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

Qiyao Xue Yuchen Dou Ryan Shi Xiang Lorraine Li Wei Gao
University of Pittsburgh
{qix63, yud105, ryanshi, xianglli, weigao}@pitt.edu

Abstract

Hate speech detection on Chinese social networks presents 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 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 results in several Chinese hate speech datasets show that MMBERT significantly surpasses fine-tuned BERT-based encoder models, fine-tuned LLMs, and LLMs utilizing in-context learning approaches.

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 under-protected (Davidson et al. 2017; Davidson, Bhattacharya, and Weber 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. 2024). These subtle manipulations exploit the structural and phonological properties of the Chinese language, making detection especially difficult for text-only models.

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, Sicilia, and Mora-Cantallops 2024; Ghorbanpour, Dementieva, and Fraser 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, Wang, and Catlin 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. 2024).

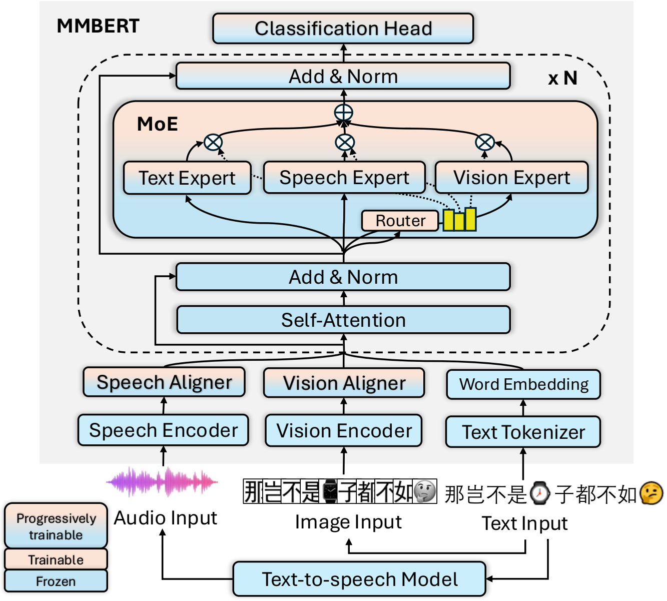


Figure 1: **Illustration of MMBERT model structure**. Compared to traditional BERT-based model, it leverages the MoE architecture to scale and effectively handle multiple modalities. A three-stage progressive training strategy is designed to ensure stable training and prevent performance degradation.

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.

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 and open-source LLM baselines and closed-source LLMs with in-context learning. Results demonstrate that MMBERT consistently achieves superior performance, particularly in detecting cloaking perturbed hate speech.

Background and Motivation

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 strategies to obfuscate offensive content while preserving its intended meaning (Xiao et al. 2024; Xiao, Bouamor, and Zaghouani 2024). It 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. Words with similar pronunciations are frequently substituted to generate alternative semantics (Tien, Carson, and Jiang 2021). For instance, Chinese internet users often replace the character “满” (meaning ‘full’) 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” “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 pinyin and emojis (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 “👅🐶” (meaning ‘licking dog’), the addition of an emoji amplifies the pejorative undertone, characterizing individuals perceived as excessively submissive in relationship contexts—analogous to the English term ‘sycophant’.

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 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, Bouamor, and Zaghouani 2024; Raza Ur Rehman et al. 2025).

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 and tonal ambiguity. 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.

Scaling Multimodal Language Models with MoE Architectures

Recent advancements in large MLLMs have increasingly explored the use of MoE (Eigen, Ranzato, and Sutskever 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 multimodal inputs, thereby establishing MoE as a compelling framework for scaling BERT-based models to complex multimodal tasks.

Methodology

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 model architecture, training setting and model efficiency information are provided in Appendix A.

