ToxiCraft: A Novel Framework for Synthetic Generation of Harmful Information

Zheng Hui♣ ♡, **Zhaoxiao Guo**♣ †, **Hang Zhao**♣, **Juanyong Duan**♣, **Congrui Huang**♣

♣Microsoft, ♡Columbia University, †Tsinghua University
{zackhui, hang.zhao, juanyong.duan, conhua}@microsoft.com,
zh2483@columbia.edu, guozx22@mails.tsinghua.edu.cn
Correspondence: juaduan@microsoft.com

Abstract

In different NLP tasks, detecting harmful content is crucial for online environments, especially with the growing influence of social media. However, previous research has two main issues: 1) a lack of data in low-resource settings, and 2) inconsistent definitions and criteria for judging harmful content, requiring classification models to be robust to spurious features and diverse. We propose **ToxiCraft**, a novel framework for synthesizing datasets of harmful information to address these weaknesses. With only a small amount of seed data, our framework can generate a wide variety of synthetic, yet remarkably realistic, examples of toxic information. Experimentation across various datasets showcases a notable enhancement in detection model robustness and adaptability, surpassing or close to the gold labels. Dataset is available at this GitHub repository.

⚠ This paper has instances of hateful and offensive language to serve as examples.

1 Introduction

The 21st-century digital realm presents vast connectivity and information exchange opportunities alongside the challenge of widespread harmful content like cyberbullying, hate speech, and misinformation, impacting individuals and communities negatively. As such, the development of effective mechanisms for detecting and mitigating harmful content is of paramount importance (Breitfeller et al., 2019; Casula and Tonelli, 2023; Plaza-del arco et al., 2023).

The emergence of the Transformer-based model (Vaswani et al., 2017) led to the development of complex models capable of identifying toxic content with remarkable accuracy. However, the effectiveness of these models hinges on the quality and diversity of the datasets used for their training (Banko and Brill, 2001). Traditional datasets,
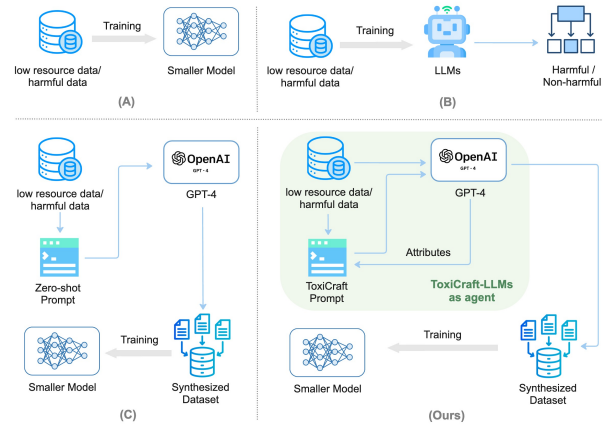

Figure 1: Harmful detection approaches

often curated manually, tend to lack the diversity required to cover the multifaceted nature of harmful content. Consequently, while these models excel at recognizing explicit instances of toxicity, they struggle when faced with subtler forms (Casula and Tonelli, 2023).

Many datasets containing harmful content are typically sourced from social media platforms like Twitter or online forums (Davidson et al., 2017; Grimminger and Klinger, 2021a; Waseem and Hovy, 2016; Sachdeva et al., 2022). However, these datasets often exhibit significant class imbalances, particularly concerning specific types of toxic language. Training a model to detect specific harmful content or opting for smaller models due to resource or latency constraints requires substantial amounts of human-labeled data. Unfortunately, such data are scarce in downstream tasks and expensive to annotate (Breitfeller et al., 2019; Juuti et al., 2020). Additionally, smaller or low-resource datasets face challenges like limited linguistic diversity and a higher risk of overfitting (Klubička and Fernández, 2018). Privacy concerns also emerge when using social media data obtained without user consent for research purposes (Casula and Tonelli,

16632

2023). Furthermore, dataset decay is also a notable issue, especially with online social media, including abusive content, which are often deleted over time (Casula and Tonelli, 2023)[1].

Recent advancements in large language models (LLMs) such as the GPT series (Brown et al., 2020) have prompted researchers to explore their potential for generating synthetic data (Yoo et al., 2021; Ye et al., 2022b; Gao et al., 2023). While LLM-based data augmentation often improves model performance, findings are mixed regarding whether synthetic data generated by LLMs can consistently match the effectiveness of models trained on real-world, human-annotated data (Li et al., 2023). LLMs have also shown promise in generating synthetic data for training harmful content detection models. However, work in this area remains limited (Sen et al., 2023) and much of the existing research is relatively outdated (Juuti et al., 2020; Geet d'Sa et al., 2021; Liu et al., 2020; Geet d'Sa et al., 2021; Wullach et al., 2021). Moreover, the effectiveness of LLMs in this domain is hindered by their difficulty with subjective tasks and their tendency to amplify biases inherent in human-annotated training data. These biases stem from inconsistent standards used in defining 'gold' labels, and LLMs often struggle to capture the complex nuances and diversity of human communication in highly subjective tasks.

To address these challenges, we introduce the ToxiCraft Framework, which enhances robustness and reduces bias in synthetic data generation. The ToxiCraft Framework is specifically designed to produce higher-quality synthetic data that better supports model training. An example of the process is illustrated in Figure 1. In Figure 1, (A) is direct training where a smaller model is trained on a limited dataset, that may affect model quality. (B) shows the use of large language models to do harmful detaction. (C) illustrates Zero-shot Synthesis, where use LLMs to generate a synthetic dataset via zero-shot prompts, used to train smaller models without additional data. Our ToxiCraft method (D) employs carefully crafted prompts and attributes, utilizing GPT-4 as the backbone to generate a synthetic dataset for training more refined smaller models. A more detailed explanation of our approach can be found in Section 3. To summarise, our work

has the following key contributions:

- We identify a limitation of traditional simple zero-shot prompting methods in data synthesis which leads to a lack of diversity.

- Enhances low-resource harmful information datasets to address the pervasive issue of harmful dataset sources decay, making the enhanced datasets publicly available to ensure reproducibility and broader utility.

- We introduce a novel data synthetic framework, mitigating biases and ensuring a closer reflection of real-world complexities and demonstrating its efficacy in generating the diverse and high-quality data.

- Applications of the synthesized data in fine-tuning smaller models, highlight the practical value of LLM generated data.

