Penetrating Linguistic Disguises: A Slang-aware Label-Aligned Framework for Fine-Grained Toxicity Extraction in Chinese Hate Speech Detection

Wei Liu[1], Xiaoliang Chen[1,2,4], Duoqian Miao[2], Xu Gu[3], Xianyong Li[1], Yajun Du[1]

[1]School of Computer and Software Engineering, Xihua University, Chengdu 610039, China,
[2]College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China,
[3]Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610041, China,
[4]School of Computer Science, McGill University, Montreal, QC H3A 0G4, Canada

Correspondence: chenxl@mail.xhu.edu.cn

Abstract

Flexible word boundaries and linguistic obfuscation, particularly slang, challenge precise span-level hate speech detection in Chinese. While benchmarks such as STATE ToxiCN demand the exact extraction of Target-Argument-Hateful-Group quadruples, generative Large Language Models (LLMs) often fail strict boundary constraints. In contrast, discriminative 2D Grid Tagging methods frequently encounter label collisions. To resolve these problems, this study presents a Slang-aware Label-Aligned Framework. A Structural-Semantic Lexicon Fusion (SSLF) module reduces ambiguity by mapping obscure slang to explicit hate semantics. Additionally, the proposed Label-Disentangled Volumetric Tagging (LDVT) projects token interactions into a volumetric space. LDVT uses task-specific branches and dedicated label channels to structurally mitigate feature interference. This approach removes label collisions without heuristic post-processing. Empirical outcomes on STATE ToxiCN indicate a Hard-F1 of 30.09%. This performance is 5.82% higher than the best fine-tuned LLM baseline and confirms the method is effective for exact-match extraction[1].

1 Introduction

The rapid growth of online hate speech demands reliable automated detection systems (Bilewicz and Soral, 2020; Araújo et al., 2016). Early research focused on sentence-level classification, but the field now favors fine-grained span-level extraction to identify Targets and Arguments (Pavlopoulos et al., 2021; Zampieri et al., 2023). The STATE ToxiCN benchmark reflects this change and requires exact *Target-Argument-Hateful-Group* quadruples (Bai et al., 2025). As Table 1 shows, mapping unstructured text to structured quadruples is difficult, especially when toxic posts use intersectional grouping or slang. In Chinese, the lack of explicit word delimiters increases this difficulty (Deng et al., 2022), while widespread linguistic cloaking (e.g., homophones) aids in bypassing moderation (Xiao et al., 2024; Lu et al., 2023).

Current methods perform poorly here. Generative LLMs have strong semantic capabilities but suffer from boundary drift. These models frequently fail hard-match evaluations despite high soft-match scores; this discrepancy leads to misalignment and hallucination (Bai et al., 2025). In contrast, discriminative Grid Tagging Schemes (GTS) (Wu et al., 2020) provide better boundary sensitivity but face label collisions. Standard 2D grids need lossy label-space reduction to assign concurrent labels when toxicity categories overlap. Rapidly changing slang adds to these limits. Encoders frequently miss new symbols (e.g., "4000+" denoting death to the whole family), and naive lexicon concatenation is unstable. Considering the strict boundary requirements of STATE ToxiCN, this study proposes a specialized discriminative framework for precise structural extraction.

This study presents the Slang-aware Label-Aligned Framework (SLAF) for toxic quadruple extraction, which integrates two primary modules. First, Label-Disentangled Volumetric Tagging (LDVT) resolves label collisions. Unlike standard methods that flatten token interactions into a 2D grid (Wu et al., 2020), LDVT maps token pairs to a volumetric space of size $L \times L \times N$. The third dimension separates label channels. Each category occupies dedicated slices in this space, which aligns spans with their attributes and allows the decoding of overlapping relations without conflicts. Second, a Structural-Semantic Lexicon Fusion (SSLF) module handles linguistic disguises. SSLF does not treat external knowledge as fixed features (Jiang et al., 2022; Lu et al., 2023). Instead, it uses structure-aware attention to weight lexicon signals according to span topology. This process re-

[1]Our code address: https://github.com/HatsuneMikuLW/SLAF

Findings of the Association for Computational Linguistics: ACL 2026, pages 21111–21123

Exp.	Post	Target	Argument	Hateful	Group	Implicit slang
1	这个赞太少了点吧。 *That's kind of short on likes.*	赞 *likes*	太少了 *kind of short*	non-hate	Non-hate	-
2	txl去死。 *txl should die.*	txl (同性恋) *homosexual*	去死 *should die*	hate	LGBTQ	txl (同性恋) *homosexual*
3	正常男人不会喜欢黑女的。 *Normal men don't like Black women.*	黑女 *Black women*	不会喜欢黑女的 *don't like Black women*	hate	Sexism, Racism	-

Table 1: Examples of fine-grained toxicity extraction from the STATE ToxiCN benchmark. The task requires extracting precise *Target-Argument-Hateful-Group* quadruples. Examples highlight challenges such as implicit slang disguises (Exp. 2) and intersectional hate categories (Exp. 3).

covers cloaked meanings while preserving original semantics.

Tests on the STATE ToxiCN benchmark confirm the framework's effectiveness. With a focus on exact-match performance, SLAF achieves a Hard-F1 score of 30.09%. This result exceeds the leading fine-tuned LLM baseline by 5.82% and confirms the model's precision in boundary-sensitive extraction. The main contributions are:

- This paper presents SLAF, a discriminative framework for fine-grained toxicity extraction. It resolves linguistic cloaking and label collisions in one end-to-end architecture.
- The study develops two core components: SSLF uses structure-aware attention to fix slang-based obfuscations, and LDVT performs volumetric decoding to stop category conflicts in multi-label extraction.
- Results from the STATE ToxiCN benchmark show clear improvements. SLAF attains 30.09% Hard-F1, which is 5.82% higher than the best LLM baseline and proves the value of exact boundary extraction.

2 Related Work

2.1 Fine-grained Hate Speech Detection

Early hate speech detection prioritized sentence-level binary classification or coarse-grained category identification (Davidson et al., 2017; Fortuna and Nunes, 2018; Ali et al., 2022). However, the lack of explanatory detail in coarse labels moved the field toward fine-grained extraction (Mathew et al., 2021). Recent studies focus on locating toxic rationales (Pavlopoulos et al., 2021) and finding explicit targets (Zampieri et al., 2023). The STATE ToxiCN benchmark (Bai et al., 2025) reflects this trend and defines the task of extracting Target-Argument-Hateful-Group quadruples.

