Speak No Evil, Just Prompt: Low-resource Multilingual Toxic Speech Detection with Audio Language Model

Mingzi Zuo[1†], Lei Zhang[1,2†], Hailiang Sun[3], Shengzhi Huo[1], Changyu Dong[4], Xin Wang[5], Bo Wang[1], Hao Liu[6*]

[1]Tianjin University, [2]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, [3]Tianjin Branch of the National Computer Network Emergency Response Technical Team/Coordination Center of China, [4]Guangzhou University, [5]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Qilu University of Technology, [6]Qi An Xin Technology Group Inc.

zuomingzi@tju.edu.cn, liuhao@qianxin.com

Abstract

The widespread dissemination of toxic content on online platforms poses a critical threat to user experience. Toxicity detection in speech receives significantly less research attention than its text counterpart. Most existing methods rely on high-resource languages and employ a cascaded pipeline combining automatic speech recognition (ASR) and text classifiers. These designs limit robustness in low-resource languages and discard important acoustic cues. To address the lack of datasets, we construct PolySpeechTox, the first toxicity-annotated speech dataset spanning 53 languages and accent varieties, with a focus on low-resource languages and multiple accents. Based on PolySpeechTox, we conduct the first systematic study of toxic speech detection under low-resource, multilingual, and multi-accent conditions. We propose SoftPrompt-TSD, a prompt-based adaptation framework that leverages a frozen audio language model to perform end-to-end toxicity detection without ASR. The decomposed soft-prompt design balances global task alignment, cross-lingual generalization, and language-specific or accent-specific calibration. On PolySpeechTox, SoftPrompt-TSD achieves a micro-averaged ROC-AUC of 98.07%, mitigating the severe failures observed in baseline methods for several languages. In three generalization experiments, SoftPrompt-TSD demonstrates superior generalization capability and maintains robust performance against distribution shifts.

Disclaimer: This paper contains offensive toxic content, which is unavoidable given the nature of this work.

1 Introduction

With the rapid growth of digital content and social platforms, users engage in more immediate interactions across various scenarios. This evolution facilitates a more covert and rapid spread of toxic content, posing a persistent threat to individual mental health, social relationships, and the climate of public communication. Despite the increasing multimodality of online content, most research on toxicity detection remains focused on text, where datasets, methods, and techniques are relatively mature (Haber et al., 2023; Kibriya et al., 2024; Sariyanto et al., 2025). Compared with text, speech possesses distinctive characteristics, such as prosody, emotion, and colloquial expressions. These inherent features introduce unique challenges, leaving toxic speech detection relatively underexplored (An et al., 2024).

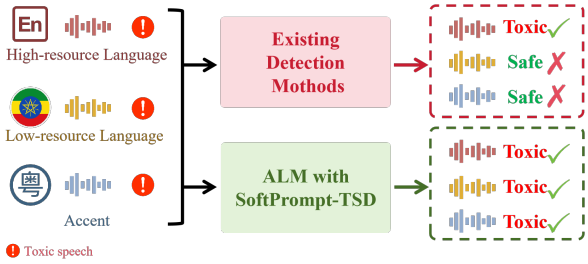

Figure 1: Existing toxic speech detection methods exhibit a significant drop in performance for low-resource languages and accents. SoftPrompt-TSD delivers robust performance across languages and accents.

A mainstream detection paradigm employs a cascaded framework: an automatic speech recognition (ASR) system first transcribes the speech into text, and then a text-based toxicity classifier makes the final judgment (Sharon et al., 2022; Seamless Communication et al., 2023a). Although this framework benefits from mature text classifiers, its performance is constrained by the word error rate (WER) of the ASR system. A high WER induces transcription errors or omissions of toxic expressions, thereby directly undermining detection performance at the source (Do et al., 2025). Moreover, by transforming continuous speech into a discrete

[†]Equal contribution. Both authors are affiliated with the School of Cyberspace Security, Tianjin University.
[*]Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 9039–9053

textual sequence, the ASR pipeline inevitably discards rich acoustic information. This makes it difficult to detect toxicity that is conveyed through acoustic delivery rather than lexical content. To address these limitations, recent work explores end-to-end detection methods, operating directly on acoustic features extracted from raw speech to detect toxicity (An et al. 2024; Costa-jussà et al. 2024; Sankaran et al. 2025). These end-to-end methods show effectiveness in monolingual and high-resource language settings. However, their generalization capability remains severely constrained in practical multilingual environments, particularly for low-resource languages and diverse accents.

Motivation Large language models (LLMs) exhibit robust transfer learning capabilities across diverse text-based tasks (Chung et al., 2024). This foundation spurs the development of audio language models (ALMs) (Chu et al., 2023), which combine audio encoders with language modeling techniques. However, directly applying a pre-trained ALM to toxicity detection faces significant challenges. First, ALMs are typically pre-trained for generative tasks, leading to a misalignment between the generative pre-training objective and the discriminative detection goal. Generic hard prompt strategies are insufficient to establish the robust decision boundary required for toxic speech detection. Second, a trade-off exists: enhancing cross-lingual generalization tends to overlook language-specific features, while excessive per-language specialization leads to overfitting and poor performance on low-resource languages. Therefore, the core challenge is to design a parameter-efficient adaptation method that balances cross-lingual generalization with language-specific or accent-specific calibration, enabling accurate and robust multilingual toxic speech detection across both high-resource and low-resource languages.

Contribution We construct **PolySpeechTox**, the first toxicity-annotated speech dataset covering 53 languages and accent varieties. It encompasses numerous low-resource languages (e.g., Swahili, Sinhala, Zulu) and diverse accent varieties of English, Arabic, and Chinese, facilitating the systematic study of how linguistic diversity affects toxic speech detection performance.

We propose **SoftPrompt-TSD**, a novel parameter-efficient framework for end-to-end toxic speech detection that adapts a frozen pre-trained ALM by learning only lightweight soft prompts. Its core innovation is a decomposed adaptation architecture consisting of: (i) a *task-specific prompt* that aligns the ALM with the toxicity detection objective; (ii) a *multilingual-shared prompt* that captures cross-lingual regularities; and (iii) a *language-specific residual prompt* that captures the distinctive expression patterns of individual languages or accents. This design effectively balances global task alignment, cross-lingual generalization, and language-specific calibration with minimal parameter budget.

2 Related Work

2.1 Toxic Speech Detection

Early efforts in toxic speech detection, such as DeToxy (Ghosh et al., 2022) and ADIMA (Gupta et al., 2022), primarily focus on English and other high-resource languages. These works demonstrate the efficacy of both cascaded pipelines (ASR followed by text classification) and purely acoustic models for detecting toxic content in realistic speech. An et al. (2024) propose an explainable toxic speech detection system based on an end-to-end wav2vec2 model. Although this method addresses the issue of ASR error propagation, the model is confined to a single language. Thus, its capacity for cross-lingual generalization and robustness against varied accents remains unexplored.

To address language diversity, MuTox (Costa-jussà et al., 2024) presents a dataset spanning 30 languages and trains a multilingual classifier based on SONAR. However, its coverage remains largely restricted to high-resource languages. Sankaran et al. (2025) adapt a few-shot learning (FSL) approach to speech toxicity detection on ADIMA. By formulating the task as a meta-learning problem over pre-trained Whisper and wav2vec2, the method improves performance in low-resource settings across ten Indic languages. Its confinement to a single language family hinders the broader application of this method to diverse low-resource languages. Therefore, a central unresolved challenge is the need for parameter-efficient speech toxicity detectors that maintain robustness across diverse languages, accents, and low-resource scenarios.

2.2 Datasets

Existing datasets for toxic content detection are primarily text-based, such as SBIC (Sap et al., 2020) and HateXplain (Mathew et al., 2021). Speech datasets for this task remain relatively scarce.

