TAMA: Target-Aware Multilingual Abuse Detection by Cascaded Conditional Multi-Task Learning

Jiyan Liu[♠], **Youzheng Liu**[♠], **Taihang Wang** [♠], **Yimin Wang**[♡], **Ye Jiang**[♠*], **Diana Maynard**[♣]

[♠]School of Information Science and Technology, Qingdao University of Science and Technology,
[♡]School of Data Science, Qingdao University of Science and Technology,
[♣]Department of Computer Science, University of Sheffield

[*]**Correspondence:** ye.jiang@qust.edu.cn

Abstract

Protecting public figures from online abuse requires models that go beyond post-level classification to determine whether abuse is directed at a designated target, characterize the abuse intent, and extract textual evidence. We introduce **T**arget-**A**ware **M**ultilingual **A**buse (**TAMA**), a benchmark of 9,386 X (Twitter) posts aimed at public figures, with aligned supervision for (i) tri-class target detection, (ii) 12-way fine-grained abuse type classification, and (iii) phrase-level abusive span localization. To exploit the hierarchical coupling of these tasks, we propose **Cascaded-MTL**, a dependency-aware multi-task framework that conditions downstream predictions on upstream beliefs via three lightweight modules: Cross-Task Feature Fusion (CTF), Task-Adaptive Gating (TAG), and Label-Guided Span Detection (LGSD). Experiments across three multilingual encoders show that Cascaded-MTL consistently yields higher average F1 than single-task and standard multi-task training and delivers robust gains on type classification and span localization. The code and the dataset are released here: https://github.com/zgjiangtoby/CASCADED-MTL.

Disclaimer: The examples presented by this paper may be considered offensive or vulgar.

1 Introduction

Online abuse poses a critical challenge for platforms, undermining political debate, news consumption, and public engagement (Galpin and Vernon, 2024). Public figures, including politicians, journalists, and activists, are disproportionately targeted by coordinated harassment and reputational attacks that often spill over into offline harm (Sobieraj, 2020; Posetti et al., 2021). Such pressure induces a significant "chilling effect", leading to self-censorship and strategic withdrawal from platforms, particularly among underrepresented groups (Vogels, 2021).

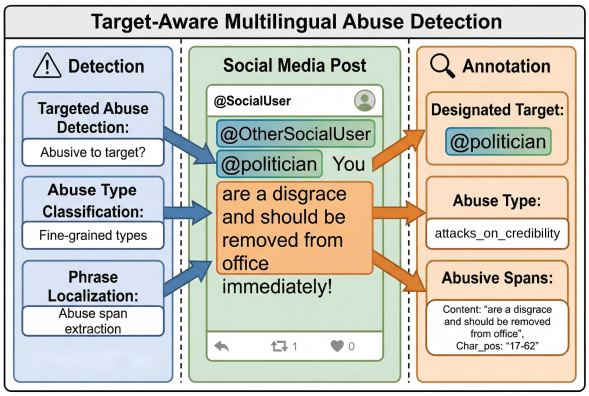


Figure 1: An overview of the TAMA benchmark. TAMA introduces three interdependent dimensions: (i) targeted abuse detection to identify whether abuse is directed at the designated public figure; (ii) abuse type classification into 12 fine-grained abuse types; and (iii) phrase localization with character-indexed abusive expressions.

Public figures act as agenda setters and information intermediaries; coordinated campaigns against them can distort deliberation and weaken democratic accountability, with visibility amplifying both the volume and organization of attacks (Steffek, 2010). Critically, most automated moderation tools treat abusiveness as a generic post-level property, overlooking *who* is attacked and *how*. This "one-size-fits-all" approach is ill-suited for protecting public figures because it fails to reliably separate legitimate policy criticism from personalised, target-directed harassment. For content explicitly directed at an individual (e.g., @-mentions), moderation requires a more nuanced paradigm: *determining not just if a post is abusive, but whether it specifically targets the designated person, while providing span-level evidence to justify the decision.*

While interest has grown in abuse targeting particular entities (Pachinger et al., 2024; Bai et al., 2025), existing safety models typically lack ex-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17842–17859
July 2-7, 2026

