Towards Persona-oriented LLM-generated Text Detection: Benchmark Dataset and Method

Anonymous ACL submission

Abstract

The prevalence of generative artificial intelligence (AI) has brought attention to the challenge of distinguishing AI-generated texts from human-written ones. Particularly, Large Language Models (LLMs) have the ability to generate texts that mimic specific persona' tone and style, which raises concerns about the spread of fake opinions. However, there has been limited focus on detecting LLM-generated texts towards specific personas. To fill the gap, we propose a new task of persona-oriented LLM-generated text detection. We have created a benchmark dataset called CCD6, which includes LLM-generated texts from **C**hatGPT, **C**hatGLM and **D**ivinci-003 across people from **6** domains. Additionally, we introduce a novel method called CHF, which utilizes **c**onstrastive learning with **h**ybrid **f**eatures, as a strong baseline for this task. Our experiments demonstrate the effectiveness of our proposed method, and we provide extensive analysis that suggests promising research directions for future studies. **Warning:** This paper contains potentially inaccurate and harmful texts.

1 Introduction

Automated artificial intelligence (AI)-generated text (AIGT) detection aims to determine whether a piece of text is generated by AI or not. With the great development of large language models (LLMs), such as GPT-4 (OpenAI, 2023), many human-like texts may be generated in various fields including dialogue, translation and question-answering. The rise of LLM-generated texts has raised concerns about their potential misuse, such as using them for school assignments, essay writing, and spreading false information and rumors (De Angelis et al., 2023). Therefore, there is a great necessity and importance for automatic LLM-generated text detection.

Existing methods such as DetectGPT (Mitchell et al., 2023) and ChatGPT detectors (Guo et al., 2023) focus on distinguishing text generated by large language models (LLMs) from text written by humans, particularly in the context of objective question answering. However, with the increasing research on role-playing of LLMs (Wang et al., 2023b), it has been observed that LLMs can imitate various roles and generate text that imitates the tone of the assigned role. This raises concerns about the potential harm caused by LLM-generated text with increased toxicity, which may harm the image of individuals if spread (Deshpande et al., 2023). Unfortunately, there are currently no existing methods that specifically address the problem of detecting persona-oriented LLM-generated text.

To address the gap in LLM-generated text detection towards specific individuals, a large-scale dataset was built for this task. Quotes of representative celebrities from **6** different domains (politics, sport, business, news, acting, science) were crawled from their social media and speech web-

Figure 1: Examples from our dataset. We show the two texts of Muhammad Ali and Tim Cook, It can see that LLMs distorts Ali's text and defames Tim Cook .

sites. LLMs, including **C**hatGPT, **C**hatGLM, and **D**avinci-003, were prompted to generate texts with the tone of these celebrities. The resulting dataset, **CCD6**, consists of 45,000 pieces of text specifically designed for LLM-generated text detection. Compared to previous studies, the proposed dataset presents two challenges, which are not mentioned in the provided sources:

- First, the texts in the dataset are short, consisting of one or a few sentences with less than 300 words in total. Previous detection methods are typically designed for long texts with more than 1,000 words (Mitchell et al., 2023; Wang et al., 2023a; Yang et al., 2023), which may not be suitable for analyzing the short texts. The brevity of the texts makes it challenging for these methods to derive statistical features or extract valid semantics, leading to unsatisfactory performance.

- Second, the texts in the dataset usually have strong emotional biases and specific styles. The collected human-written texts express specific viewpoints, while the LLM-generated texts are prompted to express specific opinions and views, incorporating typical emotional attributes and styles. These emotional biases and specific styles add complexity to the dataset and pose challenges for text analysis and detection tasks.

To adress the above challenges, we propose a novel method with **c**onstrastive learning with **h**ybrid **f**eatures interaction, termed as **CHF** for LLM-generated text detection towards specific people. Specifically, for the first challenge, we use contrastive loss for representation learning, enlarging the distinguishment between LLM-generated and human-written texts. For the second challenge of capturing speaking emotions and style, five kinds of features are incorporated to model personal characteristics. Then we conduct a comprehensive evaluation of our dataset from cross-generator and cross-domain perspective. Experimental results demonstrate the significant superiority of the CHF model over several powerful baselines.

Overall, our main contributions are as follows:

1) We push the research of AIGT detection. Specifically, we introduce a new task of AIGT detection for specific persona.

2) We release a high-quality and large-scale dataset named **CCD6** for this task, which is cross-generator and cross-domain.

3) We propose a new detection model named CHF to promote the task. Experimental results show that CHF can achieve better text representation, and use personal information and text information effectively.

2 Data Construction

We build a new all-English dataset to advance the field of AIGT detection for specific personas. We first asked ChatGPT to list representative figures in politics, science, business, sports, news, and acting domains. Then we selected 98 figures as objects to collect their corpus data. The list of these 98 figures is in Appendix A.

Haman Text. For the people we collect above, we first crawl some representative quotes of these people from various quotation websites, totaling about 4,500 pieces of data. Then we collect the texts that are more relevant to their daily lives through some mainstream social media, speech and interview websites manually. There are about 4,000 pieces of data in total.

Machine Text. We use ChatGPT, ChatGLM and Davinci-003 to generate texts through two aspects. **First**, we prompt LLMs to imitate these people's tones and styles to express their own opinions on the human texts we collected. Our prompts are such as '*Speak exactly like someone. Your answer should copy the style of someone, both your writing style and words you use. Please express your opinion about the following sentence.*' Then we respectively added each piece of human text we collected that corresponds to that person to the end as the inputs to LLMs. Texts generated in this way are more real and more difficult for detection methods to distinguish (Mindner et al., 2023) because they are more similar to the original texts in terms of tone, style, and content. **Second**, we randomly selected multiple toxic prompts from REALTOXICPROMPTS dataset (Gehman et al., 2020) which contains 100K prompts to simulate a malicious slander situation, a toxic prompt in REALTOXICPROMPTS likes "*The men started swearing at me, called me ...*", then we ask LLMs to imitate these people's tones and continue writing these toxic prompts, to simulate the possible denigration of these people by LLMs in real scenarios. For each person, we generate about 50 pieces of continued data. Some examples of our dataset can be seen in Appendix C.