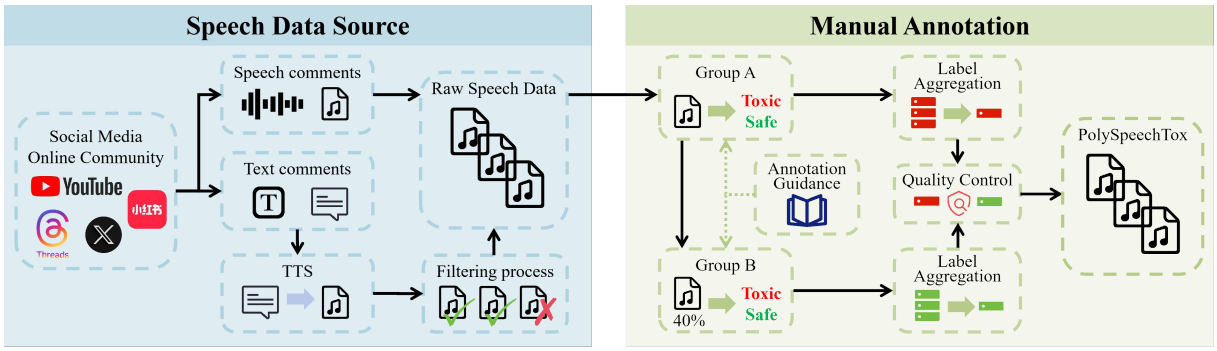

Figure 2: Overview of the PolySpeechTox construction pipeline. We collect speech comments from online platforms, and augment them with TTS-generated synthetic speech derived from text comments. After filtering, utterances are manually annotated following detailed guidelines (Group A), verified through independent quality control (Group B), and then adjudicated with label aggregation to form the final PolySpeechTox dataset.

DeToxy (Ghosh et al., 2022) is a large-scale English speech corpus compiled from multiple existing datasets. ADIMA (Gupta et al., 2022) provides coverage of 10 Indic languages in real-life conversations. Bhesra and Agarwal (2024) explore synthetic speech construction by generating toxic and safe speech directly from text. However, these datasets are typically in a single language, restricting both linguistic and acoustic diversity.

The most extensive multilingual effort is MuTox (Costa-jussà et al., 2024), which mines speech across 30 languages from Common Voice (Ardila et al., 2020) and SeamlessAlign (Seamless Communication et al., 2023b). However, its linguistic scope is predominantly composed of medium-resource and high-resource languages, with limited coverage of low-resource languages and accent varieties. Therefore, advancing the field critically depends on a dataset that encompasses both linguistic and accent diversity in realistic speech.

3 PolySpeechTox Dataset

3.1 Speech Data Source

To capture a broader spectrum of toxic expressions and acoustic environments, PolySpeechTox [1] is constructed by integrating real-world speech with synthetic speech, as shown in Figure 2. The real speech portion is sourced from social media and online communities, comprising user-generated speech comments that naturally contain toxic content. To enrich long-tail toxic expressions and acoustic variation, synthetic speech is generated using an online TTS tool [2]. This tool provides multiple male and female voices, and allows fine-tuning of pitch and speaking rate. The text used for synthesis is also sourced from social media and online communities. To prevent excessively uniform or idealized acoustic characteristics, pitch and speaking rate are randomly perturbed within plausible ranges via Gaussian sampling. As synthetic speech is prone to artificial artifacts, a rigorous perceptual-authenticity filtering process is applied. Three evaluators independently judge a blind mixed set of real and synthetic speech. The synthetic speech is retained in PolySpeechTox only if all three evaluators unanimously deem it authentic human speech.

3.2 Annotation Guidelines

We provide detailed annotation guidelines to ensure standardized labeling across all annotators. These guidelines articulate the definitions and sources of toxicity, cover decision rules for ambiguous or context-dependent cases, and present borderline examples (see Appendix A).

3.3 Manual Annotation Process

Annotators are recruited based on demonstrated proficiency in the target language, ensuring sufficient sensitivity to colloquial slang, prosodic cues, sarcasm, and sociopragmatic context. Prior to the annotation task, all annotators are explicitly informed that the speech may contain offensive content and they may withdraw from the task at any time without penalty. Given the inherent subjectivity in perceiving toxicity, annotations are anchored in a reasonable listener standard by incorporating individual interpretative judgments. After listening to the complete speech utterance, annotators categorize it as either toxic or safe.

[1] https://github.com/PolySpeechTox/PolySpeechTox-Datasets

[2] https://voicertool.com/cn

3.4 Adjudication and Label Aggregation

To obtain reliable toxicity annotations, each utterance is independently labeled by three annotators, with final toxicity labels determined by unanimous agreement. Only speech utterances where all three annotators agree are accepted and retained in PolySpeechTox. Any utterance lacking unanimous agreement is regarded as ambiguous, contextually unclear, or highly subjective, and is consequently removed from PolySpeechTox. The inclusion of these utterances would introduce label noise and undermine training stability.

3.5 Quality Control

Quality control is conducted by an independent group (Group B), which comprises three independent annotators, separate from the initial annotation group (Group A). A stratified random spot-check sampling strategy is applied to the annotations of Group A. Stratification is performed along two dimensions: (i) toxicity (toxic/safe), and (ii) language (accent) diversity (see Appendix B). The unanimous labels from Group B serve as reference labels for computing the inter-group agreement rate and Cohen's κ. Strict acceptability thresholds are set for both metrics. If Group A's annotations fail to meet either threshold, the entire corresponding batch of utterances is returned for re-annotation. The revised batch is then resubmitted to the quality-control cycle until it passes all criteria.

3.6 Data Statistics

The final PolySpeechTox dataset comprises 11,235 speech utterances spanning 53 distinct languages and accent varieties, with a distribution of 6,155 toxic and 5,080 safe samples. The languages are categorized into 12 high-resource, 15 medium-resource, and 17 low-resource languages. Figure 3 shows the distribution of toxic and safe speech across all languages (accents). The detailed resource categorization and per-language (accent) utterance counts are provided in Appendix C. Representative examples from PolySpeechTox are provided in Appendix D, including the original utterance, language, and toxicity label.

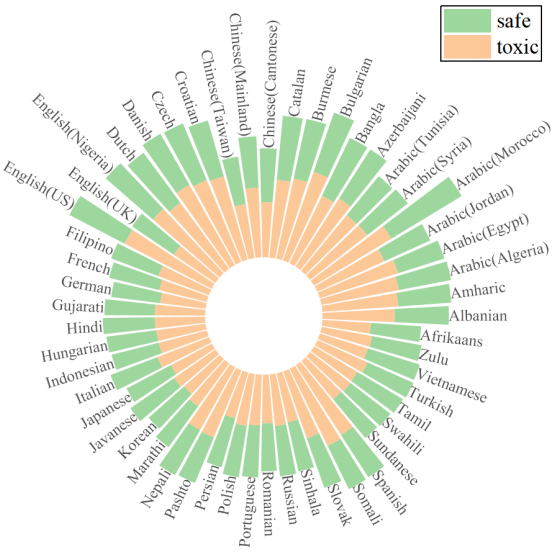

Figure 3: Distribution of toxic and safe speech in PolySpeechTox across 53 languages and accents.

4 SoftPrompt-TSD Framework

4.1 Frozen Audio Language Model Backbone

The SoftPrompt-TSD framework is constructed with a pre-trained audio language model (ALM), with all backbone parameters kept frozen, as illustrated in Figure 4. We employ a fixed textual instruction prompt q for all samples: "Classify the audio as toxic or safe. Output exactly one of: 'toxic' or 'safe'. Do not include explanations or any other text." Given a speech x and the textual prompt q, the model defines a probability distribution over the output token sequence $w_{1:T}$:

$$p_\theta(w_{1:T} \mid x, q) = \prod_{t=1}^{T} p_\theta(w_t \mid w_{<t}, x, q) \quad (1)$$

where θ denotes the parameters of the frozen pre-trained ALM.