Benchmark	Source	#Posts	Span	Target	Multilingual
TBO (Zampieri et al., 2023)	X	4,673	✓	✓	
AustroTox (Pachinger et al., 2024)	DerStandard	4,562	✓	✓	
HATECAT-TR (Şeker et al., 2025)	X	2,981	✓	✓	
GameTox (Naseem et al., 2025)	Online game chat	53,000	✓		
HateBRXplain (Salles et al., 2025)	Instagram	7,000	✓		
SCCD (Yang et al., 2025)	Weibo	38,999		✓	
STATE ToxiCN (Bai et al., 2025)	Zhihu & Tieba	8,029	✓	✓	
OffensEval (Zampieri et al., 2019)	X	14,100		✓	
HASOC (Mandl et al., 2021)	X & Facebook	17,657		✓	✓
UA-HSD-2025 (Ahmad et al., 2025)	X	5,240			✓
AfriHate (Muhammad et al., 2025)	X	90,437		✓	✓
HateDay (Tonneau et al., 2025)	X	240,000		✓	✓
TAMA (Ours)	X	9,386	✓	✓	✓

Table 1: Comparison of abuse detection benchmarks based on number of posts (#Posts), span-level annotations (Span), inclusion of Target (Target), and languages (Multilingual).

plicit target identification and phrase-level localization. To fill this gap, we first construct the **T**arget-**A**ware **M**ultilingual **A**buse (**TAMA**) benchmark; an overview is shown in Figure 1.

Unlike previous datasets, TAMA is directed at prominent public figures and provides aligned supervision for: (i) tri-class target detection (non-abusive, abusive towards target, or abusive towards others); (ii) 12-way fine-grained abuse types; and (iii) phrase-level span marking the abusive expressions. Table 1 shows the differences between TAMA and previous abuse detection benchmarks.

We also introduce a Cascaded Conditional Multi-Task Learning (**Cascaded-MTL**) framework for jointly detecting multilingual, targeted abusive language against public figures. The framework uses a shared encoder with three heads for targeted detection, fine-grained abuse type classification, and phrase-level abusive span localization, and regulates inter-task information flow via complementary conditional strategies.

First, Cross-Task Feature Fusion (CTF) injects target detection decision evidence into the fine-grained abuse type classifier via logit-to-feature fusion. Second, Task-Adaptive Gating (TAG) enables differentiable information flow from target detection to abuse type classification through probability-based representation gating. Third, Label-Guided Span Detection (LGSD) conditions abuse span localization on the abuse type distribution to perform type-aware span extraction. This design yields efficient representation sharing, span-level interpretability, and robust performance across languages.

Our main contributions are:

- We release **TAMA**, a target-aware multilingual abuse benchmark for public figures with 9,386 posts and three aligned supervision signals per instance: tri-class target detection; 12-way fine-grained abuse types; and phrase-level abusive span.
- We propose **Cascaded-MTL**, a dependency-aware multi-task framework that follows the hierarchy *target* $\rightarrow$ *type* $\rightarrow$ *span* via three lightweight conditioning modules, yielding efficient representation sharing, span-level interpretability, and robust performances.
- We benchmark different encoder backbones and large language models (LLMs), and run targeted ablations over **CTF/TAG/LGSD** to isolate how each dependency link contributes to different task performance, establishing Cascaded-MTL as a strong, practical baseline for target-aware multilingual abuse moderation.

2 Related Work

The evolution of automated abuse detection has moved from coarse-grained classification toward increasingly granular taxonomies that account for target identity and linguistic evidence.

2.1 Shared Tasks and Target Taxonomies

Early benchmarks focused on binary or multi-class classification of hate and offense (Waseem and Hovy, 2016; Davidson et al., 2017). This was later formalised by major shared tasks such as OffensEval (Zampieri et al., 2019) and HatEval

(Basile et al., 2019), which established a hierarchical standard where models first detect offensive posts and then categorise the target (e.g., *individual* vs. *group*). Subsequent tasks like HASOC (Mandl et al., 2021) expanded these taxonomies to multilingual contexts, yet they largely maintain this categorical distinction.

Recently, resources with explicit target annotations and offensive span labels have started to emerge. TBO (Zampieri et al., 2023) introduced an English Twitter dataset with post-level harmfulness labels and token-level annotations of the TARGET(the person or group being targeted) and ARGUMENT(the span containing the offensive expression), moving beyond purely post-level offense classification. Tillmann et al. (2023) subsequently built on this line and reported a German TBO annotation set constructed from offensive German tweets.

While these benchmarks were instrumental in advancing the field, they treat "individual" as an anonymous category. In contrast to these category-based benchmarks, and extending the more recent target- and span-aware paradigm exemplified by TBO and German TBO, TAMA introduces explicit entity targeting, requiring models to attribute abusive content to a specific named public figure specified *a priori*. This reformulates abuse detection from broad category recognition to precise, target-aware detection.