Baselines for this task usually follow two main approaches. Fine-tuned generative extractors based on open-source models (e.g., mT5-base, Mistral-7B, LLaMA3-8B) use task supervision but frequently miss strict span matching because of unconstrained generation. In contrast, in-context prompted LLM APIs (e.g., GPT-4o, DeepSeek-v3) offer strong semantic priors but lack stability under exact-match evaluation. Here, small boundary shifts cause heavy penalties in Hard-F1 scores and produce boundary-inconsistent fields (Bai et al., 2025). To fix this instability, this work revisits discriminative structured extraction. This approach focuses on boundary fidelity and maintains global consistency across quadruple fields.

2.2 Linguistic Cloaking and Knowledge Infusion

Chinese social media users frequently use linguistic cloaking (e.g., pinyin acronyms, homophones, emojis) to bypass keyword-based moderation. Consequently, implicit slang becomes a primary difficulty for hate speech detection (Xiao et al., 2024; Ahn et al., 2024; AlKhamissi et al., 2022). Recent studies have developed Chinese hate lexicons and fine-grained taxonomies to offer external lexical priors (Jiang et al., 2022; Lu et al., 2023; Bai et al., 2025). Yet, most lexicon-enhanced baselines inject these priors through simple matching-based fusion or concatenation. This method treats lexicon cues as static signals (Pamungkas et al., 2023). Such a design is unstable under compositional and overlapping slang spans because boundary ambiguity and spurious matches mislead predictions. Therefore, this work proposes SSLF, a structure-aware fusion mechanism. It conditions lexicon priors on contextual interactions to increase boundary-relevant evidence and reduce lexicon noise.

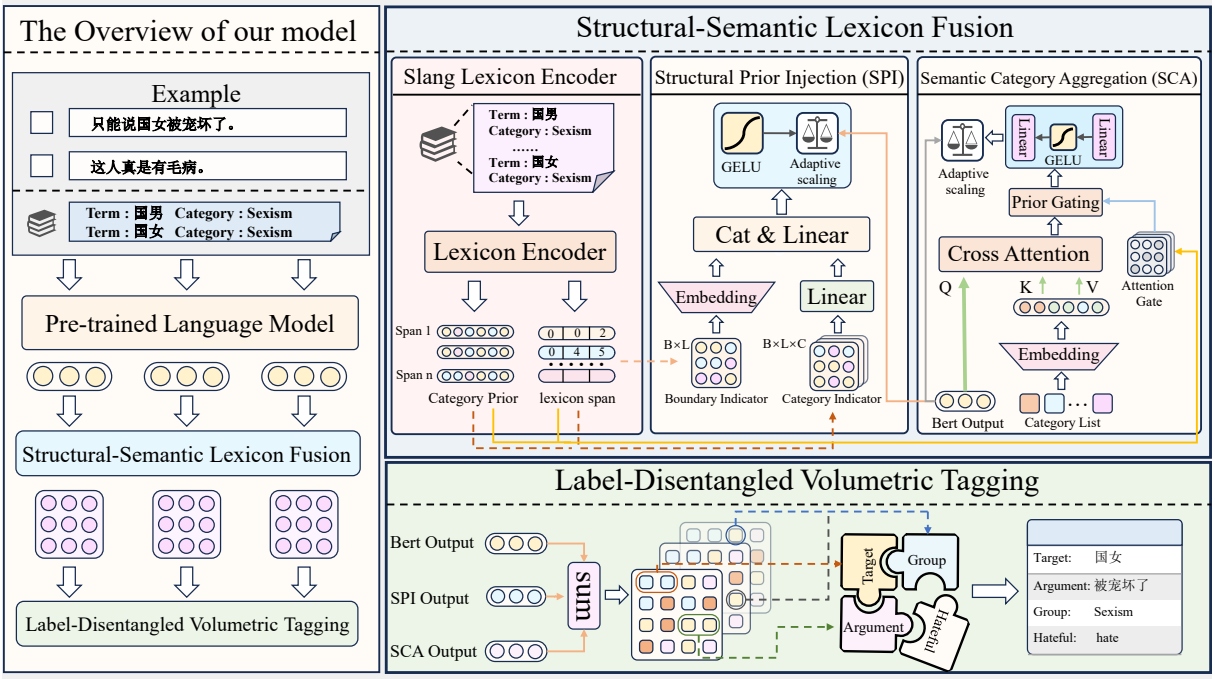

Figure 1: The overall architecture of SLAF.

2.3 Unified Information Extraction Architectures

Extracting toxic quadruples functions as Joint Entity and Relation Extraction. A unified model predicts spans and relations together to stop the error propagation found in pipeline designs. Recent generative paradigms, such as InstructUIE (Wang et al., 2023), try to unify extraction through instruction tuning. However, prior evaluations show that Large Language Models (LLMs) are unreliable under strict span-boundary requirements and schema-specific constraints (Han et al., 2023). As a result, discriminative methods are often better for exact-match evaluation. Representative 2D grid-based approaches, including GTS (Wu et al., 2020) and TPLinker (Wang et al., 2020), encode word-pair interactions and handle nested entities effectively. But in multi-category annotation, a single grid cell (span pair) might need to express multiple category labels for the same relation instance. This situation causes label collisions. Existing work reduces this issue via composite-label flattening or heuristic conflict resolution. These methods either inflate the label space or discard information. This study proposes LDVT to lift 2D grid tagging into a volumetric prediction space. It uses category-specific channels to explicitly separate overlapping labels, a process that removes the need for heuristic post-processing.

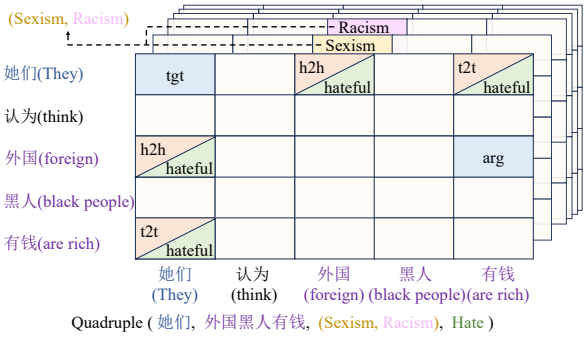

Figure 2: Example of grid labels with multi-label toxicity categories.

3 Methodology

Figure 1 illustrates the architecture of the Slang-aware Label-Aligned Framework (SLAF). This framework consists of three main parts: a PLM-based encoder, a Structural-Semantic Lexicon Fusion (SSLF) module for external knowledge injection, and a Label-Disentangled Volumetric Tagging (LDVT) decoder. The LDVT decoder specifically handles label collisions.

