MMBERT: Scaled Mixture-of-Experts Multimodal BERT for Robust Chinese Hate Speech Detection under Cloaking Perturbations

Anonymous ACL submission

Abstract

Hate speech detection on Chinese social media platforms poses distinct challenges, particularly due to the widespread use of cloaking techniques designed to evade conventional text-based detection systems. Although large language models (LLMs) have recently improved hate speech detection capabilities, the majority of existing work has concentrated on English-language datasets, with limited attention given to multimodal strategies in the Chinese context. In this study, we propose MMBERT, a novel BERT-based multimodal framework that integrates textual, speech, and visual modalities through a Mixture-of-Experts (MoE) architecture. To address the instability associated with directly integrating MoE into BERT-based models, we develop a progressive three-stage training paradigm. MMBERT incorporates modality-specific experts, a shared self-attention mechanism, and a router-based expert allocation strategy to enhance robustness against adversarial perturbations. Empirical evaluations on multiple Chinese hate speech datasets demonstrate that MMBERT substantially outperforms both fine-tuned BERT-based encoder models and prompt-based in-context learning with LLMs.

Disclaimer: *This paper includes descriptions and analyses of violent and discriminatory language, which may be offensive for some readers.*

1 Introduction

Hate speech poses a persistent threat to online communities, exacerbated by the anonymity and scale of digital platforms (Dixon et al., 2018). While automated hate speech detection has advanced significantly in recent years, most efforts remain concentrated on English, leaving other major languages like Chinese relatively under-resourced and underprotected (Davidson et al., 2017, 2019). Some researchers have attempted to leverage LLMs for Chinese hate speech detection (Chao et al., 2024; Sun et al., 2021; Zhou et al., 2023). However, on Chinese social media platforms, many hate speech disseminators employ various cloaking perturbations to escape detection, making it challenging for existing models to identify such expressions accurately (Xiao et al., 2024b). These subtle manipulations exploit the structural and phonological properties of the Chinese language, making detection especially difficult for text-only models.

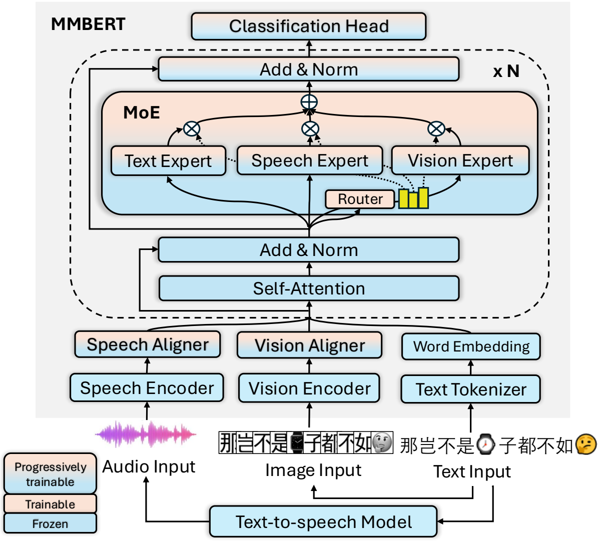


Figure 1: **Illustration of MMBERT model structure**. Compared to traditional BERT-based model, it leverages the MoE architecture to scale and effectively handle multiple modalities. A three-stage progressive training strategy is designed to ensure stable training and prevent performance degradation.

While LLMs have shown promise in content moderation, BERT-based architectures have consistently outperformed decoder-only LLMs in hate speech detection tasks, owing to their deep bidirectional encoding and strong capacity for fine-grained semantic understanding (Benayas et al., 2024; Ghorbanpour et al., 2025). Their superior performance can be attributed to the ability to generate fine-grained contextualized representations, which are especially well-suited for classification tasks that require discerning subtle semantic distinctions and interpreting nuanced language—both of which are common in adversarial or implicitly encoded hate speech (Liu et al., 2024). The architecture optimized for discriminative tasks enables more efficient and accurate detection of toxic content across various hate speech detection benchmarks (Deng et al., 2022; Xiao et al., 2024b).

To address the challenge of detecting cloaked hate speech in Chinese, we propose MMBERT, a novel multimodal BERT-based architecture that incorporates visual and speech modalities alongside text, depicted in Figure 1. To enhance scalability and specialization, MMBERT integrates the MoE mechanism, enabling dynamic routing of representations to modality-specific experts. However, naïvely inserting MoE into BERT leads to severe training instability and degraded performance, particularly in the multimodal setting (Zhang et al., 2021). To overcome this, we introduce a progressive three-stage training strategy. In the first stage, we pretrain modality aligners using synthetic multimodal data to map visual and auditory inputs into the BERT language space. In the second stage, we train modality-specific experts and continue refining aligners using task-specific supervision. In the final stage, we jointly fine-tune the full MoE-augmented architecture on real multimodal hate speech data. This phased design ensures stable optimization and effective cross-modal integration.

To overcome this, we introduce a progressive three-stage training strategy. In the first stage, we pretrain modality aligners using synthetic multimodal data to map visual and auditory inputs into the BERT language space. In the second stage, we train modality-specific experts and continue refining aligners using task-specific supervision. In the final stage, we jointly fine-tune the full MoE-augmented architecture on real multimodal hate speech data. This phased design ensures stable optimization and effective cross-modal integration.

Our experiments across three benchmark Chinese hate speech datasets demonstrate that MMBERT achieves state-of-the-art performance, significantly outperforming both fine-tuned BERT-based baselines and LLMs with in-context learning. In particular, MMBERT shows superior robustness in detecting cloaked adversarial content, highlighting the value of multimodal modeling and progressive training for Chinese hate speech detection.

We summarize the main contribution of this paper as follows:

- We propose **MMBERT**, a novel multimodal BERT-based framework for Chinese hate speech detection that integrates textual, visual, and speech modalities through a Mixture-of-Experts (MoE) architecture, enhancing robustness against cloaking-based adversarial perturbations.

