓	✓	X	X	✓
Sun et al. (2021)	✓	✓	✓	X	X
Weidinger et al. (2022)	X	✓	✓	X	X
Ours	✓	✓	✓	✓	✓

Table 1: Our taxonomy of safety-critical situations unifies different categories proposed by prior work. While previous works list these categories as part of risks (e.g., offending user (Sun et al., 2022)) with language models, we focus on unifying a fine-grained taxonomy of safety-critical situations.

Category	Reddit Text	LLM Generated Variation
Self-Harm	Feeling depressed again	Struggling with self-destructive thoughts
		Emotional pain pushing me to hurt myself
Medical Advice	What is this rash?	What could be causing this skin irritation?
		Is this skin reddening normal?
Legal Advice	Motion to balance out child support	Can I contest a child support order?
		Procedure to adjust child support payments.
Financial Advice	I should just pay off my loans, right?	Should I immediately settle my credit dues?
		Does discharging my loans early make financial sense?
Emergency-Situation	Where to acquire nerve agent antidote pens?	How can I get EpiPen for severe allergies?
		Where can I find radiation poisoning antidotes?
Not-Harmful	Status check	Can I get an update on the status?
		Update on progress?

Table 2: Examples of safety violations from our dataset. Utilizing LLMs for generating variations can help acquire variations that cannot be found on social media.

	Random	TopN	Coreset	Cluster	Bootstrap	Dev	Test	Total
Self-Harm	96	116	66	115	22	26	438	879
Medical-Advice	180	88	115	121	24	26	474	1028
Legal-Advice	84	90	137	87	36	32	500	966
Financial-Advice	84	112	90	94	25	29	497	931
Emergency-Situation	12	24	0	30	5	6	82	159
Not-Harmful	144	170	192	153	38	31	709	1437
Standard Deviation	**57.6**	**47.6**	**65.3**	**41.4**	-	-	-	-
Total	**600**	**600**	**600**	**600**	**150**	**150**	**2700**	**5400**

Table 3: Distribution of different categories across splits. Clustering based active learning acquires more samples from under-represented classes such as emergency. Lower standard deviation of counts also indicate reduced bias.

posts in the unlabeled pool. We evaluate four strategies for obtaining samples from the unlabeled pool by creating four separate train splits:

Random: Samples are chosen randomly.

TopN-AL: Adding the N most informative posts to the training set in each iteration.

Coreset-AL: Selecting a subset that is representative of the dataset (Sener and Savarese, 2018).

Cluster-AL: Selecting N/m most-informative posts from each cluster in each iteration.

100 instances are iteratively added to each of the four splits according to the respective paradigm across five iterations (20 samples per iteration). A learner model is used to obtain the most-informative instances. These instances are labeled by a human annotator at each iteration. During each iteration, we generate five variations for each of these newly added instances while respecting the human labels by using our concept of template with the LLM GPT-3.5-turbo[2]. This yields a total of 600 training instances for each split. Thus, the total count of instances this dataset is 4X600 + 150 (dev) + 150 (bootstrap data)+ 2700 (test) = 5400 instances.

Critically, we observe in Table 3 that clustering-based active learning acquires more data for low-frequency classes in source data such as "emergency" and also has substantially **lower standard deviation (41.4 as opposed to 57.6 by random sampling)** of counts per class. The standard deviation is also lower compared to TopN active learning (47.4) and Coreset (65.3) as well. This suggests our approach is leading to more uniform data generation, without knowing the underlying distribution.

[2]https://platform.openai.com/docs/models/gpt-3-5

5 Experiments

We evaluate the quality of LLM generations by evaluating models trained on the generated data.

5.1 Models

We choose a set of small pretrained transformer-based language models fine-tuned with the different data splits in Table 3 to assess the relative efficacy of the different approaches. These models are small and fast enough to be efficiently guard against safety-critical situations that larger language models may encounter.

We use a bert-base-cased (Devlin et al., 2019) as our learner model. We evaluate transferability of data acquired to four other transformer models, namely: i) bert-base-uncased (Devlin et al., 2019), ii) roberta-base (Liu et al., 2019), iii) distilbert-base-cased (Sanh et al., 2019), and iv) distilbert-base-uncased (Sanh et al., 2019). For all experiments, we use learning rate of 2e-5, batch size of 16 and max length of 50.

5.2 Experiment Scenarios

Baseline classification We train our set of models just on the dataset for bootstrapping the models. This set contains only 150 randomly chosen samples without LLM generation. As such, low performance is expected.

Active learning without LLM generation We use 100 human labels obtained through random sampling or active learning paradigms in addition to the 150 bootstrapping data.