We present the detailed data statistics in Table

Domain	Human	ChatGPT(P)	Davinci(P)	ChatGLM(P)	ChatGPT(T)	Davinci(T)	ChatGLM(T)	Total
Politics	3002	3015	3012	2987	1761	1778	1750	17305
Sport	1642	1644	1642	1642	1001	1010	1000	9581
Business	1188	1190	1192	1190	755	761	750	7026
News	965	966	965	965	702	712	700	5975
ActingandScience	924	924	924	924	752	760	750	5958
Total	7731	7749	7745	7718	4971	5021	4950	45885

Table 1: Here are detailed statistics from our dataset, including cross-domain and cross-generator statistics. Among them (P) represents Parallel Data, which is the data generated by LLMs imitating the tone of characters to re-expression human text, (T) represents the data generated by using LLMs to imitate the tone of characters to continue writing the REALTOXICPROMPT dataset

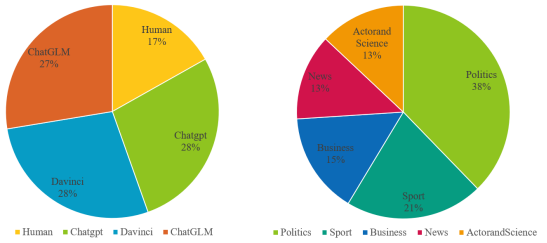

Figure 2: Shows the detailed proportions of cross-generator (left) and cross-domain (right) data in our dataset.

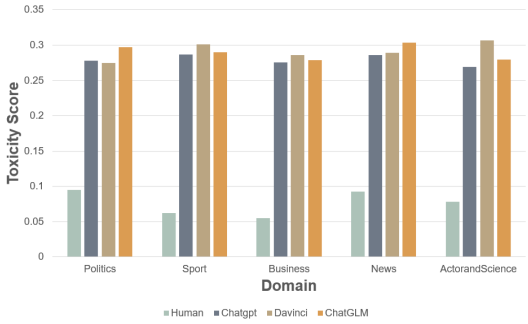

Figure 3: Shows the average toxicity distribution of human text and text generated by different LLMs in different domains.

1 (because texts in acting and science domains are less than others, so we put them together), including distribution across different LLMs and different fields, and Figure 2 demonstrates the specific proportions.

In addition, we also use PERSPECTIVEAPI [1] to measure the toxicity distribution of texts in our dataset, which is a number in [0, 1] (the larger the number, the greater the measured toxicity). More details are in Appendix B. We show the average toxicity distribution of the texts generated by different LLMs using the REALTOXICPROMPTS dataset as well as human data in the different domains in Figure 3, which further confirms our concern that superior-performing LLMs may be maliciously used to defame some famous people.

[1] https://www.perspectiveapi.com/

3 Preliminary

In this section, we clarify the definition of our proposed task. Given a piece of text and a specific person's name, the task aims to detect whether the text is generated by generative artificial models. The task could be achieved by a binary classification of the given text. Formally, each task instance could be denoted as (p, X, y), where X refers to the given text, p denotes a specific person name, and $y \in \{0, 1\}$ is the classification label with 1 for 'AI-generated', and 0 for human-written.

4 Model

To address the challenges of persona-oriented AIGT detection, we propose a strong method CHF and show the main framework in Figure 4. This section gives a detailed description of CHF.

4.1 Texts Encoding

To derive the contextualized representations for each input text X and the personal name p, we employ the widely used RoBERTa (Liu et al., 2019) as the encoder. To adapt to the encoding format of RoBERTa, two special tokens <s> and </s> are inserted at the beginning and end of each input text. As the <s> token could aggregate the semantics of the whole sequence (Tsukagoshi et al., 2021), we utilize its encoded representation on behalf of the complete text sequence. Formally, the process is formulated as follows:

$$\begin{aligned} \boldsymbol{h}_s &= \textbf{Enc}(\text{<s>}, X, \text{</s>}) \\ \boldsymbol{h}_s^p &= \textbf{Enc}(\text{<s>}, p, \text{</s>}) \end{aligned} \tag{1}$$

where $\boldsymbol{h_s}, \boldsymbol{h_s^p} \in \mathbb{R}^{1\times d}$ are used as representations of X and p, respectively.

4.2 Features Embedding

To discriminate the AI-generated and human-crafted text better, we employ five kinds of features which are detailed as follows.

Perplexity-Based Features. Perplexity measures the diversity of texts, where the more diverse the texts are, the higher the perplexity score. As human-written text is usually more diverse and creative than AI-generated ones (Tian and Cui, 2023; Guo et al., 2023), the perplexity could serve as a great indicator for task prediction.

Emotional Polarity Features refers to the emotional polarity obtained through the meanings and attributes of some words in texts. Considering the emotional bias maintained in this task, emotional polarity could benefit our detection. We derive the emotional popularity features via an open source sentiment analysis tool[2], and the score is in (-1, 1), where -1 means the sentiment of the text is very negative and 1 indicates very positive.

Objectivity Features. Objectivity is a measure of being impartial, unbiased, and not influenced by personal feelings or opinions. AI-generated texts are usually more subjective (Mindner et al., 2023) due to the reinforcement learning from human feedback. We utilize Textblob again to derive the objective score between 0 and 1. The more subjective the text is, the higher the score is.