- We design a **progressive three-stage training strategy** that first aligns multimodal inputs to the BERT language space, then specializes modality-specific experts, and finally fine-tunes the complete model. This approach ensures stable training and effective cross-modal representation learning.

- We conduct **extensive experiments** on three benchmark datasets, comparing MMBERT against fine-tuned BERT-based baselines and LLMs with in-context learning. Results demonstrate that MMBERT consistently achieves superior performance, particularly in detecting cloaking perturbed hate speech.

2 Related Work

2.1 Cloaking Perturbations in Chinese Hate Speech

Cloaking perturbations in Chinese online discourse represent a growing challenge for automated hate speech detection systems, as users employ various linguistic strategies to obfuscate offensive content while preserving its intended meaning (Xiao et al., 2024b,a). These perturbations can be mainly categorized into several types:

Deformation. As Chinese characters are logographic, their meanings can be altered by decomposing or reconfiguring individual components, often imparting specific emotional or ideological connotations (Lan, 2006). For example, the character "默" (meaning "silence") comprises the radicals "黑" (meaning "black") and "犬" (meaning "dog"), which in certain contexts have been used to convey derogatory implications toward the Black community.

Homophonic Substitution. Similar to English, words with similar pronunciations are frequently substituted to generate alternative semantics (Tien et al., 2021). For instance, Chinese internet users often replace the character "满" (meaning "Manchu") with "蛮" (meaning "barbarian"), as both share a phonetic resemblance to "man".

Abbreviation. The contraction of sensitive terms enhances conciseness while maintaining semantic clarity (Lan, 2006). A notable example is "txl," where each letter corresponds to the pinyin initials of "同", "性", and "恋", collectively denoting "homosexuality."

Code-Mixing. To intensify expressive tone and circumvent automated content moderation, Chinese social media users frequently incorporate non-Chinese linguistic elements—such as English vocabulary, pinyin transliterations, and emojis—into online discourse (Li et al., 2020). These code-mixed constructs not only obscure semantic intent from detection systems but also reinforce the emotive or derogatory force of the message. For instance, the term "ni哥" (meaning "ni brother") phonetically approximates the English racial slur "n*gger". Similarly, in the phrase 👅🐶 ("舔狗"), the addition of an emoji amplifies the pejorative undertone, characterizing individuals perceived as excessively submissive in relationship contexts—analogous to the English term "simp".

These perturbations exploit the unique structural and phonological characteristics of the Chinese language to conceal offensive intent (Lu et al., 2023). For instance, visually altering character radicals can introduce ideological connotations, while homophones and abbreviations obscure meanings through phonetic similarity or reduction. Code-mixing with foreign words, pinyin, or emojis further complicates semantic interpretation. Text-only models often fail to capture these manipulations due to their limited capacity to disambiguate subtle visual and phonological cues (Xiao et al., 2024a; Raza Ur Rehman et al., 2025).

2.2 Enhancing Chinese Language Modeling through Multimodal Pretraining

Text-only approaches in Chinese language modeling often face limitations in capturing the full linguistic complexity of the language, particularly with respect to character homographs, tonal ambiguity, and the lack of explicit word boundaries. These challenges hinder the model's ability to accurately interpret semantic and phonetic nuances inherent in Chinese.

To address these limitations, several studies have explored the integration of additional modalities, such as visual and phonetic information, into the pretraining process. For instance, ChineseBERT (Sun et al., 2021) integrates both glyph and pinyin embeddings, enriching the representation of Chinese characters by capturing visual features through multiple font variations and phonetic information to resolve the heteronym phenomenon. This dual-embedding approach has shown significant improvements in various Chinese natural language processing tasks, such as named entity recognition and sentiment analysis. Similarly, models like ERNIE-M (Ouyang et al., 2020) and GlyphBERT (Li et al., 2021) have demonstrated the benefits of incorporating external modalities, such as entity knowledge and visual cues, to enhance language understanding.

However, existing multimodal approaches predominantly rely on embedding-level fusion of heterogeneous input modalities within a fixed BERT encoder architecture. While such integration enhances input representations, the processing and interaction of multimodal information remain largely static and inflexible. Specifically, the fixed fusion mechanism in standard BERT layers may limit the model's capacity to dynamically adapt to context-dependent linguistic challenges, such as homographs and tonal ambiguity in Chinese. This rigidity restricts the model's ability to effectively leverage the complementary strengths of each modality in a nuanced and input-sensitive manner.

2.3 Scaling Multimodal Language Models with MoE Architectures

Recent advancements in large MLLMs have increasingly explored the use of MoE (Eigen et al., 2013) architectures to enhance scalability, efficiency, and specialization across modalities. Early generations of MLLMs, such as Flamingo (Alayrac et al., 2022) and GPT-4V (Yang et al., 2023), are grounded in dense architectural paradigms that encounter scalability limitations as data volume and modality complexity increase. To address this, MoE-based frameworks such as CuMo (Li et al., 2024) and Uni-MoE (Li et al., 2025) introduce sparsely-activated expert modules, allowing modality-specific processing while maintaining low inference overhead. CL-MoE (Huai et al., 2025) further extends MoE for continual learning in vision-language tasks, employing dual routers to balance generalization and retention. Furthermore, MoExtend (Zhong et al., 2024) introduces modular extension mechanisms that facilitate the adaptation of pretrained models to new tasks and modalities, thereby significantly reducing the computational cost associated with full model retraining.

These approaches illustrate that MoE architectures not only enhance computational efficiency but also offer increased flexibility in handling heterogeneous multimodal inputs, thereby establishing MoE as a compelling framework for scaling BERT-based models to complex multimodal tasks.