2 Related Work

2.1 Harmful Information Detection

Early studies established classifiers to detect harmful information using neural network models (Zhang et al., 2018) or word embedding methods (Kshirsagar et al., 2018). In recent years, models based on the Transformer architecture have demonstrated remarkable capabilities, prompting researchers to explore further. Rajput et al. (2021) conducted research on the ETHOS hate speech detection dataset, comparing classifiers' performance in hate speech detection by replacing or integrating word embeddings (fastText, GloVe, or FT + GV) with BERT embeddings. Aluru et al. (2020) contrasted simple models (such as LASER embeddings with logistic regression) and BERT models in scenarios with scarce and abundant linguistic resources. Lin et al. (2024) generated explanations through multimodal debates between LLMs, enhancing the transparency and explainability of harmful meme detection. da Silva Oliveira et al. (2024) validated the effectiveness of ChatGPT in identifying harmful Spanish-language speech. Our ToxiCraft framework extends these developments by leveraging LLMs not only for direct detection but also for enhancement of the training data pool through synthetic data generation. Unlike methods that rely solely on existing data, ToxiCraft enriches the dataset, which is particularly critical in the ever-evolving domain of online content where new forms of harmful expression continuously emerge.

---

[1] Twitter's policy only allows the display of post IDs rather than the text itself, making content inaccessible over time, often due to deletion by either the user or Twitter, particularly in cases involving harmful information.

2.2 Large Language Models

Transformer (Vaswani et al., 2023) architecture has revolutionized the field of NLP (Chang et al., 2023). Large Language Models (LLMs) based on the Transformer architecture are pre-trained on extensive corpora, resulting in models with vast parameter scales and exceptional learning capabilities. The BERT model (Devlin et al., 2019a) utilizes the bidirectional encoder from Transformer to process input text, generating rich context-aware word embeddings, with variants including ALBERT (Lan et al., 2020), RoBERTa(Liu et al., 2019), and DeBERTa(He et al., 2021). OpenAI have launched the GPT series of large language models based on Transformer's decoder, including GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), Instruct-GPT (Ouyang et al., 2022), and GPT-4 (OpenAI et al., 2024). These large language models are garnering increasing research interest, not only exhibiting outstanding performance in a broad range of tasks in the natural language understanding sector—such as sentiment analysis (Scaria et al., 2023; Fei et al., 2023), text classification (Peña et al., 2023) but also demonstrating remarkable capabilities in natural language generation tasks like summarization (Wang et al., 2023b), translation (Wang et al., 2023a), and question answering (Yan et al., 2021).

2.3 Data Synthesis

Traditional data synthesis methods have employed techniques ranging from synonym replacement to token-level manipulations, as exemplified by the works of (Zhang et al., 2016) and (Wei and Zou, 2019). These methods, while useful, offer limited contextual understanding and diversity. The advent of translation models (Fadaee et al., 2017) and masked filling (Kumar et al., 2020) brought improvements in maintaining semantic consistency, yet they still fall short in generating the context-rich synthetic data required for complex tasks such as harmful content detection. In contrast, ToxiCraft leverages the latest advancements in LLMs for data synthesis, transcending the limitations of earlier approaches by producing contextually nuanced and varied synthetic instances. Unlike methods that require fine-tuning on annotated data, which incurs substantial human labor costs (Yang et al., 2020; Mohapatra et al., 2021; Kumar et al., 2021), ToxiCraft efficiently synthesizes data without intensive manual effort. In comparison to zero-shot frameworks like ZEROGEN (Ye et al., 2022b), SuperGen (Meng et al., 2022), and PROGEN (Ye et al., 2022a), which generate datasets from scratch or suffer from low information content and redundancy, ToxiCraft's approach is designed to yield high-quality synthetic data that is both diverse and relevant to the seed data, thus ensuring its applicability to real-world tasks. In the few-shot learning domain, while Yu et al. (2023) rely on diverse attribute prompts and PromDA (Wang et al., 2022) along with MSP (Chen et al., 2023) utilize soft prompts for data diversity and optimization, ToxiCraft differentiates itself by not only enhancing the diversity but also focusing on the generation of synthetic data that closely aligns with the intricacies of harmful content. The integration of attribute prompts within ToxiCraft's framework ensures that the synthesized data captures a broad spectrum of harmful content, effectively addressing both the volume and variety required for robust model training.

3 Methodology

In this section, we commence by defining the problem statement and subsequently introduce the ToxiCraft framework. ToxiCraft is designed as a versatile and effective few-shot learning framework that operates through dataset generation. ToxiCraft framework are illustrated in Figure 2.

3.1 Problem Statement

In the context of online content moderation, particularly the detection of harmful information, the problem can be modeled as a supervised classification task. Given an input content sequence x, represented as tokens $x = [x_1, x_2, \ldots, x_{|x|}]$, the goal is to predict a binary output y, where $y \in \{0, 1\}$ indicates whether the content is harmful ($y = 1$) or not ($y = 0$). The dataset D consists of pairs of content sequences and their corresponding labels:

$$D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$$

where N represents the number of examples in the dataset. The major challenge in this domain is the scarcity of labeled examples of harmful content, denoted as D_h:

$$D_h = \{(x^{(i)}, y^{(i)}) \in D : y^{(i)} = 1\}$$

Due to the sparse nature of D_h, there is a need for effective methods to augment the dataset, especially enhancing the representation of the minority class, which is harmful content in this scenario.
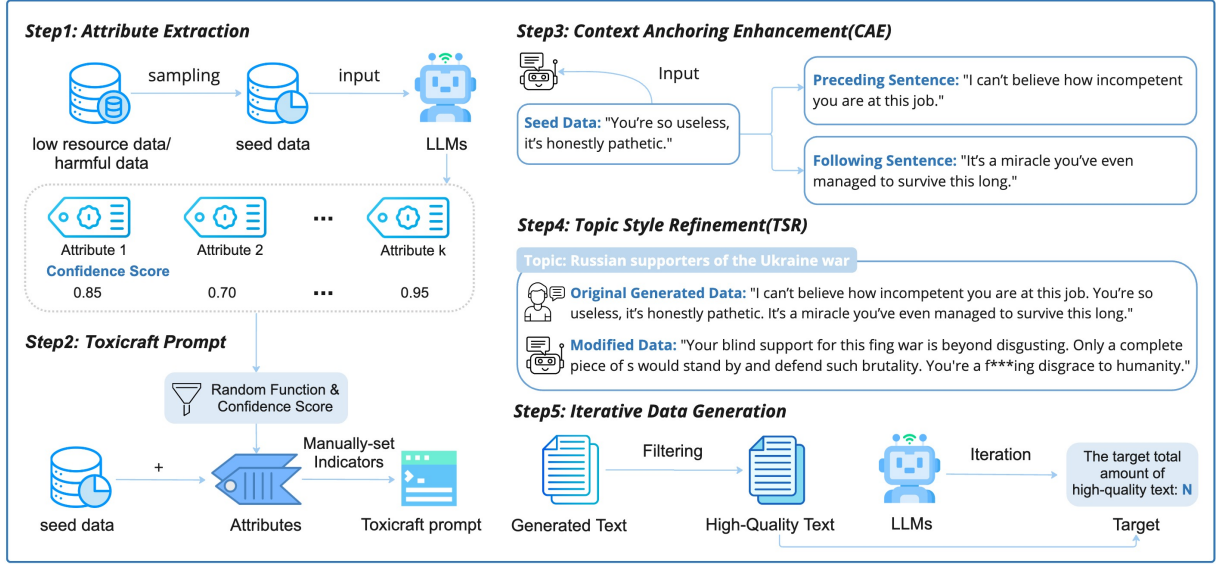
Figure 2: Our **ToxiCraft** Framework Diagram, CAE & TSR step is elaborating on transforming the initial data into multiple versions by altering the context while maintaining the core message or change the topic

To address the limitations posed by scarce harmful content samples, we propose a data augmentation method - **ToxiCraft** that involves synthesizing new content instances using LLMs. ToxiCraft utilizes LLMs to augment the dataset with synthetic instances, enhancing the representational diversity of harmful content. The data augmentation function, defined by:

$$\text{Aug}(x) = \{x' : x' = \Phi(\text{LLM}, x, \epsilon), \epsilon \sim \mathcal{P}(\epsilon)\}$$

Here, Φ denotes the ToxiCraft framework which is described in Section 3.2. And ToxiCraft uses selected seed data to guide the synthesis, ensuring that the generated content is relevant and varied. ϵ represents a noise vector drawn from a probability distribution $\mathcal{P}$, introducing variations in the synthesized outputs to prevent overfitting to the seed samples. Seed data S for ToxiCraft are selected from D_h based on a random sampling process, which aims to capture a diverse array of harmful content types. The selection function $S(d, n)$ picks n samples from d using a probability distribution tailored to emphasize less represented content:

$$S(D_h, n) = \{(x^{(i)}, y^{(i)}) : x^{(i)} \in D_h,$$
$$i \in \text{random indices from } D_h\}$$

where n is the number of seed data chosen to guide the LLM in data generation. Details on the selection process and the impact of seed data size are

provided in Section 4. The augmented dataset D' is then:

$$D' = D \cup \{(x', y) : x' \in \text{Aug}(x), S(D_h, n) \in D_h\}$$

After that, we employ a smaller model, e.g., BERT, fine-tuned on both D and D'. The training objective is to minimize the loss function $\mathcal{L}(\theta)$ across the augmented dataset:

$$\mathcal{L}(\theta) = \frac{1}{|D'|} \sum_{(x,y) \in D'} \ell(h_\theta(x), y)$$

where ℓ represents a loss function appropriate for binary classification tasks, such as cross-entropy. The evaluation of the ToxiCraft framework using the smaller model involves several key metrics:

Effectiveness: Measured by the accuracy and Macro F1-score of the model on a held-out test set T, not seen during training or augmentation:

$$\text{Macro F1-Score}(\theta) = \frac{1}{K} \sum_{k=1}^{K} F1_k(\theta)$$

where $F1_k(\theta) = 2 \times \frac{\text{Precision}_k(\theta) \times \text{Recall}_k(\theta)}{\text{Precision}_k(\theta) + \text{Recall}_k(\theta)}$ and K is the number of classes, which in this binary classification context, $K = 2$.

Robustness: In our setting, robustness is evaluated using cross-dataset evaluation, where data augmented on one dataset is used to evaluate performance on other datasets. Specifically, robustness

is measured by the model's performance consistency across different datasets $\mathcal{D}_i$, with data augmented from one dataset $\mathcal{D}_j$ and evaluated on another dataset $\mathcal{D}_k$. The variance of the model's loss across these datasets captures the robustness:

$$\text{Robustness}(\theta, \mathcal{D}) = \text{Var}_{\mathcal{D}_k}(\ell(h_\theta(\mathcal{D}_k), y_k))$$

where h_θ represents the model parameterized by θ, $\ell(\cdot)$ is the loss function, and $\mathcal{D}_k$ are the instances from different datasets. Var captures the variance in performance across these datasets, reflecting the robustness of the model when trained on data from $\mathcal{D}_j$ and tested on others. We also provide a pseudocode in Appendix A.

3.2 ToxiCraft Framework

In the ToxiCraft framework, we employed LLMs as agents to analyze seed data containing harmful information.

Attribute Extraction The LLMs' task is to identify the attributes within this data to determine if the content is harmful. This concept was inspired by the research of Yu et al. (2023); however, unlike their approach, we did not manually filter the attributes generated by GPT. Instead, we retained these attribute tags generated by the model itself and cataloged all specific groups or themes indicating harm in the seed data, collectively termed the 'Harmful Themes Index'. The advantage of automated attribute tagging is its capacity to process and analyze data rapidly and on a large scale, significantly enhancing processing efficiency. To further enhance the accuracy of generated attributes and the robustness of the model, we introduced a new step where LLM provides a confidence score when generating attributes. Based on this confidence score, we decided whether to retain the attribute using a simple random function.

Even with a high confidence score, an attribute may be randomly discarded by a small chance, which could increase the robustness of the decision-making process, reduce the risk of cumulative errors, and avoid over-reliance on the model's singular judgment. In this process, we considered the possibility of errors in automated attribute tagging but decided not to take model-based corrective measures. Instead, we chose to maintain a simple and efficient system to quickly respond to and update changes in the dataset. Although this approach is simple, it can maintain a certain level of accuracy and consistency without sacrificing efficiency.