Our focus on public figures builds upon early entity-centric studies, most notably Gorrell et al. (2018), who pioneered the use of **explicit target detection** through gazetteers and linguistic rules. However, this approach prioritises precision at the expense of recall, struggling with the high lexical variance of social media abuse. Our multilingual transformer backbone allows our model to capture subtle, non-keyword-based harassment, while maintaining the same commitment to explicit, identity-aware monitoring.

2.2 The Group-Centric Mismatch

Recent research has refined group-level detection to include specific ethnic and religious minorities in diverse sociocultural contexts (Şeker et al., 2025; Muhammad et al., 2025). However, a critical "mismatch" remains: Tonneau et al. (2025) demonstrate that models trained on fixed "academically popular" groups fail to generalise to real-world environments where victims vary wildly by region and time. This suggests that pre-defined taxonomies are inherently limited when it comes to protecting individuals from minority demographics.

Our work addresses this by moving beyond fixed categorical targets. Similar to the individual-focused analysis of cyberbullying (Yang et al., 2025), we center the person as the unit of analysis. However, by enabling explicit target specification, TAMA creates a system resilient to the "target mismatch" by adapting to the specific identity of high-risk public figures, such as journalists or activists. This granular focus is vital because, as already noted, the lack of explicit target modelling often prevents automated systems from distinguishing between vigorous public debate and the coordinated, identity-based harassment used to silence journalists and activists. By centering the specific entity, we provide the technical mechanism necessary to defend the "agenda setters" of democratic discourse from the chilling effects of personalised abuse.

2.3 Fine-grained Abuse and Explainability

As abuse becomes more sophisticated, interest has grown in identifying specific abusive behaviors and providing supporting evidence (Pachinger et al., 2024; Bai et al., 2025). While recent works add explanatory rationales (Salles et al., 2025), they often treat classification and span extraction as independent tasks. TAMA differs by providing a 12-way fine-grained taxonomy of harm, capturing nuances like reputational attacks and threats often conflated in coarse schemas, and aligning these with phrase-level abusive spans. Our Cascaded-MTL framework shows further innovation by conditioning span extraction on the identified abuse type, ensuring the highlighted evidence is semantically consistent with the model's decision.

3 Dataset

This section describes the data sources and filtering procedures, details the annotation methodology and quality control measures, and presents statistics for the final benchmark.

3.1 Task definition

We study multilingual targeted abusive language detection against public figures on X. Each instance corresponds to a post and a designated public figure, represented by their official post handle. Although posts may mention multiple accounts, we focus on whether abuse is directed specifically at the pre-

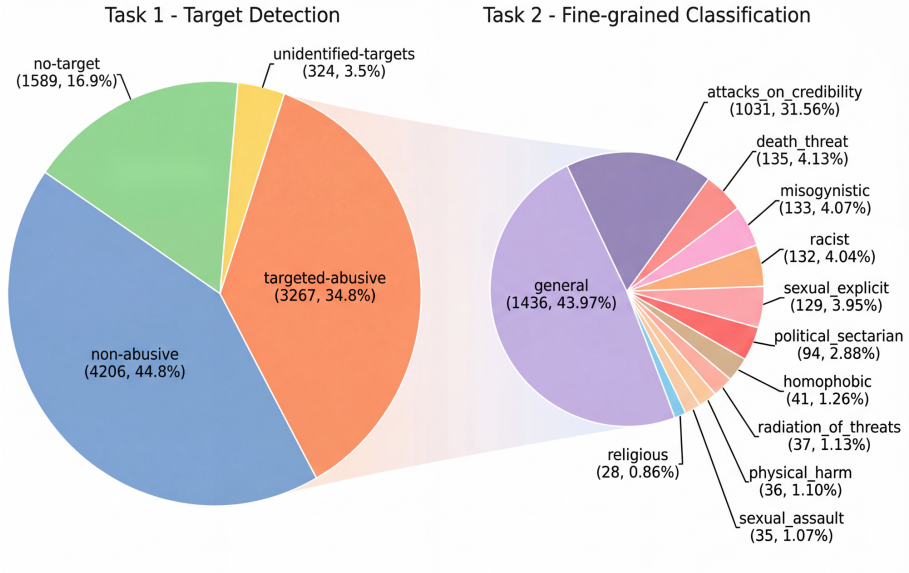


Figure 2: The overall data statistics of TAMA.

specified anchor target. This setup gives rise to three related tasks:

Task 1 (T1): Targeted abuse detection. For each post-target pair, we assign a tri-class label from the set *non-abusive*, *targeted-abusive*, or *unidentified-targets*, distinguishing between non-abusive content, abuse directed at the designated figure, and abuse aimed at other entities. Note that we remove *no-target* posts in the train phase as they do not explicitly point at any designated figures. Detailed category definitions of T1 labels are provided in Appendix D.