3 Methodology

3.1 Overview

As shown in Figure 1, the MMBERT framework consists of a text tokenizer, word embedding layer, vision and speech encoders, modality aligners, MoE-scaled BERT blocks, and a classification head. Modality aligners project non-text inputs into a shared linguistic space, enabling effective multimodal fusion. The MoE layers are integrated into the BERT encoder to dynamically route representations across modalities, improving detection accuracy. MMBERT is trained in three sequential stages: Modality aligner training, modality-specific expert training, and MMBERT tuning using a diverse collection of multimodal Chinese hate speech data.. The detailed architectural and training settings are provided in Appendix A

3.2 MMBERT Architecutre

Multimodal data generation. To synthesize the visual and audio data of corresponding text input, we employ the Kokoro text-to-speech model (Kaneko et al., 2022) to generate speech data corresponding to the input text. For the visual modality, we render a sequence of word-level font images representing each token in the text, thereby producing a visual analogue of the input.

Aligners. To enable the effective transformation of heterogeneous modality inputs into a unified linguistic representation space, MMBERT leverages the pretrained visual-language framework LLaVA (Liu et al., 2023) and the speech-language framework SpeechT5 (Ao et al., 2021). Specifically, for visual encoding, we adopt the CLIP-base-Chinese model (Yang et al., 2022), followed by a linear projection layer that maps the extracted visual features into soft image tokens compatible with the embedding space of BERT (Devlin et al., 2019). For speech, we utilize the encoder from the Whisper-base-Chinese speech recognition model (Radford et al., 2023), likewise augmented with a linear projection layer to project speech features into the same shared linguistic space. The alignment process is formally defined as follows:

$$X = \{T, \{I_1, \dots, I_k\}, S\} \tag{1}$$

$$T = \text{WordEmbedding}(\text{Tokenizer}(T)) \tag{2}$$

$$S = \text{SpeechAligner}(\text{Whisper}(S)) \tag{3}$$

$$I_i = \text{VisionAligner}(\text{CLIP}(I_i)) \tag{4}$$

$$V = [I_1, \dots, I_k] \tag{5}$$

where $\{T, \{I_1, \dots, I_k\}, S\}$ represents the text, images and speech inputs respectively. The $SpeechAligner$ and $VisionAligner$ modules are implemented as learnable linear projections that transform modality-specific features into a shared language embedding space. The sequence of word-level font image embeddings is concatenated to form the final visual token sequence.

MMBERT blocks. By the above aligners, we could obtain the encoded embedding of different modalities aligned in unified language domain. We concatenate the different modality embeddings as the final input to the MMBERT blocks. We denote the text, speech, vision embedding representations to $T = \{T_1, \dots, T_n\}$, $S = \{S_1, \dots, S_m\}$ $V = \{V_1, \dots, V_k\}$ respectively, where n, m, and k correspond to the respective sequence lengths of each modality. The MMBERT block computation proceeds as follows:

$$X_{l_0} = [T_1, \dots, T_n;\ S_1, \dots, S_m;\ V_1, \dots, V_k] \tag{6}$$

$$X_{l_j}^a = \text{Self-Atten}(\text{LN}(X_{l_{j-1}})) + X_{l_{j-1}} \tag{7}$$

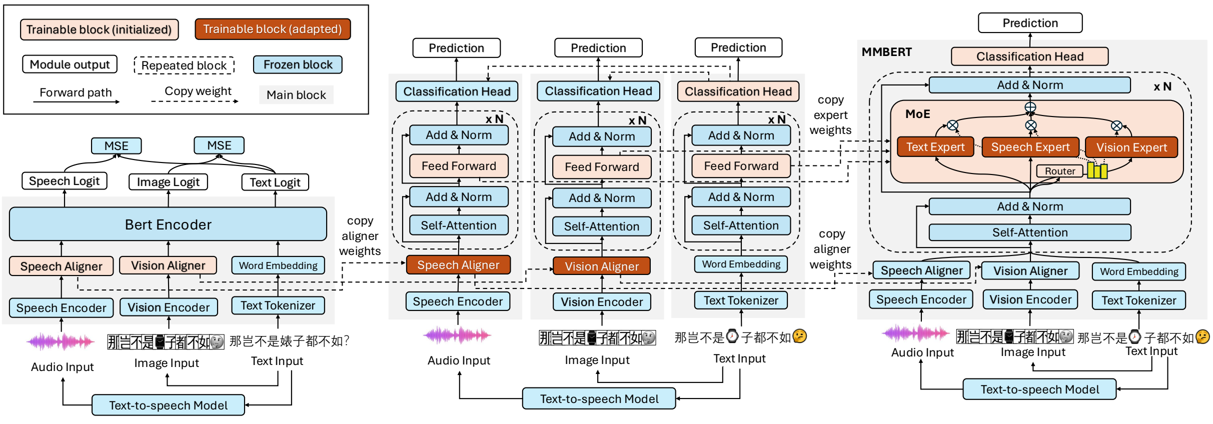
$$X_{l_j} = \text{MoE}(\text{LN}(X_{l_j}^a)) + X_{l_j}^a \tag{8}$$

where $LN(\cdot)$ refers to layer normalization, the $X_{l_j}^a$ represents the output latent of the self attention layer in the j th MMBERT block, X_{l_j} represents the output latent of j the MMBERT block. The MoE mechanism incorporates a set of experts $E = \{E_T, E_S, E_V\}$ each implemented as a feedforward neural network. A lightweight routing module, implemented as a linear transformation, computes the routing weights that determine the contribution of each modality-specific expert. The process is formally defined as:

$$P(X_l^a)_T = \frac{e^{f(X_l^a)_T}}{\sum_{i=\{T,S,V\}} e^{f(X_l^a)_i}} \tag{9}$$

$$P(X_l^a)_S = \frac{e^{f(X_l^a)_S}}{\sum_{i=\{T,S,V\}} e^{f(X_l^a)_i}} \tag{10}$$

$$P(X_l^a)_V = \frac{e^{f(X_l^a)_V}}{\sum_{i=\{T,S,V\}} e^{f(X_l^a)_i}} \tag{11}$$



(a) Stage 1

(b) Stage 2

(c) Stage 3

Figure 2: **Illustration of MMBERT Training strategy.** (a) Stage 1: Aligner training, (b) Stage 2: Expert training, (c) Stage 3: MMBERT tuning

$$\mathrm{MoE}(X_l^a) = \sum_{i=\{T,S,V\}} (P(X_l^a)_i \cdot E_i(X_l^a)) \quad (12)$$

where the $f(\cdot)$ denotes the routing function of different modalities implemented as a linear layer, the output weight logits are normalized by a softmax function. The final MoE output is weighted combination of the different modality-specific expert outputs.