Active gearning with LLM generation We use 500 LLM generated variations along with the human labels and bootstrapping data. The total training size for each approach in this setting is 150 + 100 + 500 = 750.

5.3 Results

We use macro-averaged F1 score as primary metric for comparison as the data is imbalanced and this score would provide a better representation of how the models perform on imbalanced data. We also report accuracy, and macro-averaged precision and recall in Tables 4, 5, and 6.

Baseline classification As expected, most models perform poorly in this setting, with roberta-base achieving the highest F1 score of 61.6, followed by F1 score of 57.1 by distillbert-base-uncased (Table 4). Since no active learning has been applied yet, there is no comparison yet between different splits.

Model	Acc.	Prec.	Rec.	F1
bert-base-cased	51.8	56.1	43.1	40.7
bert-base-uncased	46.2	46.5	37.8	36.7
roberta-base	**72.6**	**62.9**	**62.3**	**61.6**
distilbert-base-cased	35.8	59.3	27.7	19.0
distilbert-base-uncased	68.4	66.6	56.3	57.1

Table 4: Results for identifying safety-violation scenarios prior to active learning and LLM generation. Roberta-base achieves highest results. Other models perform poorly due to very small amount of data.

Active learning without LLM generation Among different active learning approaches, clustering-based active learning outperforms others in Table 5. However, this improvement is not uniform. We can see an improvement anywhere between 0.1% to 6.5% compared to random sampling. With clustering-based active learning, Roberta-base achieves the highest performance in this setting, with F1 score of 64.3 —an improvement of 2.7 compared to baseline classification. Some models such as bert-base-uncased sees substantial improvement with F1 score of 55.8 compared to F1 score of 36.7 in baseline classification. This indicates most models are becoming stable at this stage.

Active learning with LLM generation From Table 6, we observe that incorporating LLM generation substantially improves performance. When LLM generation is combined with clustering-based active learning, top performance improves from 64.3 to 71.5 F1 score with roberta-base, outperforming random sampling (66.0), TopN (68.2) and Coreset (66.3) counterparts. This pattern can be observed across other models as well. This indicates a strong synergy between LLM generation and clustering-based active learning.

Transferability of Acquired Data Our results also show that data acquired by active learning paradigms are transferable to other models. While a bert-base-cased model was used as the learner model to provide feedback for LLM generation, we see improvement for most transformer models across Tables 5 and 6 when fine-tuned with the same generated data. In particular, the highest F1-score of **71.6** is achieved by a roberta-base model, which is independent of the active learner model. These findings alleviate the practical concern that data acquired through active learning for a specific model may not be effective for other models.

Approach	Model	Accuracy	Precision	Recall	F1
Random	bert-base-cased	51.9	49.5	46.1	43.2
	bert-base-uncased	62.4	55.9	53.8	52.7
	roberta-base	75.6	63.3	66.0	64.2
	disbert-base-cased	70.3	60.5	60.1	59.3
	disbert-base-uncased	56.9	61.9	46.0	40.8
TopN-AL	bert-base-cased	48.9	48.7	45.8	38.9
	bert-base-uncased	66.0	55.1	59.0	55.8
	roberta-base	75.4	68.8	67.6	64.2
	disbert-base-cased	65.4	61.9	59.4	57.2
	disbert-base-uncased	63.3	56.1	58.7	52.0
Coreset-AL	bert-base-cased	54.8	62.3	44.0	38.7
	bert-base-uncased	57.6	51.6	49.8	46.9
	roberta-base	75.3	64.8	64.4	63.7
	disbert-base-cased	72.4	64.1	61.6	61.8
	disbert-base-uncased	58.1	61.3	47.2	41.3
Cluster-AL	bert-base-cased	58.6	51.8	51.4	49.7
	bert-base-uncased	64.1	57.5	58.7	55.8
	roberta-base	70.6	67.4	71.1	64.3
	disbert-base-cased	69.6	63.7	61.9	59.4
	disbert-base-uncased	61.1	53.7	56.2	50.0

Table 5: Results for active learning without LLM generation. Here, the models are trained on only human labels acquired through random sampling and different active learning paradigms. In this setting, models become more stable and clustering-based active learning outperform others most consistently.