3.1 Problem Formulation

Following the grid tagging scheme in Li et al. (2023), the extraction task decomposes into four joint classification sub-tasks. Let $\mathcal{Y}^t$ be the set of predefined labels for task t, with the predicted label $y^t \in \mathcal{Y}^t$. The label spaces are: entity

boundary labels $y^{\text{ent}} \in \{tgt, arg, others\}$; entity pair labels $y^{\text{rel}} \in \{h2h, t2t, others\}$; hate labels $y^{\text{hate}} \in \{hate, non\text{-}hate, others\}$; and category labels $y^{\text{cat}} \in \{$*Sexism*, *Racism*, *Region*, *LGBTQ*, *Others*, *Non-hate*$\}$.

Figure 2 shows that a *tgt* label connecting the head and tail of "她们(They)" marks the target span. Similarly, an *arg* tag links the head "外国(foreign)" to the tail "有钱(are rich)" to define the argument span "外国黑人有钱(foreign black people are rich)". Labels *h2h* and *t2t* join the target head "她们(They)" with the argument head "外国(foreign)", and the target tail "她们(They)" with the argument tail "有钱(are rich)". This structure establishes a hateful relation between the components. Unlike these mutually exclusive tags, category prediction y^{cat} functions as a multi-label task. This design identifies co-occurring toxicities, such as simultaneous "Sexism" and "Racism".

3.2 PLM-based Encoder

Consider an input token sequence $X = \{x_1, \ldots, x_n\}$ of length n. Here, $x_1 = [\text{CLS}]$ represents the start token and $x_n = [\text{SEP}]$ serves as the separator. The original sentence tokens correspond to $\{x_2, \ldots, x_{n-1}\}$. A pre-trained language model (PLM) encodes X to produce a sequence of contextual token representations $\mathbf{H}$:

$$\mathbf{H} = \text{PLM}(X) \tag{1}$$

Let $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_n]$. In this representation, $\mathbf{h}_i$ denotes the contextual embedding of token x_i.

3.3 Structural-Semantic Lexicon Fusion

Standard PLMs frequently miss the obscure semantics of slang because of distribution shifts from general pre-training corpora (Caselli et al., 2020; Hanu and Unitary team, 2020). To fix this limitation, the Structural-Semantic Lexicon Fusion (SSLF) module integrates external knowledge. It operates through two sub-modules: Structural Prior Injection and Semantic Category Aggregation.

3.3.1 Slang Lexicon Encoder

The system uses a slang lexicon to integrate external knowledge by linking slang terms to hate categories. Matching these terms against the input yields two outputs: a lexicon span set $S^{\text{lex}} = \{(f_s, e_s)\}_{s=1}^{N_{\text{lex}}}$ containing N_{lex} identified matches, and a category prior matrix $P^{\text{lex}} \in \{0, 1\}^{N_{\text{lex}} \times N_{\text{cat}}}$. N_{cat} represents the total category count. In this matrix, the s-th row is a multi-hot vector that marks candidate categories for the s-th span.

3.3.2 Structural Prior Injection

This module embeds the boundaries and categorical attributes of potential slang terms directly into token-level representations. This integration solves the issue where standard tokenization fragments slang terms and breaks their semantic unity.

Lexicon Feature Construction. Lexicon encoder outputs transform into two sequence-level feature matrices: a boundary indicator matrix $\mathbf{B}^{\text{lex}} \in \mathbb{R}^{n \times 1}$ and a category indicator matrix $\mathbf{M}^{\text{lex}} \in \mathbb{R}^{n \times N_{\text{cat}}}$. $\mathbf{B}^{\text{lex}}$ applies discrete boundary indices (e.g., start, end) to tokens. Simultaneously, $\mathbf{M}^{\text{lex}}$ projects the multi-hot category vectors of matched spans onto their constituent tokens. A dense lexicon representation $\mathbf{G} \in \mathbb{R}^{n \times d}$ integrates these features through:

$$\mathbf{G} = \text{GELU}\left([\mathbf{E}_{bnd}; \mathbf{M}^{\text{lex}}\mathbf{W}_{cat}^{\top}]\mathbf{W}_{g}^{\top}\right) \tag{2}$$

Here, $\mathbf{E}_{bnd}$ denotes the matrix of boundary embeddings retrieved based on the lexicon indices $\mathbf{B}^{\text{lex}}$. $\mathbf{W}_{cat}$ and $\mathbf{W}_{g}$ represent learnable projection matrices, and $[\cdot\,;\,\cdot]$ signifies concatenation along the feature dimension.

Norm-based Adaptive Scaling. A generic Norm-based Adaptive Scaling function is proposed to mitigate feature dilution by regulating injection intensity. Given a source matrix $\mathbf{V}$, a target matrix $\mathbf{T}$, and hyperparameters $\Theta = \{\theta_b, \theta_a\}$, the gating coefficient $\mathbf{\Lambda}$ is defined as:

$$\mathbf{\Lambda} = \theta_b + \theta_a \cdot \tanh\left(\frac{\mathcal{O}(\mathbf{V})}{\mathcal{O}(\mathbf{T}) + \epsilon}\right) \tag{3}$$

where $\mathcal{O}(\cdot)$ denotes the row-wise L_2-norm operator and ϵ represents a smoothing term. Regarding structural injection, the same mechanism is applied using hyperparameters Θ^{str} to derive $\mathbf{\Lambda}^{\text{str}}$ from inputs $\mathbf{G}$ and $\mathbf{H}$. This results in the structurally enhanced representation:

$$\mathbf{H}^{\text{str}} = \mathbf{H} + \mathbf{\Lambda}^{\text{str}} \odot \mathbf{G} \tag{4}$$

where $\odot$ corresponds to the Hadamard product applied column-wise to the feature matrix.

3.3.3 Semantic Category Aggregation

Although Structural Prior Injection captures explicit boundary signals, it treats category labels as static identifiers. To address the fine-grained

semantics of hate speech categories (e.g., distinguishing Region bias from Racism), we introduce the Semantic Category Aggregation module. This component employs an attention mechanism to dynamically aggregate semantic features from a learnable category embedding space, guided directly by lexicon-induced span priors $(S^{\text{lex}}, P^{\text{lex}})$ produced by the Slang Lexicon Encoder.