3.3 MMBERT three-stage training strategy

To capitalize on the effectiveness of multi-expert collaboration—where each expert possesses distinct capabilities—while retaining the rich contextual and syntactic knowledge encoded in the original BERT model through large-scale pretraining, we propose a three-stage progressive training strategy to facilitate the incremental development of MMBERT. As shown in Figure 2, the training process is structured into three progressive stages to enhance the efficacy of multi-expert collaboration through an incremental learning strategy.

Stage 1: Aligner Training. The primary objective of the initial stage is to establish effective interoperability between heterogeneous modalities and linguistic representations. Modality-specific MLPs serve as aligners that project inputs from speech and vision into soft token embeddings. These aligners are trained by minimizing the mean squared error between the modality embeddings and the BERT-encoded textual representations. To improve the model's sensitivity to perturbed speech samples, speech and image representations generated from the perturbed text are aligned with those derived from the corresponding unperturbed text representations during the training process.

Stage 2: Expert Training. In this stage, modality-specific experts are trained independently using cross-modal data to specialize in their respective domains. Training continues to be guided by the minimization of cross-entropy loss, while the trained aligners weights in the first stage are adapted and further trained to better capture and represent the unique characteristics inherent to their respective modalities on the Chinese hate speech classification task. To facilitate the projection of heterogeneous modality data into a unified linguistic representation space by both the aligners and experts, the classification head originally trained on textual input is shared across other modalities.

Stage 3: MMBERT Tuning. The final stage integrates the trained experts into the MoE layers of MMBERT. A context-aware routing mechanism dynamically assigns input representations to appropriate experts based on semantic relevance. To prevent unbalanced expert weight distribution, an auxiliary loss is applied to encourage uniform expert utilization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cross-entropy}} + \alpha \cdot \mathcal{L}_{\text{aux}} \quad (13)$$

$$\mathcal{L}_{\text{aux}} = N \cdot \sum_{i=1}^{N} p_i \cdot f_i \quad (14)$$

where N denotes the total number of experts, α represents the weighting coefficient p_i represents the proportion of sequences routed to expert i, and f_i is the average gating probability assigned to expert i. The classification head is fine-tuned jointly, to improve multimodal fusion and generate the final prediction.

Model	ToxiCloakCN				ToxiCN				COLD			
	Acc	Pre	Rre	F1	Acc	Pre	Rre	F1	Acc	Pre	Rre	F1
Finetuned Models												
BERT	80.6	80.5	80.7	86.6	87.8	88.0	87.7	87.8	81.2	80.7	82.1	80.9
BERT-wwm	80.0	80.4	80.3	87.9	88.0	88.1	88.9	88.0	82.0	81.6	83.2	81.8
RoBERTa	81.1	82.4	81.3	82.6	88.8	88.9	89.5	89.6	82.6	81.9	83.7	82.5
ChineseBERT	86.3	87.5	86.2	86.8	90.8	89.4	90.3	90.6	82.4	81.3	83.1	82.2
MMBERT (ours)	**94.3**	**94.4**	**95.7**	**95.2**	**93.3**	**91.4**	**93.2**	**92.2**	**84.2**	**84.1**	**86.3**	**85.8**
LLM APIs (Prompt template with examples)												
GPT3.5	55.5	60.5	55.5	49.5	60.7	63.7	60.7	58.5	65.2	73.6	64.9	61.3
GPT4-o	64.5	68.8	64.6	62.4	78.1	79.9	78.1	77.8	71.5	73.4	71.5	70.9
LLAMA-3-8B	68.2	68.2	68.1	68.1	74.2	74.2	74.1	74.1	70.6	70.8	70.6	70.6
Qwen2.5-7B	66.0	66.7	66.0	65.6	76.4	77.3	76.4	76.2	74.7	76.1	74.7	74.3
DeepSeek-v3	64.6	72.3	64.5	61.2	72.9	77.5	72.8	71.7	73.1	75.4	73.1	72.5

Table 1: Performance comparison across models and datasets, including accuracy, macro precision, macro recall, and macro F1 Score.

4 Experiment

4.1 Baselines

To establish a comprehensive evaluation framework, we consider both encoder-based and decoder-based language models as baselines. Specifically, we adopt several BERT-based models with a fully connected classification layer as encoder-based baselines, and utilize LLMs with structured task-specific prompts as decoder-based baselines.

Encoder-Based BERT Models. As representative encoder-based BERT models, we select three widely adopted Chinese pretrained BERT-based encoders: **BERT**[1] (Devlin et al., 2019), **BERT-wwm**[2] (Sun et al., 2019) and **RoBERTa**[3] (Liu et al., 2019). Each model is fine-tuned by attaching a fully connected layer on top of the pooled output from the encoder to perform classification. In addition, we include **ChineseBERT** (Sun et al., 2021), a recently proposed model that integrates lexicon and phonological features into the standard BERT architecture, to examine its performance under the same experimental settings.

Decoder-Based LLMs. For LLM baselines, we assess the performance of several state-of-the-art LLMs, including **GPT-3.5** (Brown et al., 2020), **GPT-4o** (OpenAI, 2024), **LLaMA-3-8B** (Meta AI, 2024), **Qwen2.5-7B** (Alibaba, 2024), and **DeepSeek-v3** (DeepSeek, 2024). These models are evaluated under a unified prompt-based inference framework. This setup ensures consistency across different models and enables a fair comparison with encoder-based models, particularly in light of the substantial differences in model scales.