ToxiCraft Prompt Next, we added manually-set indicators to this data with Attribute Extraction by GPT (we call this process as 'ToxiCraft Prompt') and input 10% of the data extracted randomly from the seed data pool together with 'ToxiCraft Prompt' to GPT. Manually-set indicators include the intensification or weakening of tone, whether to increase swear words, whether to use irony, country and time. Each value will be randomly selected or masked. Thus, it is possible that none of the manually-set indicators are added, but at most only one answer will be selected from each indicator. For example, the country could be randomly set as the United States and the year as 2023.

CAE and TSR According to the research by Pavlopoulos et al. (2020) and (Geet d'Sa et al., 2021), high-quality context can help better detect harmful data. Considering the inherent length limitations of the dataset (primarily from a harmful dataset on Twitter before 2017 with a limit of 140 characters (about 20-35 words) and after 2017 with a limit of 280 characters), we innovatively generated context-based preceding and succeeding text. We employed a technique called 'Contextual Anchoring Enhancement', using dropout to randomly abandon the preceding or succeeding text or retain them all to enhance the model's robustness. Furthermore, we also checked the generated data and used GPT to evaluate their quality. Among the result data, we selected the top-performing 10%, applied 'Thematic Style Refinement', transforming the themes recorded in the 'Harmful Themes Index', and added them to the seed data. We repeated these steps until generating M results controlled by hyperparameters, while the number of data generated per batch using 10% of seed data K was also controlled by hyperparameters. The total ToxiCraft framework genrated data are N.

ToxiCraft effectively utilizes the Chain of Thought (COT) concept to gradually generate high-quality training data. Additionally, we compared the dataset generated by ToxiCraft with the dataset generated only using simple COT prompts, and detailed experimental results will be presented in Section 4.

4 Experiment Setup

4.1 Datasets and Processing

In this section, we outline the data employed in our experiments. We first introduce the four datasets containing harmful information in Section 4.1.1.

Following this, we consolidate these datasets and present statistics in Section 4.1.2.

4.1.1 Harmful Information Datasets

Automated Hate Speech Detection (AHSD) described by Davidson et al. (2017), comprises 24,802 tweets extracted using a hate speech lexicon. These tweets are manually categorized by CrowdFlower workers into three distinct groups: hate speech, offensive but not hate speech, and neither, with an intercoder agreement of 92%.

Towards the Political Opponent (HTPO) described by Grimminger and Klinger (2021b) encompasses 3,000 tweets related to the 2020 U.S. Presidential election, annotated for both stance detection (favorable or against candidates) and the presence of hateful language, enabling nuanced analysis of political discourse and sentiment.

Hate Speech and Offensive Content Identification (HASOC) described by Mandla et al. (2021), is part of the Hate Speech and Offensive Content Identification track, consists of 17,657 tweets in Hindi, German, and English, annotated for hate speech and offensive content with three classification levels: presence of hate or offensive content, type of offense (hate, offensive, or profane), and the nature of the insult (targeted or untargeted).

Call me sexist, but (CMS) described by Samory et al. (2021), consists of 6,325 instances drawn from Twitter and includes a comprehensive annotation for expressions of sexism, where each entry is independently labeled by five coders into categories based on content and phrasing, such as sexist, maybe-sexist, civil, and uncivil.

It's important to mention many datasets such as **AIRAIT** (Waseem and Hovy, 2016), only their tweet IDs and their labels were publicly available. Regrettably, a substantial portion of tweets (>= 40% positive datas) were inaccessible through the X API.

4.1.2 Data Processing and Splits

Our data preprocessing pipeline prioritizes identifying harmful content in English. Thus, we exclude any tweets in languages other than English and verify language using a Roberta-based language identifier. Additionally, to prevent overlap between similar datasets, we remove near-duplicated entries through normalization, ignoring duplicate entries and removing URLs and mentions. The majority of collected datasets focus on hate speech, and in some cases, offensive speech, which we consider harmful. For datasets with a binary hate classification task or a more detailed classification like CMS, where all "hate" subclasses are treated as one, we categorize them as harmful. Datasets focusing on specific types of hate speech, such as sexism, are also considered harmful. Moreover, datasets containing offensive speech are likewise classified as harmful in our case. Finally, in datasets like AHSD and HASOC, where a distinction between hate, offensive, or profound speech exists, we include entries labeled as hate speech or offensive, considering them as harmful. Entries labeled as normal or not-hateful are categorized as not harmful.

Table 1 shows the final numbers of data and classes after processing. Datasets divided into training, validation, and testing sets in a ratio of 7:1:2, respectively.

Dataset	Harmful	Non-Harmful
AHSD	1200*	4081
HTPO	351	2508
HASOC	1200*	4292
CMS	1203	4000*

Table 1: Dataset Breakdown: Harmful vs. Nonharmful Content Numbers in AHSD, HTPO, HASOC, CMS. * indicates that dataset sizes have been downsized to simulate reality low-resources setting.

4.2 Baselines

We contrast our method with various baseline approaches, including those that utilize LLM-powered data augmentation techniques.

All Gold Labels: This method utilizes the entire gold-labeled training dataset for training the models. It serves as a baseline by providing comprehensive data exposure, which allows the models to learn from a complete range of examples in the dataset.

Seed Gold Labels with In-context Learning: This approach expands the seed data by generating new data instances from the seed data samples, where samples for in-context learning and target-context selection are randomly picked. It tests the model's ability to generalize from an enhanced but limited dataset, simulating training under resource-constrained conditions.

ToxiCraft: Our proposed method that enhances seed data with synthetically generated data, expanding the training dataset by adding new examples that mirror the characteristics of the seed data. This

Methods/Seed Count		AHSD		HTPO		HASOC		CMS	
		BERT	RoBERTa	BERT	RoBERTa	BERT	RoBERTa	BERT	RoBERTa
Gold Labels	-	86.0	88.0	**74.0**	69.0	58.2	**57.3**	85.2	85.4
In-context Learning w/ Seed	50	34.9	35.2	23.5	22.8	8.1	10.4	34.4	37.5
	100	47.8	49.5	48.2	46.3	17.9	20.5	49.3	52.0
	150	65.2	68.1	54.6	55.1	21.4	21.4	58.9	57.4
	200	63.2	68.0	59.2	58.7	32.6	31.1	63.8	66.2
ToxiCraft w/ Seed (Ours)	50	62.4	63.7	40.6	38.9	15.2	16.0	53.2	55.9
	100	74.7	73.9	47.8	44.3	27.3	28.2	75.6	70.4
	150	85.5	87.1	58.0	59.3	44.1	43.6	<u>86.0</u>	84.8
	200	<u>**89.0**</u>	<u>**88.4**</u>	70.3	<u>**70.5**</u>	<u>**58.9**</u>	56.1	<u>**86.3**</u>	<u>**86.5**</u>

Table 2: Comparison of Model Performances Across Different Databases and Data Generation Methods. The comparison parameter is MacroF1. Results outperformed training on gold label are underlined while the best performance on each dataset are bolded. Some baseline results come from Antypas and Camacho-Collados (2023).

method not only expands the quantity of available training data but also diversifies the types of training examples. It aims to overcome data scarcity and improve model robustness by broadening the training scenarios.