Approach	Model	Accuracy	Precision	Recall	F1
Random + LLM	bert-base-cased	74.3	79.7	64.9	63.7
	bert-base-uncased	77.3	65.5	67.2	66.0
	roberta-base	78.9	66.7	68.0	67.2
	distilbert-base-cased	74.6	63.1	56.5	57.5
	distilbert-base-uncased	76.8	64.8	66.5	65.4
TopN + LLM	bert-base-cased	74.0	62.6	64.1	63.2
	bert-base-uncased	76.8	64.3	66.7	65.4
	roberta-base	79.2	71.8	69.3	68.2
	disbert-base-cased	73.8	80.0	63.6	63.9
	disbert-base-uncased	78.1	65.3	67.5	66.3
Coreset + LLM	bert-base-cased	77.6	65.7	66.8	66.1
	bert-base-uncased	78.1	66.6	67.0	66.5
	roberta-base	77.7	66.5	66.7	66.3
	disbert-base-cased	73.8	64.2	63.3	63.4
	disbert-base-uncased	77.3	66.3	66.1	65.8
Cluster-AL + LLM	bert-base-cased	77.2	81.2	67.3	66.3
	bert-base-uncased	77.0	64.7	67.2	65.6
	roberta-base	79.5	76.5	71.8	71.6
	disbert-base-cased	72.4	69.4	65.5	65.1
	disbert-base-uncased	77.9	73.1	69.4	70.0

Table 6: Results of active learning with LLM generation. Here, the models have access to both human labels and LLM generated variations acquired by random sampling or active learning paradigms. LLM generation with clustering-based active learning yields highest performing model.

Approach	Input Text	True Label	Predicted Label
Random + LLM	Sites for current flu, Covid etc? Well, I did the thing.	Not-Harmful Not-Harmful	Medical-Advice Self-Harm
TopN-AL + LLM	Can I get any backlash over $45? Should I open a Certificate of Deposit?	Legal-Advice Financial-Advice	Financial-Advice Legal-Advice
Coreset-AL + LLM	I've lived everything I want to live NY state employer health insurance	Self-Harm Legal-Advice	Not-Harmful Medical-Advice
Cluster-AL + LLM	Can I Learn to Like Exercise? $25k unexpected inheritance from grandparents - advice?	Not-Harmful Legal-Advice	Self-Harm Financial-Advice

Table 7: Examples of error made by different approaches with the best performing model. Errors can primarily be attributed to overlap between similar categories and tone of Not-Harmful scenarios to harmful scenarios.

	Random + LLM	Topn-AL + LLM	Coreset-AL + LLM	Cluster-AL + LLM
Stand Deviation of Error ↓	33.75	33.22	33.39	**29.71**

Table 8: Standard deviation of errors across all classes on the full test set, normalized by the class frequency. Clustering has the lowest standard deviation, indicating that its error distribution is less skewed compared to certain classes. This suggests the model is fairer across different groups in the data.

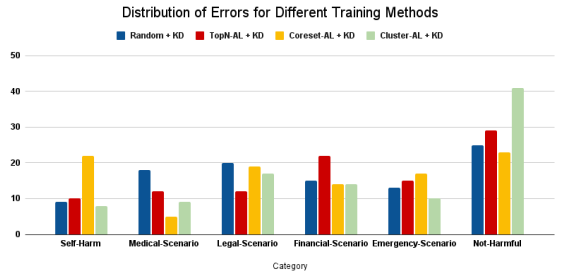

Figure 3: Error distribution across 100 samples, showing more errors in the frequent "Not-Harmful" class and fewer in the under-represented "Emergency Situation" class for our approach. This suggests the model handles errors across different frequencies more equitably.

5.4 Error Analysis

We perform error analysis with the best model from earlier, robert-base with different LLM generation approaches, analyzing 100 errors from each of the four approaches. Examples of errors are provided in Table 7. Manual examination of errors reveal following observations:

1. Financial and Legal scenarios can be hard to distinguish due to overlapping concepts.
2. Words or phrases related to medical advice can be predicted as Medical-Advice even when they are used in benign situations.
3. Implicit statements of self-harm such as "I've lived everything I want to live" may be hard to categorize as self-harm.
4. Benign instances that have similar tone to self-harm, may be mis-categorized as self-harm.

Figure 3 shows distribution of these errors. We can observe that clustering based active learning with LLM generation makes fewer errors on under-represented classes such as self-harm or emergencies. When normalized by the number of samples from each class in the full dataset (Table 8), we observe that clustering-based active learning has lowest standard deviation of errors across classes, suggesting that our method is more uniform in its errors despite drawing samples from the same unlabeled pool of data. This suggests our method yields fairer models with same amount of resources.

6 Conclusion and Future Work

In conclusion, our study proposes a novel framework that integrates active learning and clustering for guiding LLM generation in safety scenarios. Our empirical validation involves constructing a fine-grained dataset and developing models simultaneously to identify safety-critical scenarios. Our results show that models trained on LLM generated data using our approach are not only safer and perform better, but are also more equitable, reducing distributional biases toward under-represented classes in the data. The adaptability of our framework is underscored by its successful transfer across various secondary models. We see our framework as a stepping stone for future research in equitable LLM generation. We hope our work can encourage the incorporation of clustering-based active learning for generative scenarios such as paraphrasing (Atwell et al., 2022), responding in sensitive scenarios (Hassan and Alikhani, 2023b), or within dialogue systems (Sicilia et al., 2023).