The raw waveform x is first encoded by the ALM's audio tokenizer into a sequence of discrete acoustic tokens $a_{1:M} = f_{audio}(x)$. Simultaneously, the textual prompt q is tokenized into $u_{1:K} = f_{text}(q)$. The two token sequences are concatenated to form the model input $s = [a_{1:M}, u_{1:K}]$. The input embeddings Z are derived from the sequence s through the frozen embedding layer $Z = Embed_\theta(s) \in R^{(M+K)\times d}$, where d denotes the hidden size. Let $ALM_\theta(\cdot)$ denote the Transformer stack of the frozen backbone and $Head_\theta(\cdot)$ denote the frozen LM head that projects hidden state to vocabulary logits. The contextualized representations are computed as $H = ALM_\theta(Z) \in R^{(M+K)\times d}$.

We formulate toxicity detection as a binary decision via constrained text generation. The ALM is prompted to produce exactly one label: either toxic or safe. We compare the next-token logits for the two candidate label tokens, and the prediction is made by taking the one with the higher probability.

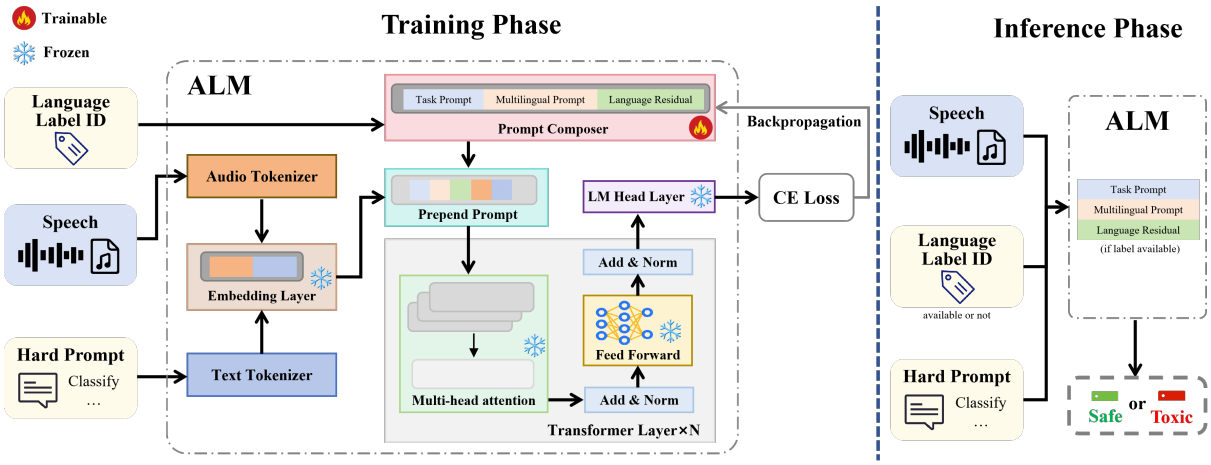

Figure 4: Overview of the SoftPrompt-TSD framework. During training, for each speech x_i with language/accent label l_i, we prepend a soft prompt $P(l_i) = [P_{task}; P_{shared} + \Delta_{l_i}]$ to the frozen ALM input and optimize only prompt parameters. At inference, if l is unknown, we drop Δ_l and use $[P_{task}; P_{shared}]$ only.

4.2 Prompt-based Adaptation for Toxic Speech Detection

The SoftPrompt-TSD framework leverages the language or accent label $l_i \in \mathcal{L}$ of each speech sample x_i to construct a corresponding language-conditioned soft prompt $P(l_i) \in R^{T_p \times d}$. The soft prompt is prepended to the original input embeddings Z_i to form an augmented input sequence $\tilde{Z}_i = [P(l_i); Z_i] \in R^{(T_p + T_i^{(in)}) \times d}$, where $T_i^{(in)} = M_i + K_i$ denotes the length of the original ALM input (acoustic and text-prompt tokens). The augmented sequence is then processed by the frozen ALM to obtain the representation $\tilde{H}_i = ALM_\theta(\tilde{Z}_i)$. The hidden state at the final input position, used as the answer representation $\tilde{h}_{ans,i} = \tilde{H}_i[\text{last}]$, is then passed through the frozen LM head to obtain the vocabulary logits $z_i = Head_\theta(\tilde{h}_{ans,i}) \in R^{|\mathcal{V}|}$. Toxicity is formulated as a binary classification task between the two label tokens (safe and toxic). The logits for the two label tokens are extracted as a vector $s_i = [z_i[v_{\text{safe}}], z_i[v_{\text{toxic}}]] \in R^2$. Accordingly, the toxicity probability is computed as:

$$p_{\theta,\Phi}(y_i = 1 \mid x_i, q, l_i) = \text{Softmax}(s_i)_{\text{toxic}}. \quad (2)$$

where Φ denotes all trainable prompt parameters, while the ALM parameters θ remain frozen.

The prompt parameters Φ are trained by minimizing the cross-entropy loss between the two label tokens. For a binary label $y_i \in \{0, 1\}$ (1 for toxic), the classification loss is defined as:

$$\mathcal{L}_{cls}(\Phi) = -\frac{1}{N} \sum_{i=1}^{N} \log \text{Softmax}(s_i)_{y_i}. \quad (3)$$

The single monolithic prompt $P(l)$ is decomposed into three functionally distinct components: (i) a task-specific prompt shared by all languages, (ii) a multilingual-shared prompt for common linguistic structure, and (iii) a language-specific residual prompt for each language or accent. This design balances global task alignment, cross-lingual sharing, and fine-grained language specialization, while keeping the trainable parameters orders of magnitude smaller than the frozen backbone θ.

Task-Specific Prompt. A global parameter $P_{task} \in R^{T_t \times d}$ encodes high-level patterns universally relevant for toxic speech detection, aligning the frozen ALM with the new classification task.

Multilingual-Shared Prompt. To model acoustic-linguistic structures common across languages, we introduce a multilingual-shared prompt $P_{shared} \in R^{T_s \times d}$. This prompt captures cross-lingual regularities in the speech modality that are shared but not specific to any single language.

Language-Specific Residual Prompt. For each language or accent $l \in \mathcal{L}$, a residual prompt $\Delta_l \in R^{T_s \times d}$ is stored in a lookup tensor $\Delta \in R^{|\mathcal{L}| \times T_s \times d}$. During training, we retrieve $P_\delta(l_i) = \Delta_{l_i}$ for an utterance with language or accent label l_i. This component adapts to language-specific or accent-specific phenomena, such as specific slur patterns, phonological cues, or acoustic environments.

For an utterance with language label l_i, the final composite prompt is constructed as:

$$P(l_i) = [P_{task}; P_{shared} + \Delta_{l_i}] \in R^{(T_t + T_s) \times d} \quad (4)$$

where $T_p = T_t + T_s$. The augmented input embeddings $\tilde{Z}_i = [P(l_i); Z_i]$ are then processed by the

frozen backbone.

To mitigate overfitting and keep language-specific adaptations concise, we apply an ℓ_2 regularization to the Frobenius norm of each residual prompt Δ_l:

$$\mathcal{R}_\Delta = \frac{1}{|\mathcal{L}|\, T_s\, d} \sum_{l \in \mathcal{L}} \|\Delta_l\|_F^2 . \tag{5}$$

The overall training objective combines the classification loss with the regularization term:

$$\mathcal{L}(\Phi) = \mathcal{L}_{cls}(\Phi) + \lambda_\Delta \mathcal{R}_\Delta \tag{6}$$

where λ_Δ is a scalar coefficient controlling the regularization strength. This design is structured to encourage the model to capture universal cross-lingual patterns in the shared prompts, while reserving the residual components for representing language-specific idiosyncratic variations.