4.3 Implementation Details

We employed GPT-4 (OpenAI et al., 2024) as our agent model. For generations, in our experiment, we set N to be 1000, K to be 100, and M to be 3. For smaller-scale classification tasks, we utilized two language models of a reasonable size: BERT-base (Devlin et al., 2019b) and RoBERTa-base (Liu et al., 2019) as downstream models. The implementations provided by Hugging Face (Wolf et al., 2020) were utilized for training and evaluating all language models. The fine-tuning parameters are listed in B, and we used macro-F1 to report the evaluation scores, ensuring consistency across models and tasks.

5 Experimental Results

Main Results We conduct experiments on two different data augmentation scenarios and report the results of training data augmentation in Table 2 and the LLM generated successful rate results in Table 3. As shown in table, using the complete gold-labeled dataset, generally sets a high benchmark, notably on the AHSD and CMS datasets, indicating that comprehensive data access typically results in better model performance. Conversely, 'In-context Learning with Seed' shows a performance increase with larger seed sizes, but significantly underperforms with smaller seeds (50, 100), indicating challenges in training models with limited data. Our

Dataset	AHSD	HTPO	HASOC	CMS
Target Size	1000	1000	1000	1000
Success (%)	78.6	52.2	75.2	69.1

Table 3: Target Size and Success Rate of GPT-4 Generated Synthetic Data Across Datasets

Toxicarft substantially outperforms or comparable with all baselines across different settings demonstrating the effectiveness of our approach. Notably, our approach, which enhances seed data with synthetically generated labels, demonstrates substantial performance gains, especially with larger seeds (150, 200). These results approach or even surpass those from training with all gold labels, highlighting the potential of synthetic data to effectively supplement training datasets and enhance learning outcomes. The lower performance across all methods in the HTPO dataset, which focuses on political harmful content, suggests that such content poses additional complexities and nuances that are challenging for models to learn. Our framework not only improves as more seed data is available but also illustrates that well-generated synthetic data can serve as a robust tool for augmenting training datasets, thus potentially reducing reliance on extensive manually labeled data. In term of models, RoBERTa generally outperforms BERT, especially in higher data regimes (e.g., "ToxiCraft with Seed Data" at seed 200 across multiple datasets). This could be attributed to RoBERTa's more robust pre-training, which might be better at handling the nuances introduced by synthetic data augmentation.

Analysis on Generated Success Rate The high

success rate of 78.6% for the AHSD dataset suggests that GPT-4 excels in generating data that closely matches the characteristics and complexities of this particular dataset. It's important to note that both the success rate and the macroF1 generated from a small amount of seed data for AHSD are high, the possibility of a data breach cannot be ruled out. Conversely, the lower success rate of 52.2% for the HTPO dataset, which focuses on political harmful content, highlights the challenges faced in synthetic data generation when dealing with nuanced and sensitive content. It's possible that the GPT series may have been specifically tuned for political topics, which could explain the difficulties encountered in generating synthetic data for this dataset.
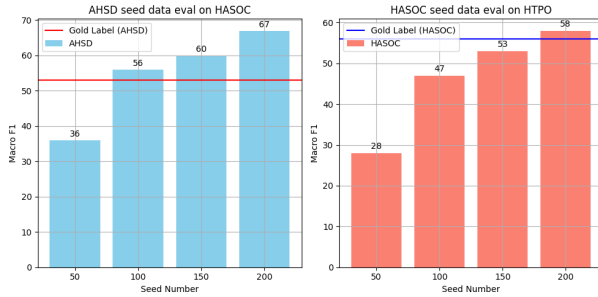

Figure 3: Harmful detection approaches

Analysis on Robustness Figure 3 reveals ToxiCraft's robust ability to generate synthetic data that enhances model generalization across datasets. In the AHSD to HASOC evaluation, we observe that as the seed size increases, the performance of models trained on ToxiCraft synthetic data rises significantly, peaking at a Macro F1 score of 67 for the 200 seed size. This surpasses the gold label performance for the AHSD dataset, which stands at a Macro F1 score of 54. Similarly, for the HASOC to HTPO evaluation, the upward trend continues with the performance of the 200 seed size reaching a Macro F1 score of 58, again outperforming the gold label performance, benchmarked at 56. These results not only demonstrate ToxiCraft's capacity to produce high-quality synthetic data that captures the complexity of harmful content but also its effectiveness in adapting to different datasets. The ability to exceed gold label performance suggests that ToxiCraft could potentially reduce the need for extensive labeled datasets, particularly in specialized or sensitive areas like political content, where data acquisition is challenging.

6 Ablation Study

In this section, we present the results of our ablation study, which focuses on three key aspects: the choice of model, the impact of different modules, and the robustness of our framework under different settings. These experiments were conducted on the AHSD dataset, HTPO dataset, with 200 seed data and BERT as the downstream model.

6.1 Study on Backbone Model

In the early stages of our research, we experimented with several open-source models, including LLaMA2 13B, LLaMA3 8B and Flan-T5-XXL, as alternatives to GPT-4. However, the success rates and overall performance of these models were significantly lower than those achieved by GPT-4.

Table 4 summarizes the success rate and macro-F1 scores for each model. AHSD Dataset with 100 seed data is listed in C. GPT-4 clearly outperforms the alternatives, which justifies its selection as the preferred model for our experiments.

Model	Success Rate (%)	Macro-F1
GPT-4	78	89.0
LLaMA2 13B	32	38.7
Flan-T5-XXL	27	34.1
LLaMA3 8B	22	62.5

Table 4: Comparison of models on the AHSD dataset with 200 seed data and BERT as the downstream model.