MMBERT Architecture

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:

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

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

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

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

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

where $\{T, \{I_1, \ldots, I_k\}, S\}$ 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 $T = \{T_1, \ldots, T_n\}$, $S = \{S_1, \ldots, S_m\}$ $V = \{V_1, \ldots, V_k\}$ respectively, where n, m, and k correspond to the respective sequence lengths of each modality. The MMBERT block computation proceeds as follows:

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

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

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

where $LN(\cdot)$ refers to layer normalization, the $X^a_{l_j}$ 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 feed-forward neural network. A lightweight routing module, implemented as a linear transformation, computes the routing

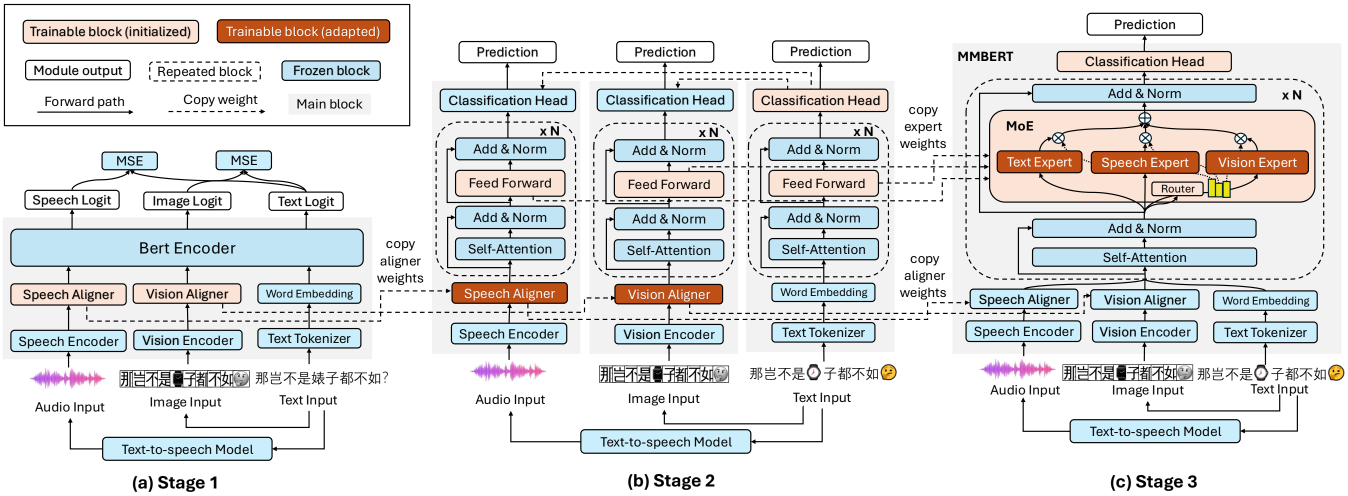


Figure 2: **Illustration of MMBERT Training strategy.** (a) Stage 1: Aligner training, (b) Stage 2: Expert training, (c) Stage 3: MMBERT tuning

weights that determine the contribution of each modality-specific expert. The process is formally defined as:

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

$$\text{MoE}(X_l^a) = \sum_{i=\{T,S,V\}} (P(X_l^a)_i \cdot E_i(X_l^a)) \tag{10}$$

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.

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:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cross-entropy}} + \alpha \cdot \mathcal{L}_{\text{aux}} \tag{11}$$

$$\mathcal{L}_{\text{aux}} = N \cdot \sum_{i=1}^{N} p_i \cdot f_i \tag{12}$$

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 generate the final prediction.

Experiments

Baseline

To establish a comprehensive evaluation framework, we consider both encoder-based and decoder-based language

Model	ToxiCloakCN				ToxiCN				COLD			
	Acc	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
Finetuned Models												
LLAMA3-8B	78.2	79.1	77.3	79.3	81.3	82.1	83.2	84.3	78.2	78.7	80.6	78.9
Qwen2.5-7B	82.1	83.6	84.1	83.7	86.8	87.1	88.2	87.9	79.6	79.8	81.3	81.1
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**

Table 1: Performance comparison of fine-tuned models across datasets with accuracy, precision, recall, and F1 scores.