Prior-Guided Category Attention. To incorporate linguistic knowledge, a token-level Attention Gate $\mathbf{\Pi}$ is constructed using the lexicon-induced span priors $(S^{\text{lex}}, P^{\text{lex}})$. For each span $s \in S^{\text{lex}}$ covering the interval $[f_s, e_s]$, let $\rho_{s,c}$ denote the entry in P^{lex} corresponding to the association strength between span s and category c. These span-level attributes are then projected onto the token sequence. For every token index i and category c, the gate entry $\Pi_{i,c}$ is calculated via the aggregation of weighted priors:

$$\Pi_{i,c} = \sum_{s \in S^{\text{lex}}} \mathbb{I}(f_s \leq i \leq e_s) \cdot \psi(i,s) \cdot \rho_{s,c} \quad (5)$$

where $\mathbb{I}(\cdot)$ is the indicator function. The term $\psi(i,s)$ reinforces structural boundaries by allocating a weight of 3.0 to single-token spans ($f_s = e_s = i$), 1.5 to span endpoints ($i \in \{f_s, e_s\}$), and 1.0 to non-boundary tokens.

Furthermore, tokens within overlapping span regions are upweighted to reflect boundary ambiguity (see Appendix A.2).

Attention scores undergo explicit modulation through the attention gate $\mathbf{\Pi}$. The encoder output $\mathbf{H}$ is used as the query, and a learnable category embedding matrix $\mathbf{E}_{\text{cat}}$ functions as both key and value. Learnable projection matrices $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ generate the raw attention scores A_{raw}. Finally, the gate is applied via element-wise multiplication before a standard renormalization step:

$$\mathbf{A_{raw}} = \frac{(\mathbf{H}\mathbf{W}_q^\top)(\mathbf{E}_{\text{cat}}\mathbf{W}_k^\top)^\top}{\sqrt{d_{\text{cat}}}} \quad (6)$$

$$\mathbf{A} = \text{RowNorm}(\text{softmax}(\mathbf{A_{raw}}) \odot \mathbf{\Pi}) \quad (7)$$

where RowNorm ensures row-wise normalization.

Consequently, the projected semantic representation matrix $\mathbf{U}$ is derived as:

$$\mathbf{U} = \text{MLP}(\mathbf{A}(\mathbf{E}_{\text{cat}}\mathbf{W}_v^\top)) \quad (8)$$

Finally, the gating coefficient $\mathbf{\Lambda}^{sem}$ is calculated (Eq. 3) using inputs $\mathbf{U}$ and $\mathbf{H}$ with parameters Θ^{sem}. The final representation is obtained through the fusion of structural and semantic features: $\mathbf{H}^{\text{final}} = \mathbf{H}^{\text{str}} + \mathbf{\Lambda}^{\text{sem}} \odot \mathbf{U}$.

3.4 Label-Disentangled Volumetric Tagging

Conventional grid tagging restricts each token pair to a single label. This constraint fails to capture co-occurring toxicity attributes within a single target–argument pair, which leads to label collisions. The Label-Disentangled Volumetric Tagging (LDVT) framework is proposed to address this limitation.

Tag-specific Representation. For each sub-task $t \in \{\text{ent}, \text{rel}, \text{hate}, \text{cat}\}$, a label channel is defined for every class in $\mathcal{Y}^t$, indexed by idx. The final token representation $\mathbf{h}_i^{\text{final}}$ is subsequently mapped to a label-specific channel representation:

$$\mathbf{z}_{i,idx}^{(t)} = \mathbf{W}_{idx}^{(t)}\mathbf{h}_i^{\text{final}} + \mathbf{b}_{idx}^{(t)}, \quad (9)$$

where $\mathbf{W}_{idx}^{(t)}$ and $\mathbf{b}_{idx}^{(t)}$ are learnable parameters.

Volumetric Scoring and Decoding. To incorporate relative position information, the position-encoded interaction term $R_{ij,idx}^{(t)}$ is calculated via Rotary Positional Embeddings (RoPE) (Su et al., 2024):

$$R_{ij,idx}^{(t)} = \text{RoPE}(\mathbf{z}_{i,idx}^{(t)})^\top \text{RoPE}(\mathbf{z}_{j,idx}^{(t)}) \quad (10)$$

For mutually exclusive tasks ($t \in \{\text{ent}, \text{rel}, \text{hate}\}$), the probability distribution $p_{ij,idx}^{(t)}$ is calculated for each token pair (i,j) by applying a softmax function over the label dimension:

$$p_{ij,idx}^{(t)} = \text{softmax}_{idx}\left(R_{ij,idx}^{(t)}\right) \quad (11)$$

Regarding the multi-label category task ($t = \text{cat}$), the prediction probability $p_{ij,idx}^{(t)}$ is derived through an element-wise sigmoid function:

$$p_{ij,idx}^{(t)} = \sigma\left(R_{ij,idx}^{(t)}\right) \quad (12)$$

3.5 Optimization Objective

The global optimization objective combines losses from all sub-tasks $t \in \{\text{ent}, \text{rel}, \text{hate}, \text{cat}\}$ through weighted summation:

$$\mathcal{L} = \sum_t \xi_t \mathcal{L}_t \quad (13)$$

where ξ_t represents the balancing coefficient. Categorical Cross-Entropy (CCE) defines the loss for mutually exclusive tasks $t \in \{\text{ent}, \text{rel}, \text{hate}\}$:

$$\mathcal{L}_t = -\frac{1}{n^2} \sum_{i,j} \sum_{idx \in \mathcal{Y}^t} \alpha_{idx}^{(t)} y_{ij,idx}^{*} \log p_{ij,idx}^{(t)} \quad (14)$$

For the multi-label classification task $t = \text{cat}$, Binary Cross-Entropy (BCE) is applied:

$$\mathcal{L}_t = \frac{1}{n^2} \sum_{i,j} \sum_{idx \in \mathcal{Y}^t} \alpha_{idx}^{(t)} \text{BCE}\left(p_{ij,idx}^{(t)}, y_{ij,idx}^{*}\right) \tag{15}$$

In these equations, $idx \in \mathcal{Y}^t$ identifies the label index for task t. The ground-truth label is denoted by $y_{ij,idx}^{*}$, and $\alpha_{idx}^{(t)}$ represents the label-specific weight (further details in Appendix A.1).