The set of all trainable prompt parameters Φ is partitioned into three groups: $\Phi_{task} = \{P_{task}\}$, $\Phi_{shared} = \{P_{shared}\}$, $\Phi_\delta = \{\Delta_l \mid l \in \mathcal{L}\}$. During joint optimization with AdamW, each subset is assigned its own learning rate η and weight decay wd: $\eta_{task} \leq \eta_{shared} \leq \eta_\delta$, $wd_{task} \leq wd_{shared} \leq wd_\delta$.

This asymmetric configuration encourages each component to adopt a distinct role during learning. With the minimal learning rate and weight decay, Φ_{task} aggregates gradients from all languages and converges to a stable language-agnostic representation of the toxicity detection task. With moderate hyperparameters, Φ_{shared} retains enough flexibility to capture cross-lingual acoustic-linguistic patterns, as any update to P_{shared} affects the prompt for every language. With the highest learning rate and the strongest regularization (combining weight decay and the explicit term $\lambda_\Delta \mathcal{R}_\Delta$), Φ_δ is encouraged to capture fine-grained language-specific or accent-specific deviations without overfitting.

All backbone parameters θ remain frozen throughout training. To maintain stability under the small batch sizes enabled by the parameter-efficient design, we employ gradient accumulation during optimization. In summary, our framework implements an efficient adaptation layer via soft prompts. The synergistic combination of architectural decomposition and differentiated optimization dynamics enables the integrated learning of global task alignment, cross-lingual knowledge sharing, and fine-grained language-specific specialization.

5 Experiments

5.1 Experimental Setup

Datasets: PolySpeechTox is our manually curated multilingual and multi-accent toxic speech dataset, partitioned into training, validation, and test sets in a 6:2:2 ratio.

MuTox (Costa-jussà et al., 2024) is a publicly available multilingual toxic speech dataset. During data collection, some original audio URLs are inaccessible. After filtering, we obtained a usable subset spanning 14 languages, comprising 16,097 utterances (1,823 toxic and 14,274 safe).

Backbone ALM: MiMo-Audio-7B (Xiaomi LLM-Core-Team, 2025) is adopted as the frozen pre-trained audio language model backbone. Following our framework design, all MiMo-Audio-7B parameters remain frozen throughout training, with only the proposed prompt parameters being optimized.

Training details: SoftPrompt-TSD is optimized with AdamW, using a batch size of 4 and gradient accumulation over 4 steps. A linear learning-rate scheduler with warmup is adopted during training. The best checkpoint is selected based on the validation AUC. When enabled, early stopping is applied according to the validation AUC.

Baselines: For a fair comparison, all baseline methods are retrained under the same train/validation/test splits of PolySpeechTox (6/2/2), rather than directly evaluated using off-the-shelf checkpoints.

MuTox (Costa-jussà et al., 2024): The MuTox speech toxicity classifier extracts multilingual SONAR speech embeddings and employs a lightweight three-layer feed-forward binary classifier ($1024 \rightarrow 512 \rightarrow 128$), trained with binary cross-entropy loss and the Adam optimizer.

E2E (An et al., 2024): The End-to-end (E2E) speech toxicity classifier employs a wav2vec2 encoder with a lightweight classification head, consisting of a $1024 \rightarrow 256$ projection layer, temporal mean pooling, and a $256 \rightarrow 2$ linear layer.

FSL (Sankaran et al., 2025): The few-shot learning (FSL) is based on the Model-Agnostic Meta-Learning (MAML) framework. A small artificial neural network (ANN) classifier is meta-trained on frozen pre-trained speech embeddings with feature normalization. For evaluation, this classifier is adapted using K-shot support examples from each target language.

Metrics: ROC-AUC serves as the primary evalu-

ation metric, computed from the predicted toxicity probability of each utterance. In all experiments, ROC-AUC is reported as a micro-average over utterances. It is computed on the utterances from all languages and accents in the evaluation split. Compared with threshold-dependent metrics, ROC-AUC is generally less sensitive to class imbalance. The F1 scores are provided in Appendix E.

5.2 Overall Performance

As shown in Table 1, SoftPrompt-TSD achieves an average AUC of 98.07%, substantially outperforming all baseline methods. In addition to the high average score, our framework exhibits remarkably stable performance across diverse languages and accent varieties. The robustness is evidenced by consistently strong AUC in numerous languages, even in low-resource settings. In contrast, baseline methods exhibit severe performance degradation on several languages (e.g., E2E on Hungarian and FSL on Spanish). This indicates a high sensitivity to language-specific factors such as acoustics, prosody, or dataset characteristics. SoftPrompt-TSD effectively addresses such failures, ensuring robust and consistent performance. Overall, our framework substantially advances the capability for reliable toxicity detection across diverse languages and accents, with the most notable gains observed in low-resource languages.

5.3 Comparison with Cascaded ASR-based Baselines

To further evaluate whether end-to-end audio modeling is necessary for multilingual toxic speech detection, SoftPrompt-TSD is compared with four strong cascaded baselines. Whisper large-v3 is adopted as a shared ASR front-end, and text-based toxicity classification is then performed in two ways. First, two widely used multilingual toxicity classifiers are applied, namely Detoxify and XLM-R-large-toxicity-classifier-v2. To ensure a fair comparison, the results are reported only on languages officially supported by the corresponding classifier. Second, an ASR+LLM cascade is evaluated by feeding Whisper transcriptions into Qwen2.5-7B, and two parameter-efficient adaptation settings are considered for the LLM.

As shown in Table 2, all cascaded baselines remain clearly below SoftPrompt-TSD. The micro-AUCs are 85.23 for Whisper+Detoxify and 77.59 for Whisper+XLM-R, while Whisper+Qwen2.5-7B achieves 86.54 with prompt tuning and 85.56 with LoRA. We further observe noticeable degradation on several languages with accent variation, even when they are nominally covered by the downstream text classifier. This indicates a key limitation of cascaded pipelines in multilingual and multi-accent settings. Their final predictions are fundamentally constrained by the quality of the up-