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), **LLaMA3-8B** (Meta AI 2024), **Qwen2.5-7B&72B** (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 fair comparison with encoder-based models.

Dataset

To evaluate the proposed MMBERT, 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. 2024) 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 toxicity. Together, these datasets form a diverse and challenging benchmark suite for assessing both accuracy and adversarial resilience in Chinese hate speech detection.

Evaluation method

We employ the widely used metrics of accuracy (**Acc**), macro precision (**Pre**), macro recall (**Rec**) and macro F_1-score (**F1**) to evaluate the classification performance of models. For the BERT-based models and open source LLMs with relatively comparable parameter size with MMBERT in the baselines, we fine-tune and reserve the best performing models with hyperparameters on the test set. All datasets are partitioned into training, validation and test sets using an 8:1:1 split ratio with early stopping strategy to prevent overfit during training. For the LLMs in the baselines, we perform few-shot learning with a basic prompt temple with different few-shot learning and chain-of-thought (CoT) settings, details can be found in Appendix B. All experiments are conducted using a NVIDIA H100 Tensor Core GPU.

Result and Discussion

Main result Table 1 and 2 presents the evaluation of fine-tuned LLMs, BERT-based models and LLM APIs across the ToxiCloakCN, ToxiCN, and COLD benchmarks using accuracy, macro precision, macro recall, and macro F1 as metrics. MMBERT consistently achieves the highest scores across all three datasets, demonstrating both strong overall performance and robustness to adversarial perturbations. Specifically, MMBERT attains macro F1 scores of 95.2, 92.2, and 85.8 on ToxiCloakCN, ToxiCN, and COLD, respectively. Compared to the strongest fine-tuned baseline, ChineseBERT, these results represent improvements of 8.4, 1.6, and 3.6 points in macro F1. These gains highlight the effectiveness of integrating textual, speech, and visual modalities through the Mixture-of-Experts framework and the progressive three-stage training strategy, which jointly enhance the model's ability to capture phonological and visual cues indicative of cloaked hate speech.

Traditional encoder-based models, including BERT, RoBERTa, and ChineseBERT, perform competitively on ToxiCN and moderately well on COLD. However, their

[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

Model	ToxiCloakCN				ToxiCN				COLD			
	Acc	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**	**Acc**	**Pre**	**Rec**	**F1**
LLM APIs (2 unperturbed hate / non-hate speech examples)												
GPT-3.5	55.5	60.5	55.5	49.5	60.7	63.7	60.7	58.5	65.2	73.6	64.9	61.3
GPT-4o	64.5	68.8	64.6	62.4	76.2	76.8	76.3	76.4	71.5	73.4	71.5	70.9
LLAMA3-8B	**68.2**	68.2	**68.1**	68.0	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	68.3	64.5	66.2	72.9	77.5	72.8	71.7	73.1	75.4	73.1	72.5
Qwen2.5-72B	67.9	**69.2**	67.2	**68.1**	**77.3**	**78.6**	**77.1**	**77.9**	74.6	**77.1**	**75.3**	**74.7**
LLM APIs ((2 unperturbed & 2 perturbed hate / non-hate examples)												
GPT-3.5	55.3	61.2	55.7	49.8	60.3	63.5	61.2	58.2	65.4	73.7	65.1	61.4
GPT-4o	66.9	71.2	68.3	67.8	78.1	**79.9**	78.1	77.8	71.5	73.4	71.5	70.9
LLAMA3-8B	67.3	68.9	67.9	68.2	75.1	74.0	74.2	74.3	71.2	70.7	72.1	71.2
Qwen2.5-7B	65.9	66.5	66.4	66.1	77.2	78.6	77.2	77.1	75.2	76.3	74.7	75.8
DeepSeek-v3	68.2	**70.2**	67.1	65.2	73.8	77.1	74.3	73.7	75.9	**77.6**	74.2	75.3
Qwen2.5-72B	**71.2**	69.7	**71.1**	**68.3**	**78.4**	79.3	**78.2**	**78.6**	**76.9**	76.9	**76.2**	**76.1**
LLM APIs (2 unperturbed & 2 perturbed hate / non-hate examples & CoT)												
GPT-3.5	57.3	62.3	58.1	51.6	62.9	65.8	61.2	59.3	66.1	73.8	63.2	63.4
GPT-4o	71.5	72.1	67.6	69.3	79.4	81.2	79.9	79.8	74.2	76.4	74.3	73.8
LLAMA3-8B	70.1	69.2	66.4	68.2	76.4	73.8	75.2	74.8	71.4	70.3	70.8	70.7
Qwen2.5-7B	68.1	67.1	65.8	66.1	77.4	76.9	77.8	77.3	75.1	75.9	75.8	74.9
DeepSeek-v3	70.6	**72.4**	72.5	**71.6**	76.6	**81.5**	78.3	77.1	78.2	**81.3**	76.9	77.3
Qwen2.5-72B	**72.3**	71.8	**72.7**	70.3	**81.1**	80.7	**81.3**	**80.1**	**78.4**	78.5	**78.1**	**78.2**