4 Experiments

4.1 Datasets

Internal Stratified Splitting. Evaluation is conducted on the STATE ToxiCN benchmark (Bai et al., 2025), which focuses on fine-grained span-level toxicity extraction. Although the original dataset provides fixed splits, direct hyperparameter optimization on the test set poses risks of data leakage and overfitting. To mitigate this, a robust internal validation set is derived from the training data ($N = 6,424$) through a Multi-Dimensional Stratified Splitting (MDSS) strategy. This algorithm stratifies samples along five orthogonal dimensions: hate category combinations, Hate Polarity balance, text length, quadruple density, and multi-label complexity.

Data partitioning yields an internal training set ($N = 5,101$) and a validation set ($N = 1,323$), maintaining an approximate 8:2 ratio (Table 2). The distribution of complex multi-label instances, such as *Racism* co-occurring with *Sexism*, shows high consistency between subsets (Diff $< 0.1\%$). Such alignment ensures that the validation set reflects the difficulty of the test set, allowing for reliable early stopping while preserving official test data integrity. Detailed stratification statistics are provided in Appendix A.3.

Evaluation Metrics. Following the STATE ToxiCN benchmark (Bai et al., 2025), **Hard-F1** is adopted as the primary metric. Unlike soft-match metrics that permit partial overlaps, Hard-F1 enforces a strict exact-match criterion. Specifically, a predicted quadruple is classified as a true positive only if the character-level boundaries of the Target and Argument spans perfectly align with the ground truth, alongside correct toxicity classification.

Table 2: Statistics of the stratified data split. The validation set closely matches the training distribution, including multi-label cases.

Metric	Train		Valid	
	Count	Ratio	Count	Ratio
Total Samples	5,101	79.4%	1,323	20.6%
Hate Polarity	3,118	61.1%	824	62.3%
Multi-label Samples	3,131	61.4%	827	62.5%

4.2 Implementation Details

The backbone encoder is instantiated with chinese-roberta-wwm-ext-large (Cui et al., 2020). Optimization was performed with AdamW at a learning rate of $2e{-}5$ and a batch size of 10. All experiments were conducted on an NVIDIA RTX 5090 GPU. An early stopping mechanism with a patience of 10 epochs was applied based on performance on the internal validation set. To ensure robustness, the experimental results are reported as the average of 5 runs. Table 5 lists the full hyperparameter settings.

4.3 Baselines

The framework is compared to established methods from STATE ToxiCN (Bai et al., 2025), and these baselines include both generative LLMs and discriminative sequence taggers. Generative evaluations follow two paradigms: (1) **open-source models** instruction-tuned on the training set, consisting of encoder-decoder (mT5-base (Xue et al., 2021)) and decoder-only architectures (Mistral-7B (Jiang et al., 2023), Qwen2.5-7B (Team, 2024b), LLaMA3-8B (Team, 2024a)), with safety-aligned versions (ShieldLM-14B (Zhang et al., 2024), ShieldGemma-9B (Zeng et al., 2024)) used to assess safety alignment transfer; and (2) **API-accessed foundation models** (GPT-4o (Hurst et al., 2024), Claude-3.5 Sonnet (Anthropic, 2024), Gemini-1.5-Pro (Gemini Team, 2024), DeepSeek-v3 (DeepSeek-AI, 2024), LLaMA3-70B (Team, 2024a), Qwen2.5-72B (Team, 2024b)) which are tested in a few-shot setting. Following the protocol in Bai et al. (2025), the "Basic Prompt and 2 Examples" template (one hateful, one non-hateful) is applied to produce structured extraction without any parameter updates.

4.4 Main Results

Table 3 compares SLAF with competitive fine-tuned baselines on the STATE ToxiCN test set. Fine-tuned models are emphasized as they generally perform better than in-context learning ap-

Model	Target Hard-F1	Argument Hard-F1	T-A Pair Hard-F1	T-A-H Tri. Hard-F1	Quad. Hard-F1
mT5-base	59.15	28.63	23.33	17.76	16.60
Mistral-7B	62.97	35.58	30.55	26.15	23.72
Qwen2.5-7B	63.96	35.42	30.63	26.51	23.70
ShieldLM-14B-Qwen	63.83	34.80	30.20	26.18	23.59
LLaMA3-8B	64.07	36.72	31.64	27.04	24.27
Ours	**69.98** (+5.91)	**43.49** (+6.77)	**39.77** (+8.13)	**34.96** (+7.92)	**30.09** (+5.82)

Table 3: Main results of the STATE ToxiCN. Performance comparison of various models across different levels of the annotated tasks, including Target, Argument, Target–Argument Pair (T-A Pair), Target–Argument–Hateful Triple (T-A-H Tri.), and Target–Argument–Hateful-Group Quadruple (Quad.), evaluated under the Hard-F1 metric. For Ours, numbers in parentheses denote absolute improvements over the strongest non-Ours baseline (LLaMA3-8B).

Model Variant	**Hard-F1 (%)**
SLAF (Full Model)	**30.09**
w/o Structural-Semantic Lexicon Fusion	21.36
⌊ w/o Structural Prior Injection	22.44
⌊ w/o Semantic Category Aggregation	27.19
w/o LDVT	23.14

Table 4: Ablation results on the STATE ToxiCN test set. "w/o LDVT" denotes replacing the multi-label volumetric decoding with a standard mutually exclusive classification scheme.

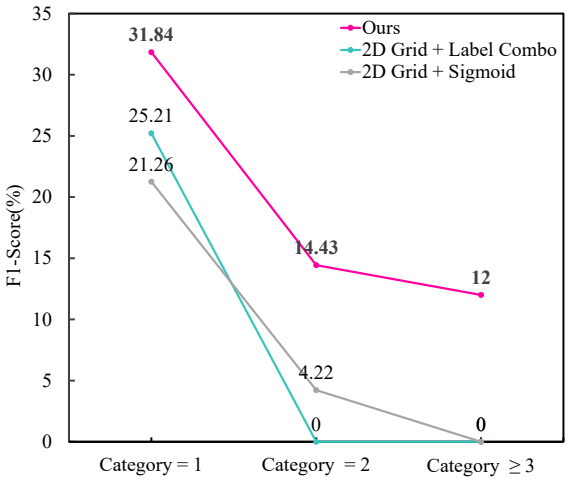

Figure 3: Performance breakdown by the number of toxicity categories per sample (Category Count).

proaches (e.g., API-based LLMs) in strict boundary extraction. Full results are provided in Appendix A.4. At the most stringent Quadruple level, SLAF achieves 30.09% Hard-F1, which is 5.82% higher than the strongest baseline (LLaMA3-8B).