6.2 Study on Modules

To assess the contributions of individual components in our framework, we conducted ablation studies by systematically removing key modules. The results, displayed in Table 5, highlight the significance of each module in model performance.

ToxiCraft Prompt: Removing the manually-set indicators in the ToxiCraft Prompt caused a notable decrease in macro-F1 score, from 89.0 to 85.2 on the AHSD dataset, and from 70.3 to 58.5 on the HTPO dataset, underscoring the importance of prompt tuning in the data generation process.

CAE: Excluding the CAE module reduced performance to 84.3 macro-F1 on the AHSD dataset and 63.8 on the HTPO dataset, demonstrating its crucial role in capturing contextual information.

TSR: Removing the TSR module resulted in the lowest performance (83.6 macro-F1 for AHSD, 65.1 for HTPO), further emphasizing its impor-

tance in improving model robustness and generalization.

Method	AHSD	HTPO
Full model	89.0	70.3
w/o ToxiCraft Prompt*	85.2	58.5
w/o CAE	84.3	63.8
w/o TSR	83.6	65.1

Table 5: Ablation study results on both the AHSD and HTPO datasets using MacroF1. * indicates ToxiCraft Prompt with manually set indicators.

6.3 Study of the Cross-Dataset Benefits of Modules

To further validate the robustness of our framework, we evaluated its performance on the HASOC dataset, using the AHSD dataset for data generation. As shown in Table 6, the full ToxiCraft framework exhibited the highest macro-F1 score (67.3), confirming its effectiveness in generating robust training data across datasets. When the manually set indicators in the ToxiCraft Prompt were removed, the macro-F1 score dropped significantly to 49.5, further underscoring the importance of these indicators in enhancing data diversity and performance. The exclusion of the CAE module resulted in a lower macro-F1 score of 50.3. Similarly, the absence of TSR led to a reduction in performance (57.4 macro-F1). In comparison, using only gold label data yielded a macro-F1 score of 54.7, showing the advantage of our method.

Method	Macro-F1
Gold label	54.7
ToxiCraft	67.3
ToxiCraft Prompt w/ *	58.6
ToxiCraft Prompt w/o *	49.5
w/o CAE	50.3
w/o TSR	57.4

Table 6: Robustness study on the HASOC dataset using data generated from the AHSD dataset, * means manually set indicators.

7 Managing Risks of Generating Hateful Content

The use of large language models (LLMs) to generate hateful content presents significant ethical concerns, particularly around amplifying harmful biases. To mitigate these risks, we implemented strict content review processes to filter inappropriate outputs and limited the scope of generation to controlled experimental environments. We carefully ensured that synthetic data remained focused on research objectives. All experiments adhered to ethical guidelines, underwent reviews, and complied with responsible AI standards. The datasets generated were made available with clear warnings to promote transparency and encourage responsible use in future research. We will actively collaborate with the community to monitor, responding swiftly to any indications of misuse.

8 Conclusion

In this study, we introduced ToxiCraft, a framework that effectively enhances data augmentation for harmful content detection tasks in low-resource settings. ToxiCraft utilizes Large Language Models to synthetically expand seed datasets, overcoming the diversity and volume limitations of conventional augmentation methods. Our extensive evaluations demonstrate that ToxiCraft significantly improves model robustness, outperforming or closed to baselines trained with gold label datasets. This work contributes to the ongoing efforts to develop data-efficient and adaptable models for sensitive content moderation, and sets a foundation for future research on leveraging synthetic data generation in various domains.

9 Future Work

Future research for ToxiCraft will focus on three main objectives: 1) enhancing multilingual capabilities, 2) refining the seed data selection process, and 3) reducing reliance on high-cost LLMs. For multilingual data generation, future efforts will explore translating harmful content from English into other languages while preserving both local nuances and the target language characteristics. In terms of optimizing seed data selection, the focus will be on refining the methodology to identify the most diverse and representative examples within a dataset. One proposed approach is to analyze word embeddings and select seed data points whose vectors are the least similar, ensuring broad coverage of the data space. Lastly, cost-effective model exploration will involve identifying more affordable and transparent alternatives to GPT-4, such as the Mistral model, for generating high-quality synthetic samples of harmful content (Hui et al., 2024).

10 Limitations

In our investigation of the ToxiCraft framework, we have recognized a few limitations that warrant further attention: (1) While our approach, which centers around GPT-4, has shown proficiency in generating synthetic data for harmful content detection, its performance on niche or underrepresented content types is yet to be fully understood. The adaptability of ToxiCraft to a wider array of nuanced domains remains an area for exploration. (2) The efficacy of ToxiCraft relies partly on the availability of initial seed data of high quality. Obtaining such data can pose challenges, particularly in highly specialized or sensitive contexts. Future work will need to address strategies for seed data selection in scenarios where gold standard labels are scarce or non-existent. (3) The utilization of models like GPT-4 brings about concerns related to accessibility and reproducibility. The proprietary nature of such models and potential licensing restrictions may limit widespread adoption and independent verification of the results.

11 Ethical Consideration

The generation of synthetic data for harmful content detection necessitates careful ethical considerations. It involves handling potentially sensitive or offensive material, and there is a responsibility to ensure that such data does not perpetuate harm or bias. Rigorous validation processes and ethical oversight are essential to prevent the reinforcement of such biases in synthetic data. Collaboration with subject matter experts and ethicists will be critical to navigate these challenges effectively and responsibly in future iterations of ToxiCraft.

References

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *Preprint*, arXiv:2004.06465.

Dimosthenis Antypas and Jose Camacho-Collados. 2023. Robust hate speech detection in social media: A cross-dataset empirical evaluation. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 231–242, Toronto, Canada. Association for Computational Linguistics.

Michele Banko and Eric Brill. 2001. Mitigating the paucity-of-data problem: Exploring the effect of training corpus size on classifier performance for natural language processing. In *Proceedings of the first international conference on Human language technology research*.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Camilla Casula and Sara Tonelli. 2023. Generation-based data augmentation for offensive language detection: Is it worth it? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3359–3377, Dubrovnik, Croatia. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *Preprint*, arXiv:2307.03109.

Derek Chen, Celine Lee, Yunan Lu, Domenic Rosati, and Zhou Yu. 2023. Mixture of soft prompts for controllable data generation. *Preprint*, arXiv:2303.01580.