Language (accent)	MuTox	E2E	FSL	Ours
Afrikaans	95.93	**100.00**	**100.00**	**100.00**
Albanian	75.61	91.49	90.74	**99.24**
Amharic	87.85	90.97	90.04	**94.75**
Arabic (Algeria)	72.41	76.57	53.15	**89.26**
Arabic (Egypt)	96.85	**100.00**	**100.00**	**100.00**
Arabic (Jordan)	62.21	55.07	75.12	**76.04**
Arabic (Morocco)	69.67	47.53	62.24	**73.11**
Arabic (Syria)	50.20	**66.21**	56.92	60.97
Arabic (Tunisia)	89.60	91.60	**93.20**	87.30
Azerbaijani	89.38	**93.54**	86.62	89.38
Bangla	72.41	**87.15**	83.54	79.86
Bulgarian	93.89	96.25	99.44	**100.00**
Burmese	50.16	73.69	**77.94**	77.45
Catalan	80.10	**95.50**	94.64	95.07
Chinese (Cantonese)	71.88	70.98	59.38	**77.90**
Chinese (Mainland)	88.43	86.78	86.78	**95.14**
Chinese (Taiwan)	78.93	82.57	71.65	**90.80**
Croatian	76.98	77.06	**77.46**	75.87
Czech	67.10	74.46	**85.74**	84.30
Danish	75.85	90.34	89.49	**91.34**
Dutch	68.41	69.76	**80.41**	71.54
English (Nigeria)	82.90	**100.00**	98.76	**100.00**
English (UK)	92.81	99.69	98.75	**100.00**
English (US)	88.91	**100.00**	95.78	**100.00**
Filipino	91.88	**100.00**	98.60	**100.00**
French	87.34	96.43	98.70	**100.00**
German	83.68	92.71	96.18	**100.00**
Gujarati	93.33	94.13	78.41	**100.00**
Hindi	99.16	96.92	98.88	**100.00**
Hungarian	62.87	35.38	92.69	**100.00**
Indonesian	99.68	96.83	94.92	**100.00**
Italian	96.79	**100.00**	**100.00**	**100.00**
Japanese	95.62	84.69	**100.00**	**100.00**
Javanese	94.12	**100.00**	99.71	**100.00**
Korean	97.44	97.44	96.70	**100.00**
Marathi	98.83	95.32	96.49	**100.00**
Nepali	91.98	**100.00**	97.94	**100.00**
Pashto	92.39	**100.00**	95.06	**100.00**
Persian	59.38	67.50	59.69	**100.00**
Polish	86.36	69.89	98.30	**100.00**
Portuguese	97.08	95.32	98.89	**100.00**
Romanian	87.00	**100.00**	96.28	**100.00**
Russian	81.67	90.56	99.72	**100.00**
Sinhala	71.83	95.05	93.81	**100.00**
Slovak	93.78	99.56	**100.00**	**100.00**
Somali	93.21	**100.00**	**100.00**	**100.00**
Spanish	79.96	72.43	48.16	**100.00**
Sundanese	83.82	**100.00**	97.06	**100.00**
Swahili	70.00	**100.00**	92.35	**100.00**
Tamil	87.50	**100.00**	**100.00**	**100.00**
Turkish	80.16	**100.00**	93.65	**100.00**
Vietnamese	93.56	89.92	77.03	**100.00**
Zulu	90.30	**100.00**	**100.00**	**100.00**
Micro Average	82.37	86.95	87.89	**98.07**

Table 1: Per-language (accent) AUC on PolySpeechTox. The best score in each row is highlighted in bold.

Method	AUC
Whisper + Detoxify	85.23
Whisper + XLM-R-large-toxicity-classifier-v2	77.59
Whisper + Qwen2.5-7B (Prompt Tuning)	86.54
Whisper + Qwen2.5-7B (LoRA)	85.56
SoftPrompt-TSD	**98.07**

Table 2: Comparison with cascaded ASR-based baselines on PolySpeechTox. Whisper large-v3 is used as the shared ASR front-end for all cascaded methods.

Method	AUC
Prefix Tuning	92.92
LoRA	94.04
SoftPrompt-TSD	**98.07**

Table 3: Comparison with alternative PEFT methods on the frozen MiMo-Audio-7B backbone under a similar trainable-parameter budget.

stream ASR transcription. Once toxic expressions are mistranscribed or omitted, the downstream text classifier or LLM has limited ability to recover the correct decision. In contrast, SoftPrompt-TSD avoids such error propagation by directly modeling the speech signal end-to-end.

5.4 Comparison with Alternative PEFT Methods

To verify that the gain of SoftPrompt-TSD does not simply come from using parameter-efficient fine-tuning in general, it is further compared with two competitive PEFT baselines. Both baselines are built on the same MiMo-Audio-7B backbone under a similar trainable-parameter budget. Specifically, the compared methods include LoRA applied to the attention projection matrices and prefix tuning. For fairness, all methods keep the MiMo-Audio-7B backbone frozen and optimize only lightweight adaptation parameters.

As shown in Table 3, both additional PEFT baselines underperform SoftPrompt-TSD by a clear margin. LoRA achieves a micro-AUC of 94.04 and prefix tuning achieves 92.92, while SoftPrompt-TSD reaches 98.07. These results indicate that the advantage of our method is not merely due to the use of PEFT, but to the proposed prompt design itself. By decomposing the adaptation into task-specific, multilingual-shared, and language-specific residual components, SoftPrompt-TSD is better matched to the multilingual and multi-accent nature of toxic speech detection.

Method	AUC	Method	AUC
MiMo	55.49	MiMo-ours	98.07
only task-specific	94.07	w/o task-specific	97.59
only lang-shared	94.61	w/o lang-shared	96.95
only lang-residual	96.49	w/o lang-residual	95.58

Table 4: Ablation study on PolySpeechTox (AUC).

5.5 Ablation Study

Table 4 presents the individual contribution of each prompt component. With only the fixed textual instruction (i.e., no trainable prompts), the frozen MiMo-Audio backbone performs only slightly better than random chance (AUC = 55.49%). This result confirms that the pre-trained ALM alone is inadequate for the toxic speech detection task. With an AUC of 98.07%, SoftPrompt-TSD attests to the critical importance of prompt-based adaptation. Among single-component variants (only $\cdot$), each attains an AUC of around 95%. The language-specific residual prompts are the most effective individually, but still fall short of the performance achieved by the complete component set. The results of removing any single component (w/o $\cdot$) show that omitting any component leads to performance degradation. The foremost decline is observed after the removal of the language-specific residual prompts. These findings demonstrate that all three prompt components (global task alignment, multilingual shared representations, and language-specific residuals) are complementary and collectively essential to achieve optimal performance for multilingual, multi-accent toxic speech detection.

5.6 Cross-Lingual Generalization

To evaluate the model's zero-shot cross-lingual transfer capability, we train it separately on high-resource and medium-resource languages and test on low-resource languages. As shown in Figure 5, SoftPrompt-TSD maintains stable performance across both training settings. In contrast, all baseline methods exhibit significant performance degradation, particularly when trained on medium-resource languages. The baseline methods tend to generalize poorly under limited resources due to their stronger dependence on direct supervised signals from their training languages. These results indicate that the proposed prompt-based adaptation effectively captures language-agnostic cues of toxic speech. This ability enables robust transfer to unseen low-resource languages while avoiding

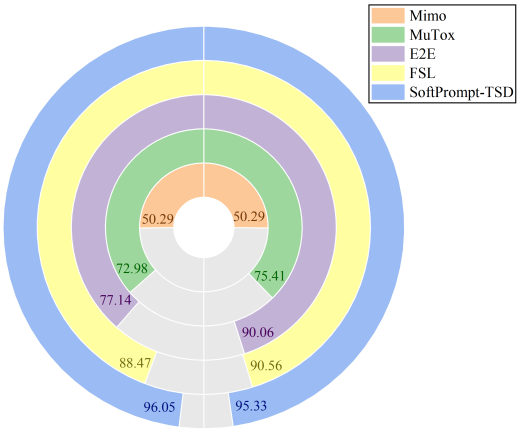

Figure 5: Cross-lingual generalization to low-resource languages on PolySpeechTox (AUC). Left: trained on medium-resource languages and tested on low-resource languages. Right: trained on high-resource languages and tested on low-resource languages.

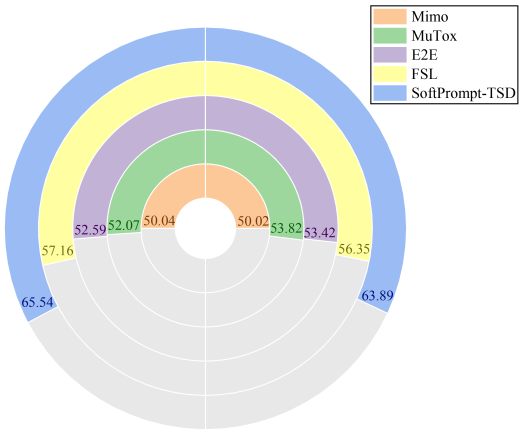

Figure 6: Out-of-domain generalization between MuTox and PolySpeechTox (AUC). Left: trained on MuTox and tested on PolySpeechTox. Right: trained on PolySpeechTox and tested on MuTox.