4.2 Dataset

To evaluate the proposed MMBERT framework, we conduct experiments on three Chinese hate speech datasets that collectively support comprehensive and robust assessment. **ToxiCN** (Lu et al., 2023) provides 12,011 samples of standard hate speech annotations for naturally occurring Chinese text, serving as a baseline for evaluating classification performance. **ToxiCloakCN** (Xiao et al., 2024b) introduces 4,582 cloaking perturbed examples in code-mixing and homophonic substitution, specifically designed to evade text-only detectors while preserving hateful intent, making it essential for testing model robustness against cloaking strategies. Finally, **COLD** (Deng et al., 2022) extends evaluation to a wider spectrum of offensive content with 37,480 samples, offering insight into a model's generalizability across various forms of online toxicity. Together, these datasets form a diverse and challenging benchmark suite for assessing both accuracy and adversarial resilience in Chinese hate speech detection.

4.3 Evaluation method

We employ the widely used metrics of accuracy (**Acc**), macro precision (**Pre**), macro recall (**Rre**)

[1] https://huggingface.co/bert-base-chinese

[2] https://huggingface.co/hfl/chinese-bert-wwm-base

[3] https://huggingface.co/hfl/chinese-roberta-wwm-ext

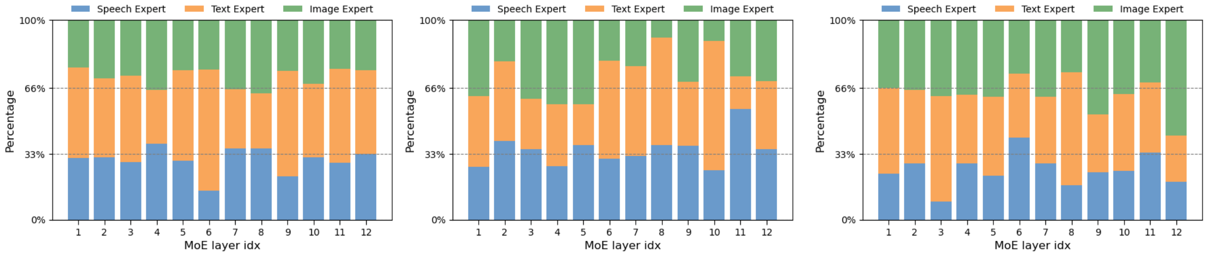


Figure 3: Distribution of expert loading with different input perturbation types, *left*: non perturbation, *middle*: homophonic perturbation, *right*: code-mixing perturbation

and macro F_1-score (**F1**) to evaluate the classification performance of models. For the BERT-based models in the baselines, we fine-tune and reserve the best performing models and hyperparameters on the test set. The models are trained using the cross-entropy loss function and optimized with the AdamW optimizer. An early stopping strategy is implemented to prevent overfitting during training. All datasets are partitioned into training and test sets using an 8:2 split ratio. For the LLMs in the baselines, we perform few-shot learning with a basic prompt including task definitions, output formats and specific prediction requirements for all elements, with a hate speech example and a non-hate speech example, details can be found in appendix B. All experiments are conducted using a NVIDIA H100 Tensor Core GPU.

4.4 Result and Discussion

4.4.1 Main result

Table 1 presents a comprehensive evaluation, MMBERT consistently outperforms the finetuned BERT-based baseline models and LLMs with in-context learning across three benchmarks, demonstrating superior performance and robustness in both standard and adversarial settings.

On ToxiCloakCN, which features cloaking perturbed hate speech, MMBERT achieves a macro F1 score of 95.2, substantially outperforming the best finetuned baseline, ChineseBERT, which reaches 86.8. Other BERT-based models such as RoBERTa and BERT-wwm show a further drop in performance. The strong results on ToxiCloakCN indicate that MMBERT is particularly effective at handling cloaking strategies such as character deformation, homophonic substitution, and code-mixing. Performance on ToxiCN, a standard hate speech benchmark, follows a similar trend. MMBERT achieves an F1 score of 92.2, improving upon ChineseBERT by 1.6 points and RoBERTa by 2.6 points. The gains are consistent across accuracy, precision, and recall, indicating MMBERT's well-rounded classification ability. On COLD, a more diverse and open-domain dataset, MMBERT again achieves the highest macro F1 score of 85.8. While ChineseBERT and RoBERTa remain competitive, they fail to match MMBERT's performance, particularly in recall, which is crucial for detecting subtle or implicit hate speech.

In contrast, LLM APIs perform significantly worse across all benchmarks. Even with prompting and examples, GPT-4o and DeepSeek-v3 achieve only 62.4 and 61.2 F1 on ToxiCloakCN, respectively. LLAMA-3 and Qwen2.5 models show similar limitations, especially in the presence of cloaked content. These results underscore the limitations of few-shot prompting approaches for domain-specific, adversarial tasks, and highlight the effectiveness of MMBERT's task-specific, multimodal training.

Overall, the results confirm that MMBERT not only outperforms existing baselines in Chinese hate speech detection but also exhibits strong resilience against cloaking perturbed samples, validating the importance of multimodal integration for the Chinese hate speech detection.

4.4.2 Routing distribution analysis

We analyze the average routing weight distribution of different experts in MMBERT 12 MoE layers under three hate speech perturbation categories in the ToxiCloakCN dataset as shown in Figure 3.