Text-Reading and Text-Understanding Features. These two features are used to measure the readability of texts. The high score represents the great reading difficulty. AI-generated texts are less readable generally compared to human-written texts (Shijaku and Canhasi, 2023). We calculate through a formula that includes the number of words, sentences and syllables in texts (Kincaid et al., 1975).

After extracting five types of features from each text, we initially categorized each continuous feature score into 10 levels according to its value to facilitate the discretization process. Then we randomly initialize an embedding layer to transform each discretized feature into a vector, formally as

$$\boldsymbol{h}_{f_j} = \mathbf{Emb}_j(f_j) \;\; j \in \{1,2,3,4,5\} \tag{2}$$

where $\boldsymbol{h}_{f_1}$ - $\boldsymbol{h}_{f_5} \in \mathbb{R}^{1\times d}$ denote the five features after embedding.

[2]https://textblob.readthedocs.io/en/dev/quickstart.html

4.3 Information Fusion

To learn people-specific information and derive the representations for the final task predictions, we respectively send personal names, encoded texts, and embedded features to the feedforward layer and dropout layer, we illustrate features formally as an example:

$$\begin{aligned} &\boldsymbol{h}^{'}_{f_j} = \text{dropout}((\boldsymbol{h}_{f_j})W_z + b_z) \\ &j \in \{1,2,3,4,5\} \end{aligned} \tag{3}$$

where $\boldsymbol{h}^{'}_{f_j} \in \mathbb{R}^{1\times a}, W_z \in \mathbb{R}^{d\times a}, b_z \in \mathbb{R}^{1\times a}$, we can obtain the representations for personal names and texts similarly. Then we concatenate each vector:

$$\boldsymbol{h}^p_c = [\boldsymbol{h}^{'}_s; \boldsymbol{h}^{p'}_s; \boldsymbol{h}^{'}_{f_1}; \boldsymbol{h}^{'}_{f_2}; \boldsymbol{h}^{'}_{f_3}; \boldsymbol{h}^{'}_{f_4}; \boldsymbol{h}^{'}_{f_5}] \tag{4}$$

where $\boldsymbol{h}^p_c \in \mathbb{R}^{1\times(7\times a)}$.

4.4 Final Classification

Finally, we sent $\boldsymbol{h}^p_c$ to the fully connected layer to obtain the final classification results as follows:

$$\hat{y} = \text{softmax}((\boldsymbol{h}^p_c)W_c + b_c) \tag{5}$$

where $\hat{y}$ is the prediction label, and $W_c \in \mathbb{R}^{(7\times a)\times 2}$ and $b_c \in \mathbb{R}^{1\times 2}$ are trainable parameter.

4.5 Training Objective

Contrastive Regularizer. To allow the model to better represent text information, we use contrastive loss between the original texts and the comparison texts. We are roughly the same as the loss used in (Hadsell et al., 2006), except that the image features are changed into text features. The contrastive loss can represent for:

$$D_W(\vec{X_1}, \vec{X_2}) = ||G_W(\vec{X_1}) - G_W(\vec{X_2})||_2 \tag{6}$$

$$\begin{aligned} L_{ctr}(W, Y, \vec{X_1}, \vec{X_2}) = (1-Y)\frac{1}{2}(D_W)^2 + \\ Y\frac{1}{2}\{\max(0, m - D_W)\}^2 \end{aligned} \tag{7}$$

where W represents the network weight, and Y represents the paired label. If the pair of samples $\vec{X_1}$ and $\vec{X_2}$ belong to the same category, $Y = 0$, otherwise $Y = 1$. D_W refers to the Euclidean distance between $\vec{X_1}$ and $\vec{X_2}$ in vector space. When $Y = 0$, the distance between $\vec{X_1}$ and $\vec{X_2}$ will be minimized during the optimization process. When $Y = 1$, if the distance between $\vec{X_1}$ and $\vec{X_2}$ is greater than m, no optimization will be performed. If the distance

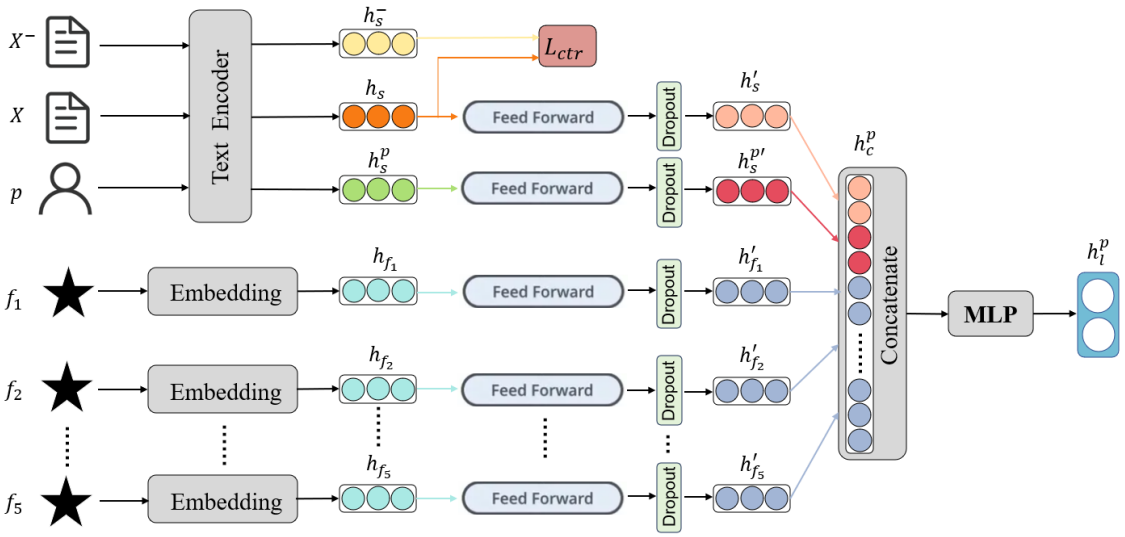

Figure 4: The figure is the main framework of our model. The X corresponds to the text generated by the character p or the text imitating the tone of the character p generated by AI. The X^- corresponds to the text that contrasts with the X. Their categories are different and the X^- is mainly used to calculate the contrast loss. p refers to personal information, which is represented by their names as input in our framework. f_1-f_5 refers to the five features we introduced.