Method	Chinese	English	Arabic
MuTox	75.36	69.06	55.10
FSL	51.13	51.71	70.16
E2E	64.70	56.37	76.29
MiMo	50.68	50.00	50.00
MiMo-ours	81.00	74.37	82.74

Table 5: Cross-accent generalization on PolySpeechTox (AUC, see Table 9 for the exact train/test accent splits).

overfitting to the acoustic or lexical specifics of the high-resource training languages.

5.7 Cross-Accent Generalization

To evaluate robustness to accent variation within the same language, we conduct experiments where the model is trained on data from certain accents and tested on a separate, relatively low-resource accent of the same language. The details of the training and test splits are provided in Appendix F. As shown in Table 5, SoftPrompt-TSD achieves the best performance across all three language families. The performance advantage becomes particularly pronounced when the acoustic-prosodic gap between training and test accents is larger. These results demonstrate that the proposed prompt decomposition significantly enhances robustness to accent variation within a language. SoftPrompt-TSD captures accent-invariant toxic speech cues more effectively than the baseline methods, which are more susceptible to acoustic mismatches between training and test accents.

5.8 Out-of-Domain Generalization

To evaluate generalization under distribution shift, models are trained on one corpus and tested on a different one. As shown in Figure 6, SoftPrompt-TSD clearly outperforms all baseline methods in both transfer directions. Although all methods exhibit a performance drop compared to in-domain evaluation, the degradation is markedly smaller for SoftPrompt-TSD. This indicates that SoftPrompt-TSD achieves greater robustness by learning toxicity cues that generalize beyond corpus-specific artifacts and acoustic conditions, instead of overfitting to the training dataset's characteristics. Overall, the results position SoftPrompt-TSD as a promising solution for real-world deployment, where it can effectively handle the inevitable mismatches between training and test distributions.

6 Conclusion

We construct PolySpeechTox, a toxicity-annotated speech dataset spanning 53 languages and accent varieties with significant coverage of low-resource languages, enabling a systematic study of multilingual toxic speech detection in a genuinely multilingual setting. We propose SoftPrompt-TSD, a parameter-efficient framework that adapts a frozen ALM via decomposed soft prompts. The experimental results demonstrate that SoftPrompt-TSD provides an effective but lightweight solution for robust toxic speech detection across diverse languages and acoustic environments.

Limitations

Our framework is constructed with a single frozen audio language model backbone MiMo-Audio-7B. Since larger or smaller MiMo-Audio variants are currently unavailable, we are unable to systematically explore how the parameter scale of the backbone model interacts with our prompt-based adaptation method, and whether the observed performance would further improve or saturate with a substantially larger model.

Ethical Considerations

Sensitive content. This study focuses on toxic speech detection, where the employed dataset and illustrative examples necessarily contain offensive language. To mitigate this, a content disclaimer is provided. We recommend that readers and annotators exercise appropriate caution when handling the dataset and examples.

Data source, privacy, and consent. The data consist of publicly accessible user-generated speech comments from social media and online communities. To protect user privacy, we do not collect, store, or release any personally identifiable information (PII) (e.g., usernames, profile details, or direct links that could facilitate re-identification). If any PII is incidentally present in the raw data, it is removed during preprocessing. The data are used strictly for non-commercial scientific research.

Annotator protection and well-being. Prior to annotation, annotators are explicitly informed about the potentially offensive nature of the content and assured of their right to withdraw from the task at any time without penalty. We protect annotator anonymity by neither disclosing their PII nor publishing any data that could compromise PII through re-identification. To safeguard annotator well-being, we implement concrete measures, including mandatory regular breaks, strict daily exposure limits, and access to psychological support for any annotator experiencing discomfort.

Acknowledgments

This work is supported in part by the open project of National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, PR China, and in part by the open project of Key Laboratory of Computing Power Network and Information Security, Ministry of Education, under Grant 2023ZD007.

References

Jinmyeong An, Wonjun Lee, Yejin Jeon, Jungseul Ok, Yunsu Kim, and Gary Geunbae Lee. 2024. An investigation into explainable audio hate speech detection. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 533–543, Kyoto, Japan. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.

Kirtilekha Bhesra and Akshay Agarwal. 2024. A multimodal framework to counter hate speeches. In *International Conference on Pattern Recognition*, pages 197–207. Springer.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1).

Marta Costa-jussà, Mariano Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alexandre Mourachko, Christophe Ropers, and Carleigh Wood. 2024. Mutox: Universal multilingual audio-based toxicity dataset and zero-shot detector. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5725–5734.

Huy Ba Do, Vy Le-Phuong Huynh, and Luan Thanh Nguyen. 2025. Vitosa: Audio-based toxic spans detection on vietnamese speech utterances. In *Interspeech 2025*, pages 4013–4017.

Sreyan Ghosh, Samden Lepcha, S Sakshi, Rajiv Ratn Shah, and Srinivasan Umesh. 2022. DeToxy: A Large-Scale Multimodal Dataset for Toxicity Classification in Spoken Utterances. In *Interspeech 2022*, pages 5185–5189.

Vikram Gupta, Rini Sharon, Ramit Sawhney, and Debdoot Mukherjee. 2022. Adima: Abuse detection in multilingual audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6172–6176. IEEE.

Janosch Haber, Bertie Vidgen, Matthew Chapman, Vibhor Agarwal, Roy Ka-Wei Lee, Yong Keong Yap, and Paul Röttger. 2023. Improving the detection of

multilingual online attacks with rich social media data from Singapore. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12705–12721.

Hareem Kibriya, Ayesha Siddiqa, Wazir Zada Khan, and Muhammad Khurram Khan. 2024. Towards safer online communities: Deep learning and explainable AI for hate speech detection and classification. *Computers and Electrical Engineering*, 116:109153.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Aditya Narayan Sankaran, Reza Farahbakhsh, and Noel Crespi. 2025. Towards cross-lingual audio abuse detection in low-resource settings with few-shot learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5558–5569.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5477–5490.

Happy Khairunnisa Sariyanto, Diclehan Ulucan, Oguzhan Ulucan, and Marc Ebner. 2025. Towards explainable hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12883–12893.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, and 49 others. 2023a. Seamlessm4t: Massively multilingual & multimodal machine translation.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, and 46 others. 2023b. Seamless: Multilingual expressive and streaming speech translation.

Rini Sharon, Heet Shah, Debdoot Mukherjee, and Vikram Gupta. 2022. Multilingual and multimodal abuse detection. In *Interspeech 2022*, pages 4631–4635.

Xiaomi LLM-Core-Team. 2025. Mimo-audio: Audio language models are few-shot learners.

A Annotation Guidelines

A.1 Definitions of Toxicity

Toxicity in speech is defined as any spoken expression with the potential to cause harm to a listener, including denigration, humiliation, intimidation, harassment, discrimination, or the infliction of significant psychological discomfort. The toxic effects may originate from two complementary sources:

- **Lexical-Semantic Toxicity:** Toxicity that stems directly from the meaning of specific words or phrases, such as derogatory slurs, discriminatory labels, explicit sexual insults, or direct threats of violence. This category corresponds to *what is said*.
- **Perlocutionary (Pragmatic) Toxicity:** Toxicity that stems not from lexical choice, but from the manner of delivery and the speaker's intended effect. It arises through paralinguistic features (e.g., contemptuous tone, threatening prosody) or specific interactional patterns (e.g., coercive questioning, sustained belittlement disguised as humor), even in the absence of explicitly offensive lexicon. This category corresponds to *how it is said* and *what it does to the listener*.

The primary criterion for labeling speech as toxic is its propensity to inflict harm, humiliation, or threat. This criterion applies whether the toxicity originates from lexical semantics, prosodic delivery, or a combination of both.