Table 2: Performance comparison of LLM prompting across datasets with accuracy, precision, recall, and F1 scores.

performance drops substantially on ToxiCloakCN, confirming their vulnerability to character deformation, homophonic substitution, and code-mixing perturbations. In contrast, LLM APIs such as GPT-3.5, GPT-4o, LLaMA3-8B, Qwen2.5-7B, and DeepSeek-v3 show limited effectiveness in few-shot and perturbed settings. For example, GPT-4o achieves only 62.4 F1 on ToxiCloakCN under basic prompting, underscoring the insufficiency of in-context learning alone for this domain-specific and adversarial task.

Providing both unperturbed and perturbed examples, as well as incorporating CoT prompting, yields modest improvements for LLMs. GPT-4o, for instance, improves from 62.4 to 69.3 F1 on ToxiCloakCN under the CoT setting. Nevertheless, these enhancements remain far below the performance of MMBERT, indicating that domain-adaptive multimodal modeling is critical for robust detection rather than relying solely on prompting.

Across datasets, ToxiCloakCN poses the greatest challenge due to heavy use of cloaking perturbations, and MMBERT is the only model to surpass 90 F1 on this benchmark. ToxiCN represents standard hate speech detection, where all fine-tuned BERT variants perform strongly and MMBERT provides consistent incremental gains. COLD, as a more diverse and open-domain dataset, produces lower overall scores, yet MMBERT maintains the best recall, confirming its generalization to nuanced and implicit toxic language.

Overall, the results validate the task-specific multimodal modeling with MoE-based expert routing and progressive training for MMBERT substantially outperforms both fine-tuned text-only models and prompt-based LLMs, particularly in adversarial scenarios involving cloaked hate speech. Detailed failure case analyses are presented in Appendix C.

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 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.

Ablation study on training strategy We conduct an ablation study to evaluate the effectiveness of the progres-