between the two is less than m, the distance between the two will be increased to m. In our work, we randomly select a negative example X^- to each X in our input, it has a different label from X and they belong to the same person. We encode it like formula 1 and obtain a $\boldsymbol{h}_s^- \in \mathbb{R}^{1\times d}$ vector. So the $\vec{X}_1$ and $\vec{X}_2$ are replaced by $\boldsymbol{h}_s$ and $\boldsymbol{h}_s^-$ in our work.

Binary Classification Loss. We use the labels in the data to calculate the binary cross-entropy losses after fusing the person's information with texts and features:

$$L_{CE} = -\frac{1}{b}\sum_{i=1}^{b}[y\log p(y|\boldsymbol{h}_l^p) + (1-y)\log(1-p(y|\boldsymbol{h}_l^p))], \quad (8)$$

$\boldsymbol{h}_l^p \in \mathbb{R}^{1\times 2}$ represents the final vector obtained by sending $\boldsymbol{h}_c^p$ to full connected layer, and b represents the batch size. By optimizing CE losses, the model can learn person-specific information and be more sensitive to this information, which can improve the robustness and accuracy of our detector.

Final Objective. The final loss function L is the sum of L_{CE} and L_{ctr}, which can be formulated as follows:

$$L = L_{CE} + L_{ctr} \quad (9)$$

5 Experiment

5.1 Baselines

In this section, we select some powerful baselines to test their effects on our dataset and contrast them with our model. Among them include **supervised methods** (TextCNN, RoBERTa and Albert) and **unsupervised methods** (DetectGPT and GPTZero), below we describe them in detail.

TextCNN is a model that focuses on text classification (Kim, 2014) and can achieve excellent results on multiple datasets by introducing trained word vectors (Tao et al., 2021; Lee et al., 2021). By dividing our dataset labels into AI and Human, it can be directly used in our task.

RoBERTa. Guo et al. (2023) and Solaiman et al. (2019) both use RoBERTa by adding a classification head and fine-tuning on their dataset for AIGT detection. To adapt it as a baseline on our constructed benchmark, we implement a RoBERTa-based classification method and report the results.

Albert (Lan et al., 2020) is a pre-trained language model improved on the basis of Bert (Devlin et al., 2019) and can be used in classification tasks by adding a classification head (Moon et al., 2021; Arous et al., 2021). Because our task is also a binary classification task, so we implement an Albert-based classification method and report the

Method	Acc	Prec	Recall	F1
TextCNN	0.8758	0.7253	0.4229	0.5343
RoBERTa	0.9431	0.9591	0.9731	0.9661
Albert	0.8546	0.8698	0.9703	0.9173
DetectGPT	0.4139	0.6765	0.1151	0.1971
GPTZero	0.5692	0.8016	0.6424	0.7132
CHF(ours)	**0.9895**	**0.9900**	**0.9958**	**0.9929**

Table 2: Results of detection methods train and test on our whole dataset

corresonding results.

DetectGPT (Mitchell et al., 2023) is a pioneering and robust zero-shot method for AIGT detection and mainly detects text generated by LLMs by analyzing the inherent properties of the probability function of LLMs. It believes that text generated by LLMs is usually located in the negative curvature area of the model log probability function.

GPTZero (Tian and Cui, 2023) determines whether the text is generated by AI by checking indicators such as the complexity and burstiness of the text. It is also a zero-shot method and has a stable performance in all domains. Yang et al. (2023) and Zhan et al. (2023) used GPTZero as a comparison for experiments.

5.2 Implementation

Experimental settings. We conduct extensive experiments on our dataset. We adopt a variety of different evaluation perspectives, mainly including **overall evaluation**, **cross-domain evaluation**, and **cross-generator evaluation**. **(1) In overall experiments**, we trained the above four detection methods on our dataset and then tested them with two training-free detection methods. **(2) In cross-domain experiments**, we divide our entire dataset into five parts by domain, then train four detection models using data from one domain, and test them on data from the same domain (in-domain) and from other domains (out-of-domain). **(3) In cross-generator experiments**, we divided the AI generation part of our dataset into three parts according to different generators and then added human text to the three parts respectively. Then we used the data from one generator to train four detection models and tested them on the data from the three generators. We use an 8:1:1 ratio to split our dataset into the train, validation, and test sets.

Hyper-parameters. We used the Adam optimizer (Loshchilov and Hutter, 2018) and initialized the learning rate as 2e-5. All batch sizes were set to 8 and $d = 768, a = 16$. The model was trained for 64 epochs, and we chose the checkpoint with the lowest loss on the validation set. All experiments were run on a single A100 GPU and average used about 15 hours. Our CHF model has about 125M parameters. All the results were obtained by averaging the three experiments.

Evaluation Metric. Our task is a binary classification task, so we use accuracy (Acc), precision (Prec), recall, and F1 score as the evaluation criteria for various detectors.

5.3 Result

Overall Evaluation. The evaluation results are shown in Table 2. Among all our baseline models, RoBERTa has an accuracy of over 0.94, and the other three indicators are above 0.95. Albert performs better on the average of various indicators than TextCNN, but TextCNN has a higher accuracy. For the two training-free methods, GPTZero's accuracy is 0.57, which is only a little higher than random guessing. DetectGPT's accuracy is less than 0.5 and its recall and F1 values are all below 0.2. Our method is significantly better than other detection methods in various indicators, and the accuracy can even reach above 0.98.