In the non-perturbed setting, the model primarily routes to the text expert, especially in middle layers, reflecting the dominance of textual semantics. Speech and image experts contribute consistently, with image usage slightly increasing in deeper layers. Under homophonic perturbation, the model shifts toward the speech expert in early and middle layers, leveraging phonetic cues to resolve ambiguities introduced by homophones. Vision expert assigned weight decreases slightly, while

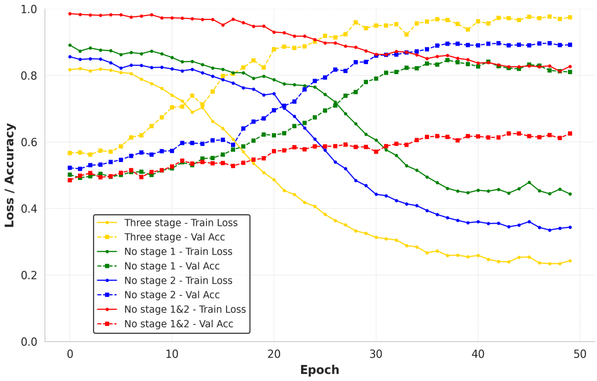


Figure 4: Ablation study evaluating the impact of each stage in the proposed three-stage training strategy

text routing remains stable. In the code-mixing scenario, image experts dominate across most layers, indicating reliance on visual context to address multilingual inconsistencies. Text experts are also more engaged in earlier layers, while speech expert weight declines.

These patterns demonstrate MMBERT adaptive routing behavior, where expert activation is dynamically adjusted based on input characteristics, enhancing robustness against modality-specific perturbations.

4.4.3 Ablation study

Training strategy. We conduct an ablation study to evaluate the effectiveness of the progressive three-stage training strategy for integrating MoE into MMBERT. Specifically, we compare the full pipeline with three variants: without aligner training stage (stage 1), without expert training stage (stage 2), and without both stages. All models are trained for 50 epochs on the ToxiCloakCN dataset under identical settings.

As shown in Figure 4, the full three-stage strategy achieves the best overall performance, with the lowest training loss and highest validation accuracy. It enables stable convergence and strong generalization, indicating that gradual modality alignment and expert specialization are both essential for effective multimodal learning. Without aligner pretraining, convergence is slower and validation performance is less stable, suggesting suboptimal cross-modal mapping. Removing expert specialization also leads to reduced accuracy and higher loss, showing that expert-specific representation learning is crucial. The worst performance is observed when both stages are removed, as the model quickly overfits and fails to generalize. These results demonstrate that each stage of the proposed training strategy plays a critical role in enabling MMBERT to effectively detect cloaked hate speech across modalities.

Dataset	**Text&Speech**		**Text&Vision**	
	Acc	**F1**	**Acc**	**F1**
ToxiCloakCN	91.2	91.1	87.7	86.6
ToxiCN	90.1	90.9	88.9	89.3
COLD	83.1	83.8	82.7	81.9

Table 2: Ablation study evaluating the impact of each modality in the MMBERT framework

Modality. To assess the contribution of each modality in the MMBERT framework, we perform an ablation study by evaluating model performance by scaling with single modality, using text paired with either speech or vision. As shown in Table 2, the text and speech combination consistently outperforms the text and vision setting across all three datasets. On the ToxiCloakCN dataset, the F1 score reaches 91.1 when using speech compared to 86.6 when using vision, indicating that speech features are more effective in capturing adversarial cues introduced by cloaking perturbations. This trend is also observed on ToxiCN and COLD, where the text and speech setting yields stronger results. These findings suggest that speech contributes more complementary information than vision and plays a critical role in improving robustness in Chinese hate speech detection.

5 Conclusion

We presents MMBERT, a multimodal framework for Chinese hate speech detection that effectively incorporates text, speech, and vision using the MoE architecture. To ensure stable integration of heterogeneous modalities, we introduce a progressive training strategy that proves critical for effective optimization. Empirical results across multiple benchmarks show that MMBERT achieves strong performance, particularly under adversarial conditions involving cloaked perturbations. Ablation studies confirm the importance of both the training strategy and modality fusion, with speech contributing most significantly to robustness. Our findings highlight the potential of task-specific multimodal modeling for addressing complex language understanding challenges, particularly in safety-critical domains like Chinese hate speech detection.

Limitation

While MMBERT demonstrates strong performance in detecting cloaked hate speech, several limitations remain. First, the current evaluation relies on a limited set of Chinese datasets, which does not fully capture the breadth and complexity of obfuscation strategies used in real-world settings. The dataset is constrained in both scale and diversity, covering only a subset of character-level, phonetic, and visual perturbations commonly found in adversarial discourse. This restricts the model's ability to generalize to more nuanced, creative, or evolving forms of cloaked hate speech. Expanding the dataset to include a wider variety of perturbation types, sociolinguistic contexts, and user-generated adversarial patterns would be essential for advancing robustness.

Second, the current study is limited to Chinese language data, and it remains unclear how well the similar method would transfer to other languages or cultural environments where obfuscation strategies may differ significantly in structure and intent. Cloaking techniques can be highly language-specific, depending on orthographic systems, phonetics, and sociocultural norms.

Future work should consider explore cross-lingual adaptations and evaluate the generalizability of similar method in multilingual or multicultural settings

Ethical Statement

This work focuses on detecting hate speech on Chinese social media platforms using a multimodal framework. Given the sensitive nature of hate speech detection, we took several ethical precautions throughout the research process. All datasets used in this study are publicly available or released under terms that permit academic use. No personally identifiable information is included in the data.

We acknowledge the potential risks associated with misuse of automated hate speech detection systems, such as censorship or the marginalization of certain user groups. To mitigate this, our model is designed for research purposes only and we do not advocate its direct deployment without thorough evaluation by domain experts and consideration of social and legal implications.

We also recognize that hate speech is a socially and culturally contextual phenomenon. While our model is tailored for Chinese-language content, we emphasize the importance of local expertise when interpreting results or extending this work to other languages or communities.

Bias mitigation and fairness were considered in model evaluation. To address this, we adopt diverse and representative datasets covering different forms of hate speech related to race, gender, region, and LGBTQ+ communities.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Alibaba. 2024. Qwen2.5: Alibaba Cloud's Open-Source Language Model. https://huggingface.co/Qwen. Accessed: 2025-05-19.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, and 1 others. 2021. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. *arXiv preprint arXiv:2110.07205*.