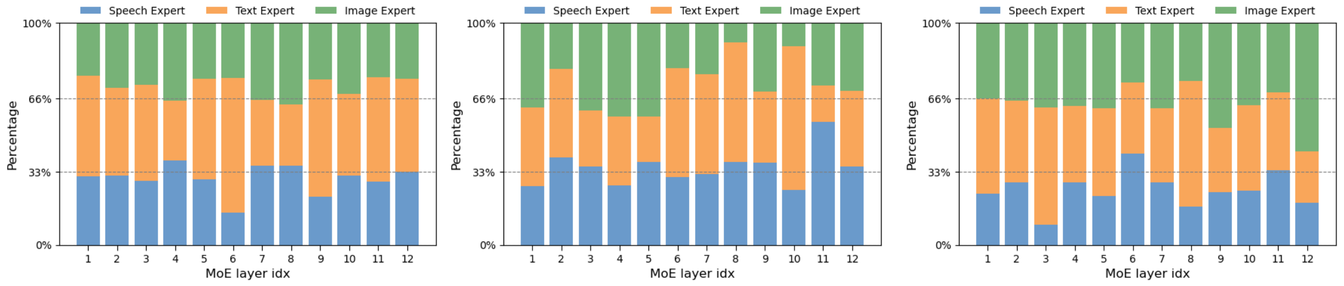


Figure 3: Distribution of expert loading with different input perturbation types, *left*: non perturbation, *middle*: homophonic perturbation, *right*: code-mixing perturbation

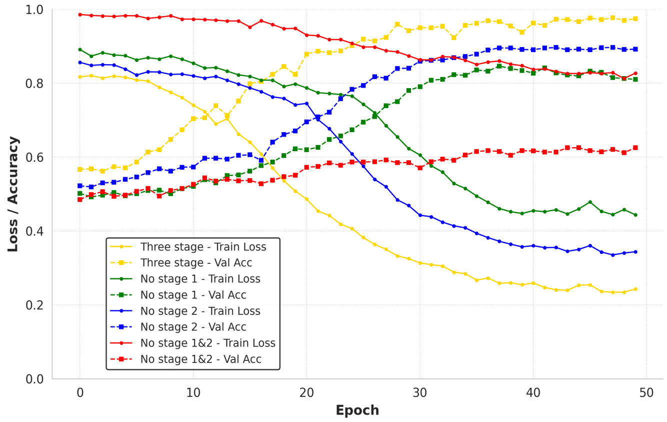


Figure 4: Ablation study evaluating the impact of each stage in the proposed three-stage training strategy

sive 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.

Ablation study on modalities To assess the contribution of each modality in the MMBERT framework, we perform an ablation study by scaling with single modality, using text paired with either speech or vision. As shown in Table 3, 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.

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

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 modalities, we introduce a progressive training strategy that proves critical for effective optimization. Ablation studies confirm the importance of both the training strategy and modality fusion, with speech contributing significantly to robustness. Empirical results across multiple benchmarks show that MMBERT achieves strong performance, particularly under adversarial conditions involving cloaked perturbations. 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.

Ethics Statement

This work involves Chinese hate speech detection with sensitive content. All datasets are publicly available and anonymized, and our models are intended solely for research to avoid potential bias and misuse.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 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.

Ao, J.; Wang, R.; Zhou, L.; Wang, C.; Ren, S.; Wu, Y.; Liu, S.; Ko, T.; Li, Q.; Zhang, Y.; et al. 2021. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. *arXiv preprint arXiv:2110.07205*.

Benayas, A.; Sicilia, M. A.; and Mora-Cantallops, M. 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*, 1–24.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chao, A. F.; Wang, C.-S.; Li, B.-Y.; and Chen, H.-Y. 2024. From hate to harmony: Leveraging large language models for safer speech in times of COVID-19 crisis. *Heliyon*, 10(16).

Davidson, T.; Bhattacharya, D.; and Weber, I. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 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, 512–515.

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

Deng, J.; Zhou, J.; Sun, H.; Zheng, C.; Mi, F.; Meng, H.; and Huang, M. 2022. COLD: A Benchmark for Chinese Offensive Language Detection. arXiv:2201.06025.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.

Eigen, D.; Ranzato, M.; and Sutskever, I. 2013. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*.

Ghorbanpour, F.; Dementieva, D.; and Fraser, A. 2025. Can Prompting LLMs Unlock Hate Speech Detection across Languages? A Zero-shot and Few-shot Study. *arXiv preprint arXiv:2505.06149*.

Huai, T.; Zhou, J.; Wu, X.; Chen, Q.; Bai, Q.; Zhou, Z.; and He, L. 2025. CL-MoE: Enhancing Multimodal Large Language Model with Dual Momentum Mixture-of-Experts for Continual Visual Question Answering. *arXiv preprint arXiv:2503.00413*.