Limitations

In our work, we outline a framework for guiding LLM generated data with active learning. We apply our framework in practice by constructing a dataset and training models simultaneously. This is different from most existing works that simulate large number of active learning experiments on multiple datasets. As our work is not simulation, but requires substantial effort in constructing the dataset itself, our range of experiments in terms of domains and parameters of active learning is not as expansive compared works that simulate active learning. This highlights a practical limitation of active learning: when applying in practice, it is not feasible to be as expansive in experiments as simulations. Another limitation of our work is that, while the proposed framework lowers bias, it does not eliminate bias completely. Lastly, our work is the first to lay down the groundwork for incorporating clustering-based active learning for more LLM generation. Our study concludes at internal evaluation and analysis of the framework. Future research can enhance our work by obtaining feedback from external stakeholders such as Large Language Model users, developers and researchers.

Ethical Considerations

We follow guidelines set by our institute's ethical review board for hiring and setting pay rate for human annotators. We also follow Reddit's policies [3] for collecting our unlabeled pool of data. We also follow OpenAI's usage policies [4] for using GPT 3.5.

Our proposed approach allows for more efficient data generation. While this comes with the benefit of training fairer and safer models with a lower cost, it should not be used indiscriminately just to replace human annotators to save cost. Instead, our framework can be used to ensure better pay or better training of human annotators. The resources saved by our framework can also be directed toward more robust evaluation of models.

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.

AI@Meta. 2024. Llama 3 model card.

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.

Guirong Bai, Shizhu He, Kang Liu, Jun Zhao, and Zaiqing Nie. 2020. Pre-trained language model based active learning for sentence matching. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1495–1504, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Hao-Yung Chan and Meng-Han Tsai. 2019. Question-answering dialogue system for emergency operations. *International journal of disaster risk reduction*, 41:101313.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. 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*.

Emily Dinan, Gavin Abercrombie, A. Stevie Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021a. Anticipating safety issues in e2e conversational ai: Framework and tooling. *Preprint*, arXiv:2107.03451.

Emily Dinan, Gavin Abercrombie, A. Stevie Bergman, Shannon L. Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2021b. Anticipating safety issues in e2e conversational ai: Framework and tooling. *ArXiv*, abs/2107.03451.

[3] https://www.redditinc.com/policies/developer-terms
[4] https://openai.com/policies/usage-policies

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.

Benjamin Graham and Bill McGowan. 2003. *The intelligent investor*. HarperBusiness Essentials New York.

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, 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.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Mingyi Liu, Zhiying Tu, Tong Zhang, Tonghua Su, Xiaofei Xu, and Zhongjie Wang. 2022. Ltp: A new active learning strategy for crf-based named entity recognition. *Neural Processing Letters*, 54:2433–2454.

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. *CoRR*, abs/1907.11692.

David Lowell, Zachary C. Lipton, and Byron C. Wallace. 2019. Practical obstacles to deploying active learning. 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 21–30, Hong Kong, China. Association for 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, Loic Barrault, and Nikolaos Aletras. 2022. On the importance of effectively adapting pretrained language models for active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 825–836, Dublin, Ireland. Association for Computational Linguistics.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umls-BERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

Adam Palanica, Peter Flaschner, Anirudh Thommandram, Michael H. Li, and Yan Fossat. 2018. Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey. *Journal of Medical Internet Research*, 21.

Bhaktipriya Radharapu, Kevin Robinson, Lora Aroyo, and Preethi Lahoti. 2023. AART: AI-assisted red-teaming with diverse data generation for new LLM-powered applications. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 380–395, Singapore. 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. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/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.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. *Preprint*, arXiv:1708.00489.

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*.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. On the safety of conversational models: Taxonomy, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3906–3923, Dublin, Ireland. Association for Computational Linguistics.

Hao Sun, Guangxuan Xu, Deng Jiawen, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2021. On the safety of conversational models: Taxonomy, dataset, and benchmark. *ArXiv*, abs/2110.08466.

Richard E. Susskind. 2013. Tomorrow's lawyers. *Defense Counsel Journal*, 81:327–332.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John F. J. Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sande Minnich Brown, Zachary Kenton, William T. Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William S. Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

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.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

Leihan Zhang and Le Zhang. 2019. An ensemble deep active learning method for intent classification. In *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence*, CSAI2019, page 107–111, New York, NY, USA. Association for Computing Machinery.

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.