Alberto Benayas, Miguel Angel Sicilia, and Marçal Mora-Cantallops. 2024. A comparative analysis of encoder only and decoder only models in intent classification and sentiment analysis: Navigating the trade-offs in model size and performance. *Language Resources and Evaluation*, pages 1–24.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

August FY Chao, Chen-Shu Wang, Bo-Yi Li, and Hong-Yan Chen. 2024. From hate to harmony: Leveraging large language models for safer speech in times of covid-19 crisis. *Heliyon*, 10(16).

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

DeepSeek. 2024. DeepSeek-V3: Open-Source Language Model. https://huggingface.co/DeepSeek-AI. Accessed: 2025-05-19.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. Cold: A benchmark for chinese offensive language detection. *Preprint*, arXiv:2201.06025.

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, volume 1 (long and short papers)*, pages 4171–4186.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. 2013. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*.

Faeze Ghorbanpour, Daryna Dementieva, and Alexander Fraser. 2025. Can prompting llms unlock hate speech detection across languages? a zero-shot and few-shot study. *arXiv preprint arXiv:2505.06149*.

Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, and Liang He. 2025. Cl-moe: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. *arXiv preprint arXiv:2503.00413*.

Takuhiro Kaneko, Kou Tanaka, Hirokazu Kameoka, and Shogo Seki. 2022. istftnet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time fourier transform. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6207–6211. IEEE.

Haixia Wang Lan. 2006. Introduction to rhetoric. *China Review International*, 13(2):533–535.

Bin Li, Yan Dou, Yingting Cui, and Yuqi Sheng. 2020. Swearwords reinterpreted: New variants and uses by young chinese netizens on social media platforms. *Pragmatics*, 30(3):381–404.

Jiachen Li, Xinyao Wang, Sijie Zhu, Chia-Wen Kuo, Lu Xu, Fan Chen, Jitesh Jain, Humphrey Shi, and Longyin Wen. 2024. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. *Advances in Neural Information Processing Systems*, 37:131224–131246.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2025. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15.

Yunxin Li, Yu Zhao, Baotian Hu, Qingcai Chen, Yang Xiang, Xiaolong Wang, Yuxin Ding, and Lin Ma. 2021. Glyphcrm: Bidirectional encoder representation for chinese character with its glyph. *arXiv preprint arXiv:2107.00395*.

Dengyi Liu, Minghao Wang, and Andrew G Catlin. 2024. Detecting anti-semitic hate speech using transformer-based large language models. *arXiv preprint arXiv:2405.03794*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

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*.

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023. Facilitating fine-grained detection of chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. *arXiv preprint arXiv:2305.04446*.

Meta AI. 2024. LLaMA 3 Technical Report. https://ai.meta.com/llama/. Accessed: 2025-05-19.

OpenAI. 2024. GPT-4o: OpenAI's Newest Multimodal Model. https://openai.com/index/gpt-4o. Accessed: 2025-05-19.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. *arXiv preprint arXiv:2012.15674*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Hafiz Muhammad Raza Ur Rehman, Mahpara Saleem, Muhammad Zeeshan Jhandir, Eduardo Silva Alvarado, Helena Garay, and Imran Ashraf. 2025. Detecting hate in diversity: a survey of multilingual code-mixed image and video analysis. *Journal of Big Data*, 12(1):1–28.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. *arXiv preprint arXiv:2106.16038*.

Adrian Tien, Lorna Carson, and Ning Jiang. 2021. *An Anatomy of Chinese Offensive Words*. Springer.

Yunze Xiao, Houda Bouamor, and Wajdi Zaghouani. 2024a. Chinese offensive language detection: Current status and future directions. *arXiv preprint arXiv:2403.18314*.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-wei Lee. 2024b. Toxicloakcn: Evaluating robustness of offensive language detection in chinese with cloaking perturbations. *arXiv preprint arXiv:2406.12223*.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2021. Moefication: Transformer feed-forward layers are mixtures of experts. *arXiv preprint arXiv:2110.01786*.

Shanshan Zhong, Shanghua Gao, Zhongzhan Huang, Wushao Wen, Marinka Zitnik, and Pan Zhou. 2024. Moextend: Tuning new experts for modality and task extension. *arXiv preprint arXiv:2408.03511*.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023. Cross-cultural transfer learning for chinese offensive language detection. *arXiv preprint arXiv:2303.17927*.

A MMBERT setting

MMBERT is built upon the `BERT-base-chinese`[4] encoder, which serves as the backbone for textual representation. For modality-specific feature extraction, we employ a vision encoder based on `chinese-clip-vit-base-patch16`[5] and a speech encoder based on `whisper-base`[6]. Each modality is passed through a dedicated aligner, implemented as a lightweight two-layer MLP, to project the modality-specific features into the BERT embedding space, thereby forming unified token representations. These representations are processed by modified BERT layers in which the original feed-forward networks are replaced by Mixture-of-Experts (MoE) layers. Each MoE layer contains modality-specific experts and a shared self-attention mechanism, with a context-aware routing function that dynamically assigns token sequences to appropriate experts. A classification head is applied to the final output to produce predictions.

Training is performed in three progressive stages. In stage 1, modality aligners are pretrained using synthetic parallel data to align visual and speech features with their corresponding textual embeddings. The learning rate in this stage is set to 1e-3. In stage 2, modality-specific experts are trained independently using cross-modal supervision, while aligners continue to adapt. During this phase, the learning rate for the aligners is maintained at 1e-3, the text expert at 5e-6, and the speech and vision experts at 5e-5. In stage 3, all components are jointly fine-tuned on the multimodal Chinese hate speech detection task using a cross-entropy loss. The learning rate in this final stage is set to 5e-4. To promote balanced utilization across experts, we incorporate an auxiliary load-balancing loss into the MoE layers, with a weighting coefficient of 1e-2. The model is trained for 50 epochs using the AdamW optimizer and a linear learning rate decay schedule. All experiments are conducted using PyTorch on NVIDIA A100 GPUs.