Kaneko, T.; Tanaka, K.; Kameoka, H.; and Seki, S. 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)*, 6207–6211. IEEE.

Lan, H. W. 2006. Introduction to Rhetoric. *China Review International*, 13(2): 533–535.

Li, B.; Dou, Y.; Cui, Y.; and Sheng, Y. 2020. Swearwords reinterpreted: New variants and uses by young Chinese netizens on social media platforms. *Pragmatics*, 30(3): 381–404.

Li, J.; Wang, X.; Zhu, S.; Kuo, C.-W.; Xu, L.; Chen, F.; Jain, J.; Shi, H.; and Wen, L. 2024. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. *Advances in Neural Information Processing Systems*, 37: 131224–131246.

Li, Y.; Jiang, S.; Hu, B.; Wang, L.; Zhong, W.; Luo, W.; Ma, L.; and Zhang, M. 2025. Uni-MoE: Scaling Unified Multimodal LLMs with Mixture of Experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–15.

Li, Y.; Zhao, Y.; Hu, B.; Chen, Q.; Xiang, Y.; Wang, X.; Ding, Y.; and Ma, L. 2021. Glyphcrm: Bidirectional encoder representation for chinese character with its glyph. *arXiv preprint arXiv:2107.00395*.

Liu, D.; Wang, M.; and Catlin, A. G. 2024. Detecting antisemitic hate speech using transformer-based large language models. *arXiv preprint arXiv:2405.03794*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lu, J.; Xu, B.; Zhang, X.; Min, C.; Yang, L.; and Lin, H. 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.

Ouyang, X.; Wang, S.; Pang, C.; Sun, Y.; Tian, H.; Wu, H.; and Wang, H. 2020. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. *arXiv preprint arXiv:2012.15674*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Raza Ur Rehman, H. M.; Saleem, M.; Jhandir, M. Z.; Alvarado, E. S.; Garay, H.; and Ashraf, I. 2025. Detecting hate in diversity: a survey of multilingual code-mixed image and video analysis. *Journal of Big Data*, 12(1): 1–28.

Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; and Wu, H. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Sun, Z.; Li, X.; Sun, X.; Meng, Y.; Ao, X.; He, Q.; Wu, F.; and Li, J. 2021. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. *arXiv preprint arXiv:2106.16038*.

Tien, A.; Carson, L.; and Jiang, N. 2021. *An Anatomy of Chinese Offensive Words*. Springer.

Xiao, Y.; Bouamor, H.; and Zaghouani, W. 2024. Chinese offensive language detection: Current status and future directions. *arXiv preprint arXiv:2403.18314*.

Xiao, Y.; Hu, Y.; Choo, K. T. W.; and Lee, R. K.-w. 2024. ToxiCloakCN: Evaluating Robustness of Offensive Language Detection in Chinese with Cloaking Perturbations. *arXiv preprint arXiv:2406.12223*.

Yang, A.; Pan, J.; Lin, J.; Men, R.; Zhang, Y.; Zhou, J.; and Zhou, C. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1): 1.

Zhang, Z.; Lin, Y.; Liu, Z.; Li, P.; Sun, M.; and Zhou, J. 2021. Moefication: Transformer feed-forward layers are mixtures of experts. *arXiv preprint arXiv:2110.01786*.

Zhong, S.; Gao, S.; Huang, Z.; Wen, W.; Zitnik, M.; and Zhou, P. 2024. MoExtend: Tuning new experts for modality and task extension. *arXiv preprint arXiv:2408.03511*.

Zhou, L.; Cabello, L.; Cao, Y.; and Hershcovich, D. 2023. Cross-cultural transfer learning for Chinese offensive language detection. *arXiv preprint arXiv:2303.17927*.