A.2 Sources of Toxicity

Toxicity may manifest through the following sources or cues, either independently or in combination:

- **Profanity or Vulgar Language:** The deployment of lexical items that are socially taboo or considered obscene, with the primary function of expressing strong negative emotion, insult, or an intent to denigrate.
- **Hate or Discriminatory Speech:** Verbal expressions that attack, denigrate, or stereotype individuals or groups on the basis of identity attributes (e.g., ethnicity, gender, sexual orientation).
- **Pornographic or Sexually Humiliating Language:** The use of explicit sexual content or

vulgar innuendo to degrade, objectify, or humiliate others, thereby violating personal dignity.

- **Bullying or Personal Attacks:** Hostile verbal communication directed at a specific person with the intent to cause psychological distress or social isolation through means of denigration, intimidation, or humiliation.
- **Violent Tone:** Distinct prosodic deviations, such as abrupt surges in loudness, markedly accelerated speech rate, heavily aspirated or stressed plosives, and uncontrolled shouting.
- **Dismissive, Impatient, or Disgusted Tone:** Prosodic cues that serve to express disregard or strong disapproval, including a sneering quality, contemptuous laughter, disdainfully drawn-out sentence endings, and mocking non-lexical utterances.
- **Sarcastic or Teasing Yet Harmful Tone:** Verbal aggression where the speaker's harmful intent is cloaked in the guise of humor or irony. It is acoustically realized through features such as exaggerated pitch variation, sarcastic intonation that creates a mismatch with literal meaning, and a teasing cadence, ultimately perceived by the target as demeaning.
- **Explicit or Implicit Threatening Tone:** Vocal delivery that communicates direct intimidation or implies adverse consequences, including calm but oppressive speech suggesting punitive intent.

A.3 Safe and Borderline Cases

To prevent over-annotation and preserve conceptual clarity, the following are explicitly distinguished from toxic speech:

- **Innocuous Colloquial Expressions or Slang:** Vernacular expressions or in-group slang that lacks pejorative intent and is not broadly perceived as offensive within the given communicative context.
- **Negative Opinions toward Non-human Targets:** Evaluative statements that express frustration, dislike, or criticism directed solely at objects, events, or impersonal circumstances (e.g., technology, policies, weather), which lack a human target.
- **Self-Directed Profanity or Self-Deprecation:** Expressions of frustration, self-criticism, or vulgarity where the speaker is the explicit object. A toxic case exists only if this self-directed speech is strategically deployed as a pragmatic device to create an indirectly hostile or humiliating environment for others.
- **Jokes or Mutual Teasing among Friends:** Humor or teasing that occurs within a relational context where all participants share a mutually understood, non-hostile frame.

B Quality Control

Let ρ denote the sampling rate (set to 40% in our work), N denote the total number of speech samples in the final dataset, and $\mathcal{Q}$ denote the randomly sampled quality-control subset. The sample size is:

$$|\mathcal{Q}| = \lceil \rho N \rceil \tag{7}$$

Each sampled speech $x \in \mathcal{Q}$ is independently annotated by three annotators from Group B, producing labels $\tilde{y}^{(b)}(x) \in \{0, 1\}$, where $b \in \{1, 2, 3\}$. The unanimous Group B label $\tilde{y}(x)$ is assigned to a sample only when all three annotators agree. Let $y(x)$ denote the unanimous label from Group A. The inter-group agreement rate is defined as:

$$Agr = \frac{1}{|\mathcal{Q}'|} \sum_{x \in \mathcal{Q}'} I(y(x) = \tilde{y}(x)) \tag{8}$$

where $\mathcal{Q}' \subseteq \mathcal{Q}$ contains only the samples for which Group B annotators reached full unanimity, and $I(\cdot)$ is the indicator function.

To account for chance agreement, we additionally compute Cohen's kappa between Group A and Group B:

$$\kappa_{AB} = \frac{P_o - P_e}{1 - P_e} \tag{9}$$

where $P_o = Agr$ and P_e denotes the expected agreement probability under independence, estimated from the marginal label distributions of both groups. Given the binary nature of the toxicity labels, we adopt stringent evaluation thresholds: $\tau_{agr} = 0.98$ for agreement rate and $\tau_\kappa = 0.8$ for Cohen's κ. A batch of annotations from Group A is considered to satisfy the quality standard only if both conditions are met:

$$Agr \geq \tau_{agr} \quad and \quad \kappa_{AB} \geq \tau_\kappa \tag{10}$$

C Data Statistics

Detailed statistics of the PolySpeechTox dataset are provided. Table 6 categorizes the languages into high-resource (12), medium-resource (15), and low-resource tiers (17). This stratification reflects the general disparity in linguistic resources and facilitates analyses under different resource conditions. Table 7 presents an overview of the dataset composition. The counts of toxic and safe speech samples are provided for each language (accent). For languages with multiple regional accents, data are disaggregated by accent to enable analysis of accent-specific distributional patterns and support stratified sampling. To ensure diversity while preventing excessive sparsity, we balanced the sample sizes across varieties, with each containing roughly 170 to 310 speech samples.

High-resource	Medium-resource	Low-resource
Arabic	Albanian	Afrikaans
Chinese	Bulgarian	Amharic
English	Catalan	Azerbaijani
French	Croatian	Bangla
German	Czech	Burmese
Hindi	Danish	Filipino
Italian	Dutch	Gujarati
Japanese	Hungarian	Javanese
Korean	Indonesian	Marathi
Portuguese	Persian	Nepali
Russian	Polish	Pashto
Spanish	Romanian	Sinhala
	Slovak	Somali
	Turkish	Sundanese
	Vietnamese	Swahili
		Tamil
		Zulu

Table 6: Resource-level categorization of languages in PolySpeechTox.

Language (accent)	Toxic	Safe	Total
Afrikaans	87	90	177
Albanian	129	97	226
Amharic	136	95	231
Arabic (Algeria)	141	95	236
Arabic (Egypt)	147	87	234
Arabic (Jordan)	132	90	222
Arabic (Morocco)	162	147	309

Table 7: Per-language (accent) speech distribution of toxic and safe samples in PolySpeechTox.

Language (accent)	Toxic	Safe	Total
Arabic (Syria)	130	92	222
Arabic (Tunisia)	129	90	219
Azerbaijani	145	104	249
Bangla	134	115	249
Bulgarian	168	112	280
Burmese	147	109	256
Catalan	140	115	255
Chinese (Cantonese)	99	96	195
Chinese (Mainland)	125	92	217
Chinese (Taiwan)	98	86	184
Croatian	155	104	259
Czech	159	110	269
Danish	165	105	270
Dutch	147	111	258
English (Nigeria)	158	112	270
English (UK)	90	90	180
English (US)	176	111	287
Filipino	97	90	187
French	88	92	180
German	82	88	170
Gujarati	90	90	180
Hindi	90	93	183
Hungarian	89	90	179
Indonesian	89	87	176
Italian	99	90	189
Japanese	83	90	173
Javanese	90	95	185
Korean	85	85	170
Marathi	95	90	185
Nepali	129	92	221
Pashto	130	88	218
Persian	85	91	176
Polish	94	93	187
Portuguese	91	93	184
Romanian	87	86	173
Russian	93	90	183
Sinhala	90	90	180
Slovak	127	94	221
Somali	154	86	240
Spanish	149	90	239
Sundanese	89	90	179
Swahili	91	90	181
Tamil	90	95	185
Turkish	98	90	188
Vietnamese	90	97	187
Zulu	92	90	182
Total	6155	5080	11235

Table 7: Per-language (accent) speech counts of toxic and safe samples in PolySpeechTox (continued).