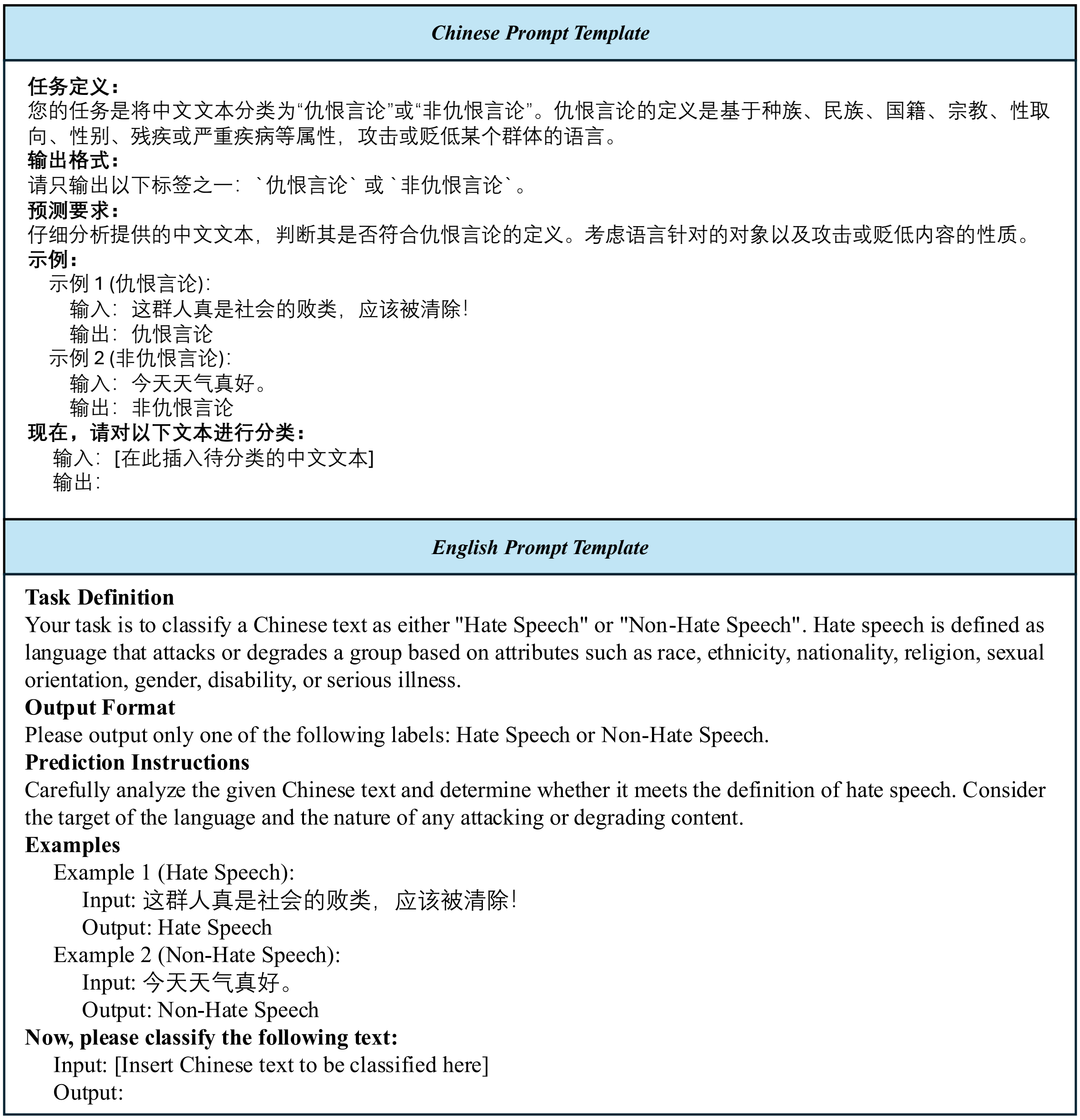
B LLM evaluation prompt template

The basic prompt of LLM for Chinese hate speech detection is shown in Figure 5

[4] https://huggingface.co/bert-base-chinese

[5] https://huggingface.co/OFA-Sys/chinese-clip-vit-base-patch16

[6] https://huggingface.co/openai/whisper-base



Chinese Prompt Template

任务定义：
您的任务是将中文文本分类为“仇恨言论”或“非仇恨言论”。仇恨言论的定义是基于种族、民族、国籍、宗教、性取向、性别、残疾或严重疾病等属性，攻击或贬低某个群体的语言。
输出格式：
请只输出以下标签之一：`仇恨言论` 或 `非仇恨言论`。
预测要求：
仔细分析提供的中文文本，判断其是否符合仇恨言论的定义。考虑语言针对的对象以及攻击或贬低内容的性质。
示例：
 示例 1 (仇恨言论):
  输入：这群人真是社会的败类，应该被清除！
  输出：仇恨言论
 示例 2 (非仇恨言论):
  输入：今天天气真好。
  输出：非仇恨言论
现在，请对以下文本进行分类：
 输入：[在此插入待分类的中文文本]
 输出：

English Prompt Template

Task Definition
Your task is to classify a Chinese text as either "Hate Speech" or "Non-Hate Speech". Hate speech is defined as language that attacks or degrades a group based on attributes such as race, ethnicity, nationality, religion, sexual orientation, gender, disability, or serious illness.
Output Format
Please output only one of the following labels: Hate Speech or Non-Hate Speech.
Prediction Instructions
Carefully analyze the given Chinese text and determine whether it meets the definition of hate speech. Consider the target of the language and the nature of any attacking or degrading content.
Examples
 Example 1 (Hate Speech):
  Input: 这群人真是社会的败类，应该被清除！
  Output: Hate Speech
 Example 2 (Non-Hate Speech):
  Input: 今天天气真好。
  Output: Non-Hate Speech
Now, please classify the following text:
 Input: [Insert Chinese text to be classified here]
 Output:

Figure 5: Chinese and English version of the LLM Chinese hate speech detection evaluation template