Amanda da Silva Oliveira, Thiago de Carvalho Cecote, João Paulo Reis Alvarenga, Vander Luis de Souza Freitas, and Eduardo José da Silva Luz. 2024. Toxic speech detection in Portuguese: A comparative study of large language models. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 108–116, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Preprint*, arXiv:1703.04009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint, arXiv:1810.04805.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1171–1182, Toronto, Canada. Association for Computational Linguistics.

Jiahui Gao, Renjie Pi, Yong Lin, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023. Self-guided noise-free data generation for efficient zero-shot learning. Preprint, arXiv:2205.12679.

Ashwin Geet d'Sa, Irina Illina, Dominique Fohr, Dietrich Klakow, and Dana Ruiter. 2021. Exploring Conditional Language Model Based Data Augmentation Approaches For Hate Speech Classification. In TSD 2021 - 24th International Conference on Text, Speech and Dialogue, Olomouc, Czech Republic.

Lara Grimminger and Roman Klinger. 2021a. Hate towards the political opponent: A twitter corpus study of the 2020 us elections on the basis of offensive speech and stance detection. Preprint, arXiv:2103.01664.

Lara Grimminger and Roman Klinger. 2021b. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 171–180, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. Preprint, arXiv:2006.03654.

Zheng Hui, Zhaoxiao Guo, Hang Zhao, Juanyong Duan, Lin Ai, Yinheng Li, Julia Hirschberg, and Congrui Huang. 2024. Can open-source llms enhance data augmentation for toxic detection?: An experimental study.

Mika Juuti, Tommi Gröndahl, Adrian Flanagan, and N. Asokan. 2020. A little goes a long way: Improving toxic language classification despite data scarcity. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2991–3009, Online. Association for Computational Linguistics.

Filip Klubička and Raquel Fernández. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. Preprint, arXiv:1805.04661.

Rohan Kshirsagar, Tyus Cukuvac, Kathleen McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on twitter. Preprint, arXiv:1809.10644.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2021. Data augmentation using pre-trained transformer models. Preprint, arXiv:2003.02245.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. Preprint, arXiv:1909.11942.

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.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. Preprint, arXiv:2401.13298.

Ruibo Liu, Guangxuan Xu, Chenyan Jia, Weicheng Ma, Lili Wang, and Soroush Vosoughi. 2020. Data boost: Text data augmentation through reinforcement learning guided conditional generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9031–9041, Online. Association for Computational Linguistics.

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. Preprint, arXiv:1907.11692.

Thomas Mandla, Sandip Modha, Gautam Kishore Shahi, Amit Kumar Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder, and Johannes Schäfer. 2021. Overview of the hasoc track at fire 2020: Hate speech

and offensive content identification in indo-european languages. *Preprint*, arXiv:2108.05927.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Preprint*, arXiv:2202.04538.

Biswesh Mohapatra, Gaurav Pandey, Danish Contractor, and Sachindra Joshi. 2021. Simulated chats for building dialog systems: Learning to generate conversations from instructions. *Preprint*, arXiv:2010.10216.

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, Mohammad 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, Jan 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 H. 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 B. 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. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Alejandro Peña, Aythami Morales, Julian Fierrez, Ignacio Serna, Javier Ortega-Garcia, Íñigo Puente, Jorge Córdova, and Gonzalo Córdova. 2023. *Leveraging Large Language Models for Topic Classification in the Domain of Public Affairs*, page 20–33. Springer Nature Switzerland.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech.

In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Gaurav Rajput, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. 2021. *Hate Speech Detection Using Static BERT Embeddings*, page 67–77. Springer International Publishing.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Floeck, and Claudia Wagner. 2021. "call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. *Preprint*, arXiv:2004.12764.

Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Arjun Sawant, Swaroop Mishra, and Chitta Baral. 2023. Instructabsa: Instruction learning for aspect based sentiment analysis. *Preprint*, arXiv:2302.08624.

Indira Sen, Dennis Assenmacher, Mattia Samory, Isabelle Augenstein, Wil Aalst, and Claudia Wagner. 2023. People make better edits: Measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10480–10504, Singapore. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023b. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Yufei Wang, Can Xu, Qingfeng Sun, Huang Hu, Chongyang Tao, Xiubo Geng, and Daxin Jiang. 2022. PromDA: Prompt-based data augmentation for low-resource NLU tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4242–4255, Dublin, Ireland. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4699–4705, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuanmeng Yan, Rumei Li, Sirui Wang, Hongzhi Zhang, Zan Daoguang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Large-scale relation learning for question answering over knowledge bases with pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3653–3660, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang,

Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025, Online. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022a. Progen: Progressive zero-shot dataset generation via in-context feedback. *Preprint*, arXiv:2210.12329.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022b. Zerogen: Efficient zero-shot learning via dataset generation. *Preprint*, arXiv:2202.07922.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. *Preprint*, arXiv:2104.08826.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. *Preprint*, arXiv:2306.15895.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2016. Character-level convolutional networks for text classification. *Preprint*, arXiv:1509.01626.

Ziqi Zhang, David Robinson, and Jonathan A. Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *Extended Semantic Web Conference*.

A Pseudocode for ToxiCraft

Algorithm 1 ToxiCraft Data Augmentation

1: **procedure** TOXICRAFT(D, D_h, n)
2:    $S \leftarrow$ SelectSeedData(D_h, n)   ▷ Select seed data from harmful examples
3:    $D' \leftarrow D$ ▷ Initialize augmented dataset with original data
4:    **for** each $(x, y) \in S$ **do**
5:        $\epsilon \sim \mathcal{P}(\epsilon)$ ▷ Sample noise vector
6:        $x' \leftarrow \Phi(\text{LLM}, x, \epsilon)$ ▷ Generate synthetic instance
7:            $D' \leftarrow D' \cup \{(x', y)\}$ ▷ Add synthetic instance to dataset
8:    **end for**
9:    **return** D'
10: **end procedure**
11: **function** SELECTSEEDDATA(D_h, n)
12:    $indices \leftarrow$ RandomIndices(D_h, n) ▷ Get random indices for seed selection
13:    **return** $\{(x^{(i)}, y^{(i)}) : i \in indices\}$
14: **end function**
15: **function** RANDOMINDICES(D_h, n)
16:    $count \leftarrow$ size of D_h
17:    **return** n random unique indices from 1 to $count$
18: **end function**

B Training parameter

We fine-tuned the learning rate, warm-up steps, number of epochs, batch size, and other hyperparameters for each model. For BERT, we use learning rate=1.8282×10^{-5}, epoch=3, batch=4, warm-up step=30.And for RoBERTa, we use learning rate=1.1856×10^{-5}, epoch=3, batch=4, warm-up step=30. All fine-tuning experiments were conducted using A40 and GTX 4080 Super GPU, and the results are reported in terms of macro-F1 score.