Task 2 (T2): Fine-grained abuse type classification. For *targeted-abusive* posts, we predict a fine-grained abuse type from the 12 categories defined by Posetti et al. (2023): *death_threat*, *sexual_assault*, *sexual_explicit*, *physical_harm*, *radiation_of_threats*, *attacks_on_credibility*, *misogynistic*, *homophobic*, *religious*, *political_sectarian*, *racist*, *general*. Detailed category definitions of T2 labels are provided in Appendix E.

Task 3 (T3): Phrase-level abusive span localization. We identify spans of specific abusive expressions using character offsets in the post. Spans are provided only for *targeted-abusive* posts.

3.2 Data source and filtering

We focus on five public figures from politics and journalism: Theresa May, Liz Truss, Maria Ressa, Yulia Navalnaya, and Vladimir Soloviev. These targets are based in multiple countries (UK, the Philippines, and Russia), receive messages spanning multiple languages (including English, Russian, Filipino, Spanish, and Italian), and are known to attract sustained online harassment.

We retrieve 218,259 original posts (excluding replies and quotes) associated with these accounts. To build a balanced candidate pool, we apply IBM's Granite Guardian content-safety models (Padhi et al., 2025) in zero-shot mode to obtain coarse binary predictions (i.e. whether a post is abusive or not), then sample approximately 2,000 posts per public figure (balanced between predicted abusive and non-abusive), yielding 10,000 candidates. Zero-shot predictions from LLM serve only as prescreening signals and are never used as ground truth.

Before annotation, we perform text cleaning: removing deleted or unavailable posts, filtering exact duplicates and extremely short content, and retaining posts suitable for subsequent annotation with respect to the five selected public figures.

3.3 Annotation

Annotators follow detailed guidelines to determine whether a post is abusive, whether abuse targets the designated figure, what type of abuse is expressed, and which phrases contain the abuse. To mitigate bias, we assemble a diverse team of four voluntary annotators with varied backgrounds in gender, age, ethnicity, and region. We employ regular cross-validation and expert arbitration to ensure consistency.

We follow a multi-stage protocol to stabilise guidelines before full-scale annotation:

Pilots. In the first calibration round, we sampled

2,500 posts (500 per public figure) for dual annotation. We achieved an average Cohen's kappa of 0.2751 across the five public figures and inspected disagreements to refine the guidelines. A second calibration round confirmed substantial improvement, with agreement scores rising to 0.5480 across all targets.

T1&T2 annotation. Using finalized guidelines, we annotate the complete candidate pool. Each post receives independent dual annotation for T1 and T2 labels. Disagreements are resolved through discussion and expert arbitration. The final average Cohen's kappa score is 0.7096, indicating substantial agreement.

T3 annotation. We adopt a human-in-the-loop approach for efficiency: we prompt GPT-5 to suggest candidate spans based on text, target, and fine-grained label. Annotators verify and edit suggestions, with only human-confirmed spans retained in the final dataset.

Inter-annotator agreement between targets and rounds is shown in Appendix B, which also discusses the main sources of disagreement and how iterative guideline refinement resolved recurring ambiguities. After removing posts that violate guidelines (unavailable content, ambiguous cases), we obtain 9,386 fully annotated posts. We perform additional quality checks using GPT-5 API to flag potential errors, manually reviewing and correcting several hundred instances to produce the final dataset.

3.4 Data description

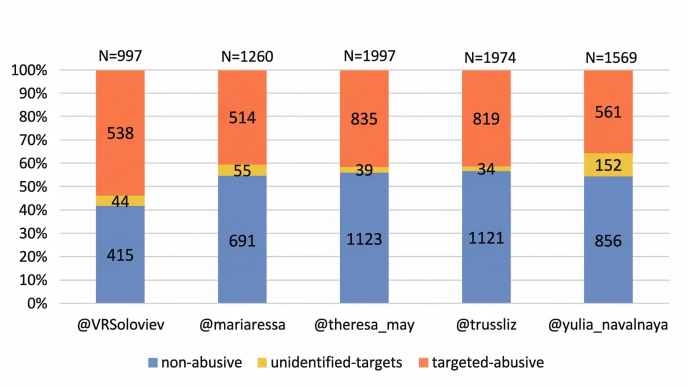
The TAMA dataset comprises 9,386 social media posts annotated with a multi-level hierarchical labeling scheme supporting three complementary tasks. As illustrated in Figure 2, T1 focuses on target detection, while T2 and T3 provide different annotation perspectives on the targeted-abusive content identified in T1, with T3 serving as a complementary annotation to T2.