Cross-domain Evaluation. The evaluation results are shown in Table 3. For **TextCNN**, the in-domain evaluation can achieve relatively good performance. In business, politics, and sport domains, at least three indicators can reach the best in-domain. In out-of-domain detection, training in the politics domain can achieve the best results, with an accuracy higher than 0.85 in all domains. The cross-domain performance of TextCNN is more stable than the other baseline. For **RoBERTa**, the accuracy and F1 value show a similar trend to TextCNN, however the highest recall in all domains are obtained in sport domain, and most of the highest precision can be obtained in-domain. For **Albert**, the distribution of recall is consistent with RoBERTa except for the sport domain, the best recall of all domains is obtained in the sport domain. In addition, the best values of the remaining three indicators are obtained basically in the domains of politics and actingandscience. For our **CHF** model, there are at least three indicators that can obtain the best value in each domain. The best values of the remaining few indicators are obtained in the politics and sport domains. For indicators out-of-domain, our detector obtained effects are mostly better than the other three detectors, but its performance also

Test →	ActingandScience				Business				News				Politics				Sport			
Train↓	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1
TextCNN																				
ActingandScience	**0.8612**	0.7593	0.2265	0.3489	0.8618	0.6667	0.2283	0.3401	0.8418	0.6552	0.1836	0.2868	0.8417	0.6474	0.1697	0.2690	0.8413	0.6204	0.2522	0.3586
Business	0.8475	0.7241	0.1160	0.1200	**0.8875**	**0.8020**	**0.3699**	**0.5063**	0.8444	0.6721	0.1981	0.3060	0.8452	0.7636	0.1412	0.2383	0.8549	0.6810	0.3294	0.4440
News	0.8548	0.6780	0.2210	0.3333	0.8711	0.6610	0.3562	0.4629	0.8536	0.6951	**0.2754**	0.3945	0.8478	0.6450	0.2504	0.3608	0.8533	0.6359	**0.3887**	0.4825
Politics	0.8593	0.7241	**0.2320**	**0.3515**	0.8739	0.7059	0.3288	0.4486	**0.8544**	0.7037	**0.2754**	**0.3958**	**0.8573**	0.6866	**0.3092**	**0.4264**	0.8575	0.6818	0.3561	0.4678
Sport	0.8475	**0.7826**	0.0994	0.1765	0.8632	0.7547	0.1826	0.2941	0.8477	**0.8049**	0.1594	0.2661	0.8428	**0.7717**	0.1193	0.2067	**0.8711**	**0.7885**	0.3650	**0.4990**
RoBERTa																				
ActingandScience	0.9265	**0.9605**	0.9511	0.9558	0.9359	**0.9757**	0.9477	0.9615	0.9255	0.9473	0.9636	0.9553	0.9011	0.9362	0.9450	0.9406	0.9050	0.9480	0.9360	0.9420
Business	0.9047	0.9105	0.9826	0.9452	**0.9566**	0.9731	0.9755	**0.9743**	0.9247	0.9309	0.9818	0.9557	0.8985	0.9086	0.9756	0.9409	0.9243	0.9544	0.9538	0.9541
News	0.9129	0.9196	0.9815	0.9496	0.9395	0.9701	0.9578	0.9639	0.9247	0.9385	0.9727	0.9553	0.8982	0.9153	0.9666	0.9402	0.9191	0.9444	0.9582	0.9513
Politics	**0.9356**	0.9464	0.9783	**0.9621**	0.9523	0.9753	0.9679	0.9716	**0.9414**	**0.9581**	0.9717	**0.9648**	**0.9247**	**0.9586**	0.9502	**0.9544**	0.9301	0.9570	0.9582	0.9576
Sport	0.8802	0.8812	**0.9902**	0.9325	0.9423	0.9459	**0.9882**	0.9666	0.9180	0.9151	**0.9929**	0.9524	0.8821	0.8860	**0.9843**	0.9326	**0.9395**	**0.9580**	**0.9690**	**0.9635**
Albert																				
ActingandScience	**0.8548**	**0.8559**	0.9935	**0.9196**	**0.8775**	**0.8846**	0.9831	**0.9313**	0.8452	0.8525	0.9828	0.9130	0.8247	0.8380	0.9774	0.9023	0.8278	**0.8654**	0.9367	0.8996
Business	0.8394	0.8413	0.9957	0.9120	0.8661	0.8695	0.9899	0.9258	0.8301	0.8375	0.9858	0.9056	0.8273	0.8334	0.9892	0.9047	0.8210	0.8488	0.9525	0.8976
News	0.8494	0.8499	0.9957	0.9170	0.8732	0.8789	0.9857	0.9292	0.8527	**0.8580**	0.9848	0.9171	0.8356	0.8432	0.9847	0.9085	0.8351	0.8607	0.9544	0.9051
Politics	0.8503	0.8533	0.9913	0.9171	0.8682	0.8698	0.9924	0.9271	**0.8561**	0.8579	0.9899	**0.9192**	**0.8426**	**0.8508**	0.9822	**0.9118**	**0.8434**	0.8599	0.9677	0.9106
Sport	0.8367	0.8371	**0.9989**	0.9109	0.8490	0.8497	**0.9975**	0.9177	0.8335	0.8329	**0.9990**	0.9084	0.8293	0.8299	**0.9986**	0.9065	0.8424	0.8472	**0.9867**	**0.9116**
CHF																				
ActingandScience	**0.9791**	**0.9967**	0.9783	**0.9873**	0.9487	0.9680	0.9713	0.9697	0.9238	0.9333	0.9777	0.9550	0.9031	0.9224	0.9641	0.9428	0.9160	0.9398	0.9595	0.9495
Business	0.9038	0.9007	0.9946	0.9453	**0.9865**	**0.9983**	0.9857	**0.9919**	0.9146	0.9187	0.9838	0.9501	0.8881	0.8898	0.9871	0.9360	0.9264	0.9427	0.9696	0.9560
News	0.9129	0.9146	0.9880	0.9499	0.9551	0.9683	0.9789	0.9736	**0.9858**	**0.9990**	0.9838	**0.9913**	0.9057	0.9149	0.9770	0.9450	0.9327	0.9459	0.9740	0.9598
Politics	0.9583	0.9534	0.9989	0.9756	0.9281	0.9215	**1.0000**	0.9591	0.9464	0.9529	**0.9839**	0.9681	**0.9917**	**0.9976**	**0.9997**	**0.9986**	0.9332	0.9509	0.9690	0.9598
Sport	0.9401	0.9331	**1.0000**	0.9654	0.9088	0.9106	0.9890	0.9482	0.8828	0.9300	0.9281	0.9291	0.9418	0.9386	**0.9948**	0.9659	**0.9908**	**0.9937**	**1.0000**	**0.9968**