Two trends are evident from these results. First, SLAF shows a significant Hard-F1 lead on the Target–Argument Pair sub-task, 8.13% above LLaMA3-8B. This result supports the LDVT scheme; since the model maps token interactions into separate subspaces, span boundary ambiguity is handled more reliably than in generative models. This advantage also persists as structural complexity increases, with gains of 7.92% on T–A–H Triples and 5.82% on Quadruples. Here, the adaptive adjustment of structural lexicon priors strengthens alignment between extracted spans and toxicity attributes (Hate and Group), which assists in addressing complex cases like implicit slang.

4.5 Ablation Study

We conduct an ablation study on the STATE ToxiCN test set to quantify the contribution of each SLAF component; Table 4 summarizes the results. Removing SSLF leads to the largest performance drop (8.73%). This shows that slang knowledge within the pre-trained encoder is vital for linguistic cloaking. Detailed analysis identifies Structural Prior Injection as the main factor. Without it, Hard-F1 drops to 22.44%. This component defines token boundaries clearly to help identify spans altered by slang during hard-match evaluation. Meanwhile, excluding Semantic Category Aggregation reduces performance by 2.90%. This suggests that attention-based semantic processing complements the structural backbone and helps distinguish similar toxicity categories.

The "w/o LDVT" variant examines label collisions by simplifying category prediction into a single-label task, similar to standard 2D grid tagging. Performance drops by 6.95%. This reflects the frequent occurrence of multi-attribute spans in hate speech detection, such as span pairs that are both Sexism and Racism. LDVT manages this problem specifically. It maps predictions into separate subspaces and applies multi-label decoding to prevent collisions and keep co-occurring hateful attributes.

4.6 Robustness to Collision and Complexity

Robustness under severe label-collision conditions is evaluated by stratifying the test set according to "Category Count" (Figure 3). Higher complexity leads to a sharp performance decline in two common grid tagging variants because of structural limitations. In particular, 2D Grid + Label Combo treats each label combination as a distinct atomic class but collapses to 0% F1 at Category = 2 owing to combinatorial sparsity. The 2D Grid + Sigmoid method uses Sigmoid activation for multi-class output yet fails on complex overlaps; its score reaches 0% at Category $\geq$3. SLAF shows better robustness; the model captures intersectional dependencies to maintain a 12% F1 score in the most challenging Category ≥ 3 scenario.

4.7 Case Study

Figure 4 displays two challenging cases from the model. In the first instance, the post targets women by associating them with foreigners ("They think foreign black people are rich"). This association creates an intersectional attack where the same target–argument pair carries both *Sexism* and *Racism*. Conventional 2D grids face problems assigning multiple attributes to a span interaction due to label collisions. In contrast, LDVT maps the interaction into a multi-dimensional prediction space to separate simultaneous toxicities. In the second case, the target "tank" serves as linguistic camouflage, specifically a derogatory slang term for overweight women. Standard encoders often map the term to its literal sense as a military vehicle. Once SSLF provides lexicon definitions, the model resolves the context and identifies the implicit Sexism signal.

id	post
1	她们以为外国黑人有钱 哈哈哈。 *They think foreign black people are rich, hahaha.*
2	上次遇到一个坦克，直接跟说自己这么漂亮怎么可能没谈过对象我就知道是什么成分了。 *Last time I met a tank, as soon as she directly said, "I'm so pretty, how could I possibly have never dated anyone," I knew what kind of stuff she was made of.*

Id	Type	Gold Label[T, A, H, G]	ours
1	Multi-label	['她们', '以为外国黑人有钱 ', 'hate', 'Sexism, Racism']	['她们', '以为外国黑人有钱 ', 'hate', 'Sexism, Racism']
2	Slang	['坦克', '什么成分', 'hate', 'Sexism']	['坦克', '什么成分', 'hate', 'Sexism']

Figure 4: Case Study: The primary target, argument in a post are colored differently. Also, to improve clarity, we color-code each element of the gold label quadruple in the case study table: target, argument, hateful, and group.

5 Conclusion

SLAF is a boundary-precise framework for Chinese hate speech detection in STATE ToxiCN, where precise span matching is necessary to prevent boundary errors. SSLF tackles semantic obfuscation through a combination of structural priors and dynamically aggregated semantic-category cues. This helps with resistance to slang and disguised toxic spans. For multi-category extraction, LDVT minimizes label interference through category-disentangled decoding. Consequently, the model extracts overlapping toxicity quadruples directly with little need for heuristic post-processing. Tests show that SLAF leads competitive baselines on STATE ToxiCN and provides a 5.82% absolute Hard-F1 gain over the strongest fine-tuned LLM. The results suggest that boundary-aware, structurally reliable architectures are vital for span-level toxicity extraction despite the rise of generative models.

Limitations

SLAF achieves high results for precise toxicity extraction, though several constraints exist. First, the method relies on external lexicon coverage. While structural prior injection reduces this need, the model might miss new slang terms absent from the gazetteer. This gap indicates that static resources often lag behind fast-changing online speech. Second, LDVT requires more parameters and memory than models with shared features. Since it assigns a separate feature subspace to each label index ($O(N \times d)$), the projection layer expands dimensionality based on the label set size. This structure prevents label collisions but might limit use on small devices if the label set is large. Third, hate detection via quadruple extraction is a recent task with few benchmarks. Only one Chinese dataset currently follows this format. This shortage of data makes evaluation across different domains difficult. Performance of the architecture beyond this specific benchmark is not yet fully verified.

Ethical considerations

This research adheres to the data usage policies of the STATE ToxiCN benchmark and the CC BY-NC 4.0 license. Dataset samples and selected examples

may involve offensive, vulgar, or hateful language. Such content is included solely for scientific investigation within the scope of toxicity and hate-speech detection and does not represent the personal views of the authors. All personally identifiable information that may appear in case studies or analyses is removed or anonymized to ensure user privacy is fully protected. For the present experiments, the dataset is only split as required by the study, and neither the dataset nor raw user-generated text is redistributed; released contributions remain limited to the model, training procedure, and evaluation results as permitted by the benchmark. Since exposure to toxic language is sensitive and potentially distressing, unnecessary reproduction of such text is avoided, and examples are provided only when required for technical analysis. Potential risks of bias and misuse are also considered: performance might differ across targets or social groups, and the system could be used in the creation of subtle attacks. Responsible use is therefore encouraged, including human oversight, safety testing, and post-deployment monitoring with clear reporting channels to minimize harm and prevent abuse.