Appendix A: MMBERT Details

Model Architecture

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 Setting

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. Excluding the parameters of the modality-specific encoders, the MMBERT architecture contains approximately 60 million trainable parameters. All experiments are conducted using PyTorch on NVIDIA A100 GPUs.

Model Efficiency

Parameter Count. The MMBERT model comprises 297.4 million parameters in total, including 162.4M in the backbone network (representing a 47% increase relative to BERT-base), 49M in the Whisper-base speech encoder, and 86M in the CLIP-base vision encoder.

Computational Cost. A single forward pass requires approximately 58.44 GFLOPs, which is the sum of 12×2.89 GFLOPs from the MMBERT layers, 21.2 GFLOPs from the Whisper-base encoder, and 2.56 GFLOPs from the

[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

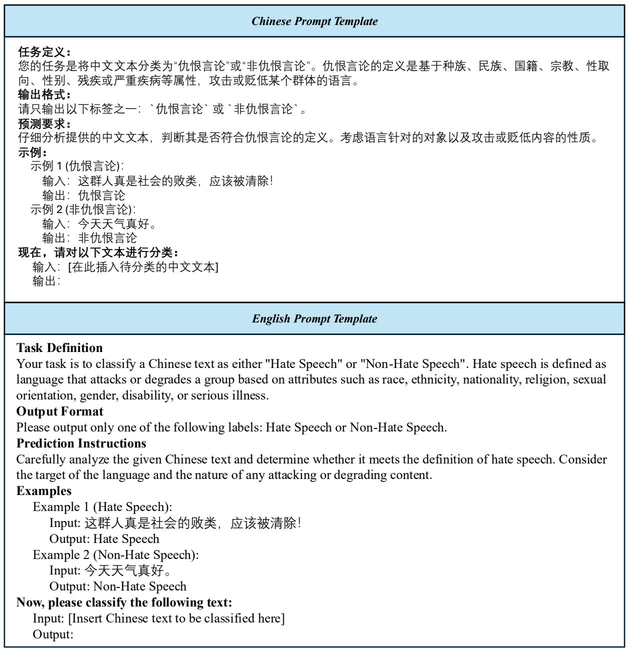
CLIP-base encoder. The contribution of the pooler and classifier heads is negligible.

Routing Overhead. The mixture-of-experts (MoE) layer routing introduces an additional 908.4 MFLOPs (12×75.8 MFLOPs), accounting for approximately 2.6% of the total computational cost.

Inference Latency. Under single-query inference with a sequence length of 128 on an NVIDIA H100 GPU, MMBERT achieves a latency of 6.3 ms in FP32 precision (compared to 3.5 ms for BERT-base) and 3.2 ms in FP16 precision (compared to 2 ms for BERT-base).

Appendix B: LLM evaluation prompt template

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



```
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

Appendix C: Failure Case Analysis

To better understand the limitations of MMBERT, we manually reviewed 50 misclassified samples from each test set. Two dominant failure modes emerged:

Cultural Context Gaps (38%)

False Positive Example (COLD):

“Taiwanese rednecks leave Weibo”

Root Cause: The model misclassifies culturally nuanced expressions as toxic due to limited coverage of regional dialects and sociopolitical context in the training data.

Mitigation Strategy: Diversify annotation teams with native speakers from multiple Chinese-speaking regions and include context-rich examples to reduce such errors.

Sarcasm and Reclaimed Terms (32%)

True Negative Example (ToxiCN):

“We gays are disgusting haha”

Root Cause: Binary toxicity labels lack contextual nuance. The model cannot distinguish reclaimed slurs or self-deprecating humor from genuine hate.

Mitigation Strategy: Introduce ternary labeling schemes (e.g., *hate*, *reclaimed*, *neutral*) or enrich the dataset with metadata such as speaker identity and intent.

These errors highlight that MMBERT is sensitive to cultural variation, sarcasm, and reclaimed language. Future work should explore context-aware annotations, richer label taxonomies, and sociolinguistic metadata to improve robustness in real-world deployment.