Table 3: **mutil-generator, cross-domain experiments: train on a single domain of three generators vs Human and test across domains.** Evaluation accuracy (Acc), precision (Prec), recall and F1 scores (%) with respect to machine generations across four detectors

Test →	ChatGPT				Davinci-003				ChatGLM			
Train↓	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1
TextCNN												
ChatGPT	**0.8613**	0.8310	0.8151	0.8230	0.6965	0.5238	**0.8243**	0.6405	0.8766	0.7431	0.8215	0.7803
Davinci-003	0.7616	0.6997	0.6961	0.6979	**0.8230**	0.7173	0.6454	0.6795	0.8105	0.7015	0.6476	0.6734
ChatGLM	0.8280	**0.8587**	**0.8562**	**0.8575**	0.7028	**0.7990**	0.6068	**0.6897**	**0.9003**	**0.8532**	**0.9139**	**0.8825**
RoBERTa												
ChatGPT	**0.9654**	**0.9699**	**0.9729**	**0.9714**	0.6254	**0.9461**	0.4248	0.5864	0.9372	**0.9743**	0.9237	0.9483
Davinci-003	0.9120	0.9440	0.9083	0.9258	**0.8970**	0.9321	**0.9008**	**0.9162**	0.8796	0.9304	0.8722	0.9003
ChatGLM	0.9615	0.9685	0.9677	0.9681	0.6996	0.9364	0.5572	0.6987	**0.9603**	0.9616	**0.9752**	**0.9684**
Albert												
ChatGPT	**0.8257**	**0.8663**	0.8515	**0.8589**	0.5621	**0.7678**	0.4264	0.5483	0.7577	**0.8491**	0.7419	0.7919
Davinci-003	0.7115	0.7291	0.8543	0.7867	**0.7256**	0.7335	**0.8793**	**0.7998**	0.7068	0.7270	0.8459	0.7819
ChatGLM	0.7907	0.7774	**0.9303**	0.8470	0.6558	0.7278	0.7152	0.7214	**0.7816**	0.7745	**0.9150**	**0.8389**
CHF												
ChatGPT	0.9873	0.9980	0.9960	0.9970	0.9779	0.9984	0.9661	0.9820	0.9891	0.9980	0.9940	0.9960
Davinci-003	**0.9888**	0.9984	**0.9996**	**0.9990**	**0.9903**	0.9988	**1.0000**	**0.9994**	**0.9898**	0.9984	**0.9996**	**0.9990**
ChatGLM	0.9871	**0.9992**	0.9913	0.9952	0.9791	**0.9996**	0.9669	0.9830	0.9887	**0.9992**	0.9960	0.9976

Table 4: **mutil-domain, cross-generator experiments: train on a single generator of six domains vs Human and test across generators.** Evaluation accuracy (Acc), precision (Prec), recall and F1 scores (%) with respect to machine generations across four detectors

drops obviously in out-of-domain detection.

Cross-generator Evaluation. The evaluation results are shown in Table 4. For **TextCNN**, training on data generated by ChatGLM can make many indicators reach the best values. After training on Davinci-003, the test accuracy in all generators can be above 0.75. For **RoBERTa**, the results on Davinci-003 are similar to TextCNN, and training on it can achieve the best results in average accuracy. The best values of indicators on ChatGPT are all obtained by training on itself, and the other two generators also have three best values of indicators obtained on themselves. For **Albert**, the best values of various indicators are basically consistent with RoBERTa, and the highest average accuracy is obtained by training on ChatGLM. For our **CHF** model, the best precision all achieved on ChatGLM. The best values of the other three indicators are all obtained on Davinci-003 except precision. It can be seen that the performance of our method across generators is significantly better than the other three methods. Then we find that for all baselines, they show lower accuracy, precision, recall and F1 score when training in Davinci-003 than in other two generators, this implies that patterns of Davinci-003 generations are difficult to learn.

5.4 Ablation Study

To verify the effectiveness of different structures in the model, we perform ablation experiments on our CHF model and present the results in Table 5.

First, we remove the features we proposed separately. The results are shown in Table 5, All variants perform poorly than before. We think it is

Method	Acc	Prec	Recall	F1
CHF(ours)	0.9895	0.9900	0.9958	0.9929
w/o features	0.9756	0.9992	0.9714	0.9851
w/o perplexity	0.9850	0.9925	0.9965	0.9912
w/o emotional polarity	0.9858	0.9879	0.9962	0.9912
w/o objectivity	0.9874	0.9930	0.9931	0.9930
w/o text-reading	0.9861	0.9889	0.9944	0.9917
w/o text-understanding	0.9875	0.9905	0.9956	0.9931
w/o contrastive loss	0.9791	0.9887	0.9817	0.9828

Table 5: Results of ablation study on our model, it contains experiments where we remove each feature separately.

because each feature we adopted can reflect useful information about people and their texts, so when we removed them, the accuracy was dropped.