References

Hyeseon Ahn, Youngwook Kim, Jungin Kim, and Yo-Sub Han. 2024. Sharedcon: Implicit hate speech detection using shared semantics. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 10444–10455. Association for Computational Linguistics.

Raza Ali, Umar Farooq, Muhammad Umair Arshad, Waseem Shahzad, and Mirza Omer Beg. 2022. Hate speech detection on twitter using transfer learning. *Comput. Speech Lang.*, 74:101365.

Badr AlKhamissi, Faisal Ladhak, Srini Iyer, Veselin Stoyanov, Zornitsa Kozareva, Xian Li, Pascale Fung, Lambert Mathias, Asli Celikyilmaz, and Mona T. Diab. 2022. Token: Task decomposition and knowledge infusion for few-shot hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2109–2120. Association for Computational Linguistics.

Anthropic. 2024. Claude 3.5 sonnet model card addendum. Technical report, Anthropic. Model card addendum (official report).

Leandro Araújo, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016*, pages 687–690. AAAI Press.

Zewen Bai, Liang Yang, Shengdi Yin, Junyu Lu, Jingjie Zeng, Haohao Zhu, Yuanyuan Sun, and Hongfei Lin. 2025. STATE ToxiCN: A benchmark for span-level target-aware toxicity extraction in Chinese hate speech detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10206–10219, Vienna, Austria. Association for Computational Linguistics.

Michał Bilewicz and Wiktor Soral. 2020. Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*.

Tommaso Caselli, Valerio Basile, Jelena Mitrovic, and Michael Granitzer. 2020. Hatebert: Retraining BERT for abusive language detection in english. *CoRR*, abs/2010.12472.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 657–668. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11580–11599. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Technical report, Google. Technical report (official PDF).

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *CoRR*, abs/2305.14450.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Elena Hanna, Simon Bressand, and 1 others. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023. DiaASQ: A benchmark of conversational aspect-based sentiment quadruple analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13449–13467, Toronto, Canada. Association for Computational Linguistics.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16235–16250. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14867–14875. AAAI Press.

Endang Wahyu Pamungkas, Dian Purworini, Diah Priyawati, and Rona Rizkhy Bunga Chasana. 2023. Exploring the impact of lexicon-based knowledge transfer for hate speech detection in indonesia code-mixed languages. In *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval, NLPIR 2023, Seoul, Republic of Korea, December 15-17, 2023*, pages 85–90. ACM.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Llama Team. 2024a. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Qwen Team. 2024b. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023. Instructuie: Multi-task instruction tuning for unified information extraction. *CoRR*, abs/2304.08085.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1572–1582. International Committee on Computational Linguistics.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online, November 16-20, 2020*, pages 2576–2585, Online. Association for Computational Linguistics.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024. ToxiCloakCN: Evaluating robustness of offensive language detection in Chinese with cloaking perturbations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6012–6025, Miami, Florida, USA. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Marcos Zampieri, Skye Morgan, Kai North, Tharindu Ranasinghe, Austin Simmmons, Paridhi Khandelwal, Sara Rosenthal, and Preslav Nakov. 2023. Target-based offensive language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 762–770. Association for Computational Linguistics.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. 2024. Shieldgemma: Generative AI content moderation based on gemma. *CoRR*, abs/2407.21772.

Zhexin Zhang, Yida Lu, Jingyuan Ma, Di Zhang, Rui Li, Pei Ke, Hao Sun, Lei Sha, Zhifang Sui, Hongning Wang, and Minlie Huang. 2024. ShieldLM: Empowering LLMs as aligned, customizable and explainable safety detectors. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10420–10438, Miami, Florida, USA. Association for Computational Linguistics.

A Appendix

A.1 Experiment Details

Table 5 lists the hyperparameters for the experiments and covers training configurations, loss weights, and decoding thresholds. The full model has approximately 0.36B parameters. In contrast, generative baselines such as LLaMA3-8B and ShieldLM-14B use 7B–14B parameters; this scale is over $20\times$ the size of this architecture. This disparity shows that the performance of SLAF stems from specific and unique architectural inductive biases rather than simple parameter scaling. Furthermore, the structured signals are prepared offline via a single preprocessing step to avoid repeated parsing during the training phase. The extra computational demand is small compared to the cost of fine-tuning or serving large language models. Therefore, the framework remains a highly suitable and efficient technical option for environments where hardware resources are limited.

A.2 Overlap handling in Π.

Tokens at overlapping lexicon span positions are designated as part of a shared overlap region, where additional upweighting is used to stress boundary ambiguity. In these specific instances, the position weight $\psi(i, s)$ is multiplied by a constant factor of 1.5. When multiple spans cover a single token, their resulting priors are integrated through a lightweight weighted fusion. This process is designed to prevent the over-amplification of individual spans while preserving multi-span evidence.

A.3 Detailed Data Splitting Strategy

A Multi-Dimensional Stratified Splitting (MDSS) strategy is used to ensure reliable internal evaluation and avoid tuning to the official test set. The original training dataset ($N = 6,424$) is partitioned into an internal training set ($N = 5,101$) and an internal validation set ($N = 1,323$). MDSS defines each sample through a feature tuple across five orthogonal dimensions. Data is split within these strata so that the validation set matches the training distribution across multiple granularities.

The first dimension, Hate Category Combination, covers all distinct toxicity categories in the quadruples of a sample, which are stored as a sorted set. This method distinguishes single-category samples from intersectional cases, such as posts with both Sexism and Racism. This approach maintains their proportions in the validation set. The second dimension is Hate Balance. This dimension uses the set of hate labels (Hate or Non-hate) to preserve the overall toxic/non-toxic ratio. The third dimension is Text Length Distribution. Posts are grouped by character count into "Short" (≤ 50), "Medium" (51–100), and "Long" (> 100). This categorization limits length-driven bias and provides an evaluation across different levels of verbosity.

Quadruple Density serves as the fourth dimension to balance extraction difficulty. Samples with two quadruples or fewer are labeled "Sparse", while those with more than two are labeled "Dense". This ensures information-dense posts are present in validation. The fifth dimension, Multi-label Complexity, indicates whether a sample includes multiple toxicity types. Based on these five dimensions, mutually exclusive strata are formed. A fixed-seed randomized split is performed within each stratum to allocate 20% of samples to the validation set. The final split aligns with the original data distribution, as shown in Table 6.