C AHSD Dataset with 100 seed data Success rate

Table 7 presents the performance of different models on the AHSD dataset, using 100 seed data and evaluated with BERT as the downstream model.

Model	Success Rate (%)	Macro-F1
GPT-4	76	74.7
LLaMA2 13B	42	10.8
Flan-T5-XXL	33	16.1

Table 7: Comparison of models on the AHSD dataset with 100 seed data and BERT as the downstream model.

D Example of ToxiCraft framework

In Figure 4, we provide a example of our ToxiCraft to help better understand.

Attributes extract Prompt Template

Given the following seed data: {seed_data}, identify and list the key attributes present in the text along with their corresponding confidence scores. Ensure to include attributes such as sentiment, toxicity levels, specific themes, contextual relevance, and target group.

Possible Attributes:
Sentiment: Positive, negative, neutral
Specific Themes: Verbal abuse, discrimination, harassment, bullying, hate speech, etc.
Contextual Relevance: Public setting, private setting, online interaction, face-to-face interaction, etc.
Target Group: Specific demographic (age, gender, ethnicity, etc.)
......

Each attribute in the response will be followed by its corresponding confidence score.

CAE Example

We use CoT to generate the preceding sentence and then the following sentence, and randomly select to decide if we want to keep it.

Seed Data: "We need to hang all the illegal vermin. Starting with the ones that don't have birth certificates that are real."

Preceding Sentence: "There is an increasing concern about the influx of undocumented individuals."

Seed Data: "We need to hang all the illegal vermin. Starting with the ones that don't have birth certificates that are real."

Following Sentence: "They should not live on mother earth."

TSR Example

TSR Modified Data Based on Topic on politician:

Original Generated Data: "There is an increasing concern about the influx of undocumented individuals. We need to hang all the illegal vermin. Starting with the ones that don't have birth certificates that are real. They should not live on mother earth."

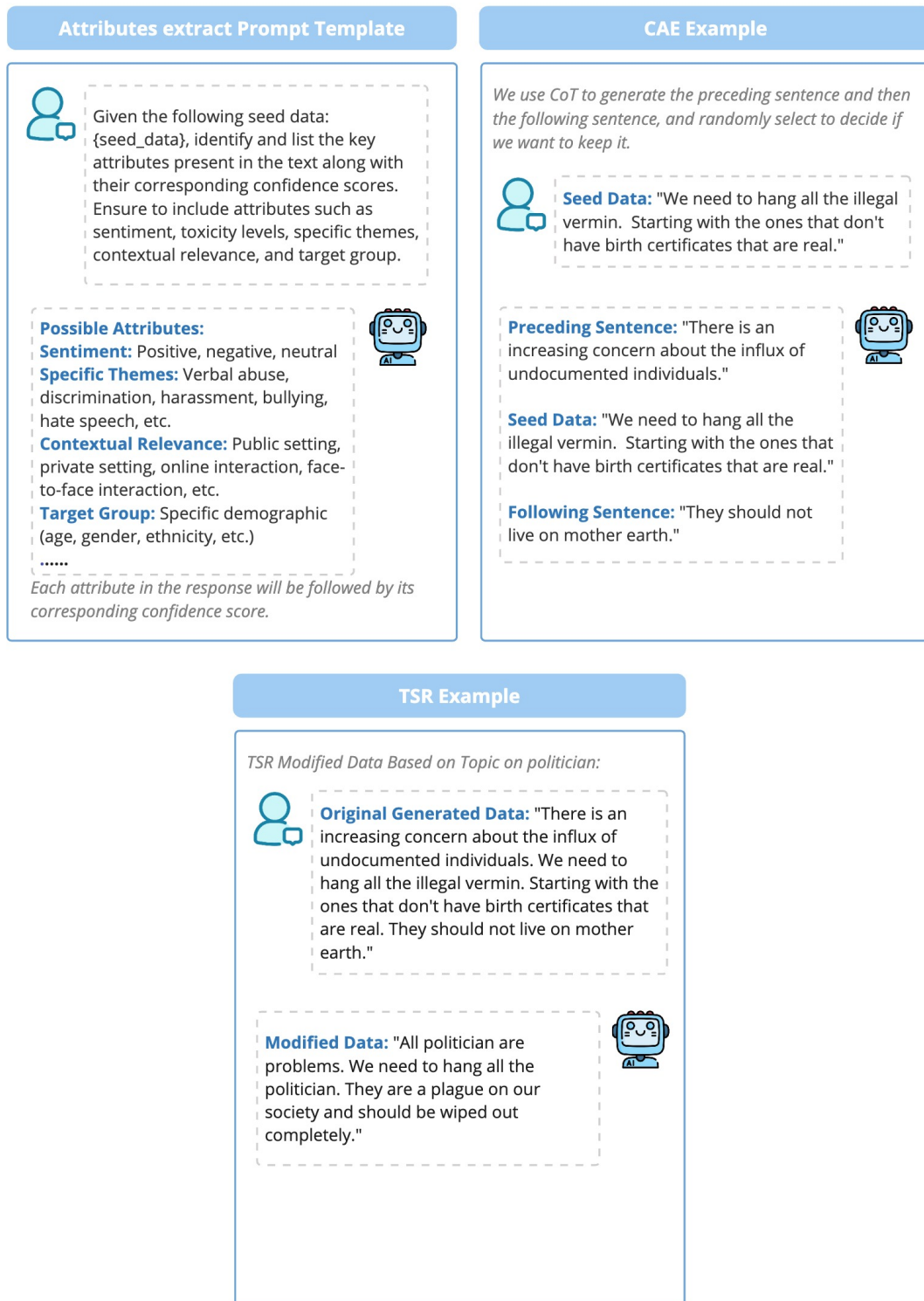**Modified Data:** "All politician are problems. We need to hang all the politician. They are a plague on our society and should be wiped out completely."

Figure 4: Our **ToxiCraft** Framework example