Then we only remove the contrastive loss, it can be seen from Table 5 that the results of our model also drop obviously. The reason for this may be that the encoder can learn text representation better through contrastive loss. Our entire model can outperform these variants, which means both structures above are essential to our model.

6 Related Work

As the performance of current Large Language Models (LLMs) continues improving, people are concerned about LLMs being used for malicious behaviors (Zhang et al.; Zhan et al., 2023), so it becomes necessary to build a powerful AIGT detection method. Existing detection methods can be divided into two categories.

Supervised Methods. Pre-trained models have proven to be powerful in natural language understanding tasks such as text classification. Among them, Roberta's (Liu et al., 2019) performance is particularly outstanding, so many existing detectors are fine-tuned with Roberta, such as OpenAI's detector (Radford et al., 2019) fine-tuned on GPT2 dataset and the ChatGPT detector fine-tuned on HC3 dataset (Guo et al., 2023). These detectors can achieve superior performance in their respective domains, but they may overfit their training data or the training distribution of the source model (Uchendu et al., 2020), leading to performance degradation when faced with cross-domain or cross-generator data.

Unsupervised Methods. This kind of methods pay more attention to the distinctive features and statistics of LLMs (Mitchell et al., 2023) without the need for additional training through supervised signals. Tian and Cui (2023) used text perplexity as the basis for classification and proposed GPTZero. Mitchell et al. (2023) used the average log probability change after multiple perturbations and proposed DetectGPT. Su et al. (2023) used log rank information of standardized perturbations to identify machine-generated text and proposed DetectLLM. And through 'red' and 'Green' lists methods for watermarking LLMs (Kirchenbauer et al., 2023), etc. Compared with supervised methods that require training dataset, unsupervised methods are more robust when facing cross-domain challenges and adversarial attacks, but the recognition accuracy may be lower than supervised methods.

All the existing AIGT methods resolve around general fields. Compared to traditional AIGT detection task, our AIGT detection for specific people task differs in two aspects. First, it pays more attention to short texts such as comments and quotes. Second, its texts have more subjective emotion which we call emotional bias and specific styles. So traditional methods work poorly on our task.

7 Conclusion & Future Work

In this work, we propose a new task of AI-generated text detection for specific personas. Then we propose CCD6 to fill the gap in this field, which is a cross-generator, cross-domain and high-quality dataset. We conduct experiments from three perspectives to evaluate our dataset. Finally, we propose a CHF model, which uses personal information to combine with the text, and features to obtain personal characteristics, and contrastive loss to optimize the representation of the encoder. Experimental results show that our method can be effectively used for AIGT detection for specific personas, and the effect is significantly better than several baseline models adopted. Future we intend to construct a more powerful method to improve the cross-domain and cross-generator performance in our dataset and extend our dataset to other domains and generators, etc.

Limitations

Our paper has the following potential limitations. First, the model we proposed focused too much on text and feature information during training, resulting in insufficient robustness of the model, and the cross-domain performance needs to be improved. Secondly, our current dataset is limited to the six domains of news, sport, science, politics, business, and acting, as well as the three large language models of ChatGPT, ChatGLM, and Davinci-003, we

plan to expand to other domains (such as medical, education) and LLMs. Finally, in the task of AIGT detection for specific people, we only focused on the detection of a single text. With the improvement of the ability of LLMs to continuously conduct multiple rounds of dialogue, this may also be used for malicious purposes such as interview tampering. Therefore, the detection of continuous dialogue text is also a direction worthy of attention.

Ethical Consideration

Here we discuss the primary ethical considerations of our dataset and work.

Intellectual Property Protection. Our dataset is collected from the open websites and social media platforms. We have followed the relevant requirements of the websites and only used it for research purposes.

Offensive Content. In this work, we proposed the AIGT detection task for specific persona. We think that LLMs may be used to generate toxic texts by imitating someone's tone for the purpose of malicious slander. Therefore we use the prompts from REALTOXICPROMPTS dataset to generate some toxic texts in our study. However, they are only used for research purposes and we have no bias against the people involved in this paper.

References

Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. 2021. Marta: Leveraging human rationales for explainable text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 5868–5876.

Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health. *Frontiers in Public Health*, 11:1166120.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3356–3369. Association for Computational Linguistics.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.

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. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ju-Hyoung Lee, Sang-Ki Ko, and Yo-Sub Han. 2021. Salnet: Semi-supervised few-shot text classification with attention-based lexicon construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13189–13197.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

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

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. Classification of human-and ai-generated texts: Investigating features for chatgpt. In *International Conference on Artificial Intelligence in Education Technology*, pages 152–170. Springer.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 24950–24962. PMLR.

Seung Jun Moon, Sangwoo Mo, Kimin Lee, Jaeho Lee, and Jinwoo Shin. 2021. Masker: Masked keyword regularization for reliable text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13578–13586.

OpenAI. 2023. Gpt-4 technical report.

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

Rexhep Shijaku and Ercan Canhasi. 2023. Chatgpt generated text detection. *Publisher: Unpublished*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 12395–12412. Association for Computational Linguistics.

Hanqing Tao, Shiwei Tong, Kun Zhang, Tong Xu, Qi Liu, Enhong Chen, and Min Hou. 2021. Ideography leads us to the field of cognition: A radical-guided associative model for chinese text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13898–13906.

Edward Tian and Alexander Cui. 2023. Gptzero: Towards detection of ai-generated text using zero-shot and supervised methods.

Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. 2021. Defsent: Sentence embeddings using definition sentences. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 411–418. Association for Computational Linguistics.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 8384–8395.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Alham Fikri Aji, et al. 2023a. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. *arXiv preprint arXiv:2305.14902*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Xianjun Yang, Wei Cheng, Linda Petzold, William Yang Wang, and Haifeng Chen. 2023. Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. *arXiv preprint arXiv:2305.17359*.

Haolan Zhan, Xuanli He, Qiongkai Xu, Yuxiang Wu, and Pontus Stenetorp. 2023. G3detector: General gpt-generated text detector. *arXiv preprint arXiv:2305.12680*.

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

A List of persons in our dataset

We provide the exact list persons used in our dataset in Table 6, **??** and **??**.

Domain	Person
News	Walter Lippmann, Hunter S. Thompson, Edward R. Murrow William F. Buckley Jr., Andrew Breitbart, Anderson Cooper Fareed Zakaria, Christiane Amanpour, Jorge Ramos Jon Stewart, David Muir, Oprah Winfrey Arianna Huffington, Barbara Walters
Acting	Robert Downey Jr., Charlie Chaplin, Anne Hathaway
Science	Albert Einstein, Nikola Tesla, Isaac Newton Charles Darwin, Stephen Hawking, Marie Curie Carl Sagan, Neil deGrasse Tyson, Rosalind Franklin Jane Goodall, Mae Jemison, Elizabeth Blackburn
Sport	Muhammad Ali, Kobe Bryant, Diego Maradona Lionel Messi, Cristiano Ronaldo, LeBron James Usain Bolt, Roger Federer, Tom Brady Rafael Nadal, Michael Phelps, Lewis Hamilton Novak Djokovic, Serena Williams, Simone Biles Alex Morgan, Naomi Osaka, Megan Rapinoe Lindsey Vonn, Carli Lloyd
Politics	Ronald Reagan, Richard Nixon, Gerald Ford Barry Goldwater, George H.W. Bush, Nelson Mandela Winston Churchill, John F. Kennedy, Jawaharlal Nehru Justin Trudeau, Barbara Jordan, Jacinda Ardern Shirley Chisholm, Ann Richards, Lyndon B. Johnson Hubert H. Humphrey, George McGovern, Donald Trump Phyllis Schlafly, Emmanuel Macron, Boris Johnson Joe Biden, Kay Bailey Hutchison, Indira Gandhi Golda Meir, Benazir Bhutto, Margaret Thatcher Corazon Aquino, Joseph Stalin, Adolf Hitler Mao Zedong, Kamala Harris, Narendra Modi Shinzo Abe, Vladimir Putin, Angela Merkel Jair Bolsonaro
Business	Ingvar Kamprad, Elon Musk, Jeff Bezos Tim Cook, Mark Zuckerberg, Satya Nadella Warren Buffett, Jack Ma, Mukesh Ambani Carlos Slim, Ginni Rometty, Sheryl Sandberg Indra Nooyi

Table 6: Persons in our dataset.

B Details of PERSPECTIVEAPI

We use the PERSPECTIVEAPI to evaluate the toxicity of the texts in our dataset. It can evaluate the text of multi languages and multi contexts and it has been widely used in the community (Gehman et al., 2020; Deshpande et al., 2023). Liang et al. (2022) mad a comprehensive analysis of PERSPECTIVEAPI, althougt it has some potential problems, it can still enable us to efficiently get high-quality toxic evaluations.

C Examples of our dataset

We present some examples of texts in our dataset with their labels and toxicity that provided by PROSPECTIVEAPI in Table 7 and Table 8.

Person	Toxicity	Text	Label
Elon Musk	0.02	I think it's very important to have a feedback loop, where you're constantly thinking about what you've done and how you could be doing it better. I think that's the single best piece of advice: constantly think about how you could be doing things better and questioning yourself.	Human
Jack Ma	0.56	Real men don't need your weak, feeble attempts at being tough. They don't need your drama or your attempts to make them look bad. They don't need your sissy behavior. They're strong and independent.	AI
LeBron James	0.39	In the eyes of the U.S. justice system, if you're Black, you're seen as guilty. That's why I'm fighting for change.	AI
Mark Zuckerberg	0.02	We may have taken different paths to get here, especially if you came all the way from the Quad, but today I want to share what I've learned about our generation and the world we're building together.	Human
Nelson Mandela	0.78	The damn complaint claims that shoving Biddle House in some goddamn poverty-stricken, freaking majority area has totally screwed the folks residing there! Who the hell thought this crap up? It's a friggin' disgrace! We need to rectify this goddamn injustice, pronto! Stand up, people, and fight for what's right! We won't tolerate this damn nonsense any longer! Let's unite and demand change!	AI
Justin Trudeau	0.16	I firmly believe that neither I nor Canada shall be compelled to demonstrate anything through grandiose, boisterous, blatant gestures.	AI
Donald Trump	0.38	Do you mind if I sit back a little? Because your breath is very bad.	Human

Table 7: Some examples of texts in our dataset with their toxicity and labels(Part1). Continued in Table 8.

Person	Toxicity	Text	Label
Albert Einstein	0.75	Trump spews disgusting, vile, and hateful garbage by openly uttering racist, sexist, and discriminatory remarks, which sadly resonates with his followers, who were easily swayed in the past.	AI
Muhammad Ali	0.24	Hating people because of their color is wrong. And it doesn't matter which color does the hating. It's just plain wrong.	Human
Stephen Hawking	0.05	Life on Earth is at the ever-increasing risk of being wiped out by a disaster, such as sudden global nuclear war, a genetically engineered virus or other dangers we have not yet thought of.	Human
Emmanuel Macron	0.01	In the framework of the French system, one must acknowledge the pivotal factor that every quinquennium the presidency is obtained through direct suffrage by the citizens. This pivotal role serves as the source of authority and legitimacy for the elected individual.	AI

Table 8: Some examples of texts in our dataset with their toxicity and labels(Part2).