A.4 Comprehensive Results of All Models

A compact comparison between representative fine-tuned baselines and the proposed model is provided in the main text. For completeness, Table 7 reports results for all twelve models. Hard-F1 is evaluated at five levels: Target, Argument, Target-Argument Pair (T-A Pair), Target-Argument-Hate Triple (T-A-H Tri.), and Quadruple (Quad.). This protocol follows the primary setting. Fine-tuned models use the basic prompt; API-based LLMs use the basic prompt with two in-context examples. Values in parentheses for the proposed method denote absolute gains over the strongest baseline (LLaMA3-8B).

Item	Value
Training settings	
Backbone encoder	`hfl/chinese-roberta-wwm-ext-large`
Optimizer	AdamW
Dropout	0.1
Random seed	42
Batch size	10
Epochs	30
Learning rate	$2e{-5}$
Early stopping patience	10
Device	RTX5090
Parameter scale (Ours)	≈ 0.36B
$\Theta^{\text{str}} = (\theta_a^{\text{str}}, \theta_b^{\text{str}}, \epsilon)$	(4.5, 0.5, $1e{-6}$)
$\Theta^{\text{sem}} = (\theta_a^{\text{sem}}, \theta_b^{\text{sem}}, \epsilon)$	(1.0, 2.0, $1e{-6}$)
$\alpha^{(\text{ent})}$ (tgt, arg, others)	[3.0, 3.0, 1.0]
$\alpha^{(\text{rel})}$ (h2h, t2t, others)	[8.0, 5.0, 1.0]
$\alpha^{(\text{hate})}$ (hate, non-hate, others)	[3.0, 3.0, 1.0]
$\alpha^{(\text{cat})}$ (Sexism, Racism, Region, LGBTQ, Others, Non-hate)	[4.0, 4.0, 4.0, 4.0, 8.0, 1.0]

Table 5: Hyperparameters used in our experiments.

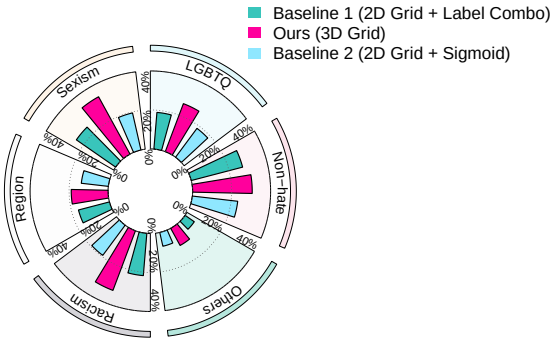

Figure 5: Hard-F1 comparison of decoding strategies across toxicity categories. Baseline 1 resolves label collisions via label-combination enumeration; Baseline 2 uses a coupled 2D grid with sigmoid activation; Ours decodes in a label-disentangled volumetric space.

A.5 Impact of Decoding Strategies

LDVT's improvement over standard grid tagging is analyzed through a comparison of the decoding strategy in SLAF with two common 2D grid variants (Figure 5).

Baseline 1 (2D Grid + Label Combo) adopts a "flattening" approach for label collisions. This method merges labels into composites such as Sexism and Racism. Its performance is lower in most categories. This result stems from sparsity in an expanded label space. Baseline 2 (2D Grid + Sigmoid) modifies 2D grid tagging for multi-label tasks. It replaces the Softmax layer with a Sigmoid activation. Predictions still come from a coupled 2D feature representation where label semantics remain mixed. SLAF shows higher scores than Baseline 2. Large differences appear in Sexism and LGBTQ. These findings show the need for volumetric disentanglement. A change in output activation alone is not sufficient. Separate label dimensions are necessary to separate subspaces. This configuration reduces cross-type interference for accurate fine-grained classification.

Stratification Dimension	**Internal Train** ($N = 5,101$)		**Internal Valid** ($N = 1,323$)	
	Count	Ratio	Count	Ratio
1. Hate Balance				
Hate	3,118	61.1%	824	62.3%
Non-hate	2,034	39.9%	518	39.1%
2. Multi-label Complexity				
Multi-label Targets	3,131	61.4%	827	62.5%
Single-label Targets	1,970	38.6%	496	37.5%
3. Text Length Distribution				
Short Text	4,118	80.7%	1,049	79.3%
Medium Text	700	13.7%	191	14.4%
Long Text	283	5.6%	83	6.3%
4. Quadruple Density				
Sparse Extraction (Few Quads)	4,966	97.4%	1,269	95.9%
Dense Extraction (Many Quads)	135	2.6%	54	4.1%
5. Hate Category Combinations				
Single Category Prevalence				
Sexism (Total)	1,227	24.1%	323	24.4%
Racism (Total)	982	19.3%	265	20.0%
Region (Total)	770	15.1%	206	15.6%
Intersectional Combinations (Top-3)				
Racism + Sexism	261	5.1%	68	5.1%
Region + Sexism	38	0.7%	11	0.8%
Racism + Others	34	0.7%	11	0.8%

Table 6: Statistics of the stratified data split. The training and validation sets exhibit closely matched distributions across five stratification dimensions.

Model	Target Hard-F1	Argument Hard-F1	T-A Pair Hard-F1	T-A-H Tri. Hard-F1	Quad. Hard-F1
LLM APIs (with Basic Prompt and 2 Examples)					
LLaMA3-70B	30.54	14.39	8.16	6.03	3.69
Qwen2.5-72B	40.94	21.10	15.66	12.48	8.74
Gemini-1.5-Pro	29.80	18.43	9.37	7.71	5.45
Claude-3.5-Sonnet	37.61	15.45	9.72	7.94	6.29
GPT-4o	46.85	22.64	17.21	13.21	9.00
DeepSeek-v3	48.16	22.79	18.68	14.95	11.48
Finetuned Models (with Basic Prompt)					
mT5-base	59.15	28.63	23.33	17.76	16.60
Mistral-7B	62.97	35.58	30.55	26.15	23.72
Qwen2.5-7B	63.96	35.42	30.63	26.51	23.70
ShieldLM-14B-Qwen	63.83	34.80	30.20	26.18	23.59
ShieldGemma-9B	63.40	34.40	29.99	25.64	23.49
LLaMA3-8B	<u>64.07</u>	<u>36.72</u>	<u>31.64</u>	<u>27.04</u>	<u>24.27</u>
Ours	**69.98** (+5.91)	**43.49** (+6.77)	**39.77** (+8.13)	**34.96** (+7.92)	**30.09** (+5.82)

Table 7: Comprehensive Hard-F1 results of all twelve models across five task levels (Target, Argument, T-A Pair, T-A-H Tri., and Quad.).