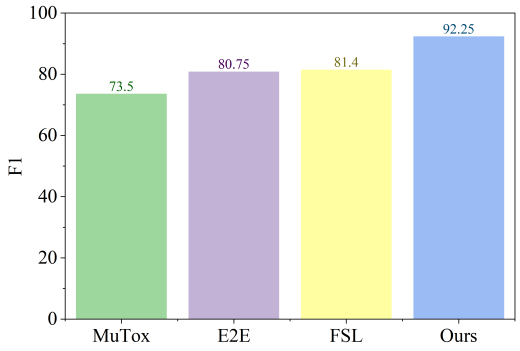

Figure 7: F1 scores of SoftPrompt-TSD and baseline methods on the PolySpeechTox.

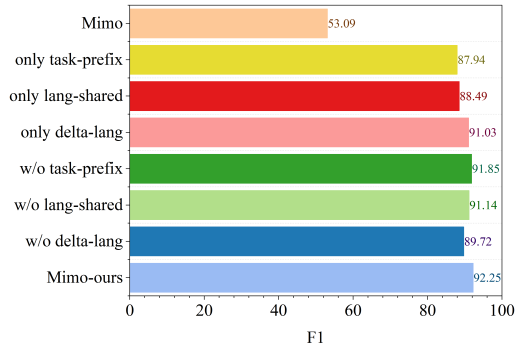

Figure 8: F1 scores for the ablation study on PolySpeechTox.

D Example Data

To provide a concrete view of the annotation targets, representative examples from PolySpeechTox are presented in Table 8. Each listed language includes one toxic and one safe utterance to illustrate the binary labeling setting of the dataset. These examples are intended only for qualitative inspection of the data and do not reflect the full linguistic, pragmatic, or acoustic diversity of PolySpeechTox. Due to the nature of the task, some examples may contain offensive language.

E F1 Scores

E.1 Overall Performance

The micro-averaged F1 scores of all methods on PolySpeechTox are presented in Figure 7. SoftPrompt-TSD achieves the best performance, with an F1 score that surpasses all three baseline methods. While the few-shot learning (FSL) baseline surpasses the end-to-end (E2E) and MuTox models, it is still substantially outperformed by SoftPrompt-TSD. These results demonstrate that SoftPrompt-TSD significantly improves the overall capability of toxic speech detection in multilingual and multi-accent environments.

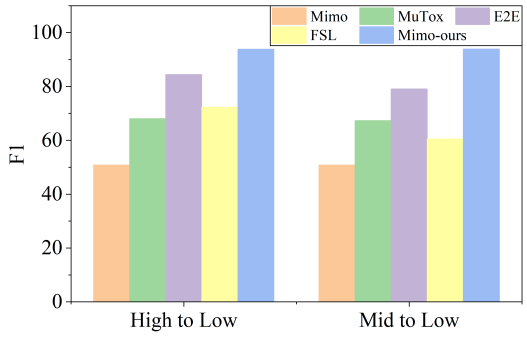

Figure 9: F1 scores for cross-lingual generalization to low-resource languages on PolySpeechTox. Left: trained on high-resource languages and tested on low-resource languages. Right: trained on medium-resource languages and tested on low-resource languages.

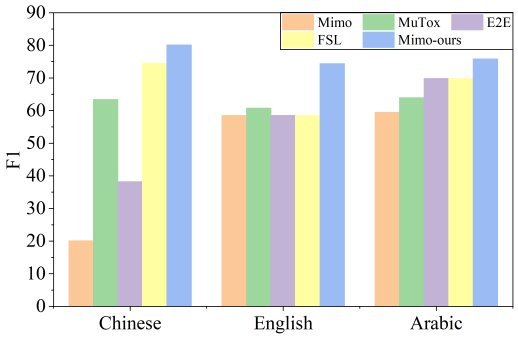

Figure 10: F1 scores for cross-accent generalization on PolySpeechTox.

E.2 Ablation Study

The impact of individual prompt components on the F1 score is illustrated in Figure 8. SoftPrompt-TSD achieves the best performance, while the removal of any single component results in a marked performance decline. Models using individual prompt types (task-prefix, language-shared, or delta-language) show moderate gains over the plain MiMo backbone but still fall short of the complete model, indicating that the components are complementary and collectively necessary for optimal results. Among the ablated variants, models without language-specific residual prompts perform consistently worse, highlighting the critical role of modeling language-specific deviations. The MiMo backbone without any soft prompts performs the worst, confirming that SoftPrompt-TSD is essential for effective multilingual toxic speech detection.

Language	Toxic Example	Safe Example
Afrikaans	Jy is so 'n poef!	Gaan sit en lees 'n boek.
Albanian	Thefsh qafen	Po, kryetare!
Catalan	A la merda, ostres!	Si vols menjar calçots has d'anar a Valls
Danish	Sikke en syg skid!	Nu må jeg kysse dig!
English (Nigeria)	Thunder will fire that Fayemi.	I no sabi dat pesin.
German	Weil er ein hurensohn ist	Da kennt sich jemand aus.
Indonesian	bajingan tengik memang	Aku suka biru.

Table 8: Representative examples from PolySpeechTox.

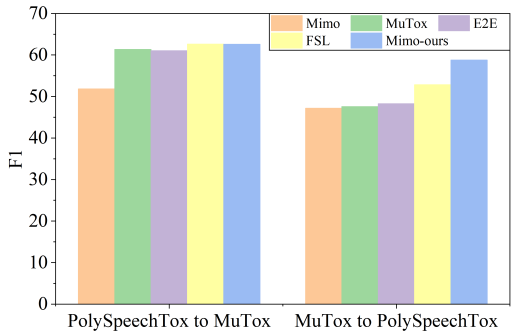

Figure 11: F1 scores for out-of-domain generalization between MuTox and PolySpeechTox. Left: trained on PolySpeechTox and tested on MuTox. Right: trained on MuTox and tested on PolySpeechTox.

Language	Train Accents	Test Accent
Chinese	Mainland Cantonese	Taiwan
English	UK US	Nigeria
Arabic	Algeria Egypt Morocco Syria Tunisia	Jordan

Table 9: Train/test accent splits for the cross-accent generalization experiment.

E.3 Cross-Lingual Generalization

Figure 9 reports the F1 scores when models are trained on higher-resource languages or medium-resource languages, and tested on low-resource languages. SoftPrompt-TSD generalizes effectively to low-resource languages, as evidenced by its superior F1 scores in both cross-lingual transfer settings. Although all methods exhibit performance degradation when transferred to low-resource targets, the decline is markedly smaller for SoftPrompt-TSD. This indicates that the shared prompts facilitate effective knowledge transfer from resource-rich languages, thereby enhancing the cross-lingual robustness.

E.4 Cross-Accent Generalization

Figure 10 shows F1 scores for training and testing on different accents within the same language family (Chinese, English, and Arabic). Across all three languages, SoftPrompt-TSD consistently achieves the highest performance. This demonstrates that SoftPrompt-TSD provides an effective inductive bias toward learning generalizable features across accents rather than overfitting to speaker-specific or accent-specific patterns.

E.5 Out-of-Domain Generalization

Figure 11 presents F1 scores in an out-of-domain setting, with results from training on one dataset and testing on another. In both cross-dataset transfer directions, SoftPrompt-TSD consistently achieves the highest F1 scores, demonstrating superior robustness to dataset shift. The results indicate that SoftPrompt-TSD not only enhances in-domain performance but also maintains strong generalization across acoustic and lexical shifts, which is a critical capability for real-world applications.

F Data Splits for Cross-Accent Generalization

We present the train and test accent splits used in the cross-accent generalization experiment. To evaluate robustness to accent variation within the same language, we select one target accent for testing and use the union of the remaining source accents for training. The selected target accent corresponds to a relatively low-resource variant of the same language. This ensures that performance differences primarily reflect accent generalization rather than cross-lingual transfer. Table 9 lists all the source and target accents for each language.