The hierarchical structure ensures that T2 and T3 annotations both correspond to the same set of targeted-abusive instances from T1, maintaining data consistency across all three tasks. Each of the 3,267 targeted-abusive samples receives annotations from both T2 (abuse type classification) and T3 (complementary annotations), enabling multi-dimensional analysis of abusive content.

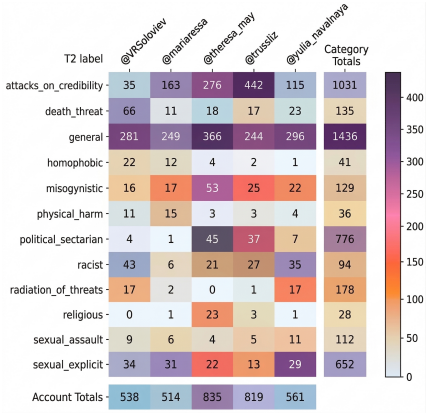
The T2 distribution exhibits notable class imbalance, with two dominant categories (general abuse and attacks on credibility) accounting for 75.5% of all targeted-abusive content. Conversely, specialized abuse types such as religious discrimination, sexual assault, and physical harm represent less than 2% each, reflecting their relatively lower prevalence in the dataset. This imbalance mirrors real-world patterns in online abuse, where generic insults and credibility attacks are more common than specific threat categories (Vogels, 2021).

Figure 3 provides a target-wise view of the hierarchical label composition in TAMA, showing the distributions of T1 and T2 labels across the five public figures. In Figure 3a, the T1 distribution shows a broadly consistent pattern across targets: *non-abusive* and *targeted-abusive* posts together account for most instances in every subset, while *unidentified-targets* remains limited. This suggests that the dataset mainly consists of either clearly non-abusive content or abuse that can be assessed with respect to the designated anchor target, rather than ambiguous cases or posts involving other entities. At the same time, the figure also reveals target-specific differences. For most targets, *non-abusive* posts are the largest category, with *targeted-abusive* posts forming a substantial secondary proportion. By contrast, one subset shows a stronger concentration of *targeted-abusive* content, while another has a more visible share of *unidentified-targets* cases. These variations indicate that, although TAMA maintains a stable first-stage label structure, abusive discourse is not distributed uniformly across public figures, which is consistent with the heterogeneous communicative environments in which they are discussed.

In Figure 3b, the T2 distribution within *targeted-abusive* posts reveals a clear long-tailed pattern. Across targets, *general* abuse and *attacks_on_credibility* are the most frequent labels, indicating that direct insults and credibility-undermining language are the dominant forms of target-directed abuse in the dataset. In contrast, categories such as *religious*, *physical_harm*, *sexual_assault*, *homophobic*, and *radiation_of_threats* remain rare and are often supported by only a few examples in specific subsets. This uneven distribution poses challenges for fine-grained modeling and evaluation, as some minority classes rely on limited, target-specific contexts rather than broadly representative evidence. Given that TAMA covers public figures from different political, linguistic, and sociocultural settings, variation in the prevalence and form of abuse is itself a realistic feature of naturally occurring online discourse. In this



(a) T1 label distribution per target. N denotes the total number of posts per target.



(b) T2 label distribution per target within targeted-abusive.

Figure 3: T1 and T2 label distributions across the five public figures in TAMA.

sense, the distribution shown in panel b not only highlights the difficulty of T2 classification, but also reflects the validity of the benchmark.

4 Cascaded Conditional MTL Framework

The overall Cascaded-MTL framework is shown in Figure 4.

4.1 Problem Formulation and Notation

For each example we observe a post x and its designated public figure (anchor target) a (e.g., an official handle). We construct an input sequence $s = \text{[TGT]}\ a\ \text{[/TGT]}\ x$, where `[TGT]` and `[/TGT]` are special tokens marking the target span.

T1 assigns a label $y^{\text{tgt}} \in \mathcal{Y}^{\text{tgt}}$, where $\mathcal{Y}^{\text{tgt}}$ are labels in T1. For posts with $y^{\text{tgt}} = \textit{targeted-abusive}$, T2 predicts a type label $y^{\text{type}} \in \mathcal{Y}^{\text{type}}$, where $\mathcal{Y}^{\text{type}}$ are labels in T2, and T3 predicts token-level span labels $y^{\text{loc}} \in \{0,1\}^L$ for a tokenized sequence of length L. By construction, y^{type} and y^{loc} are undefined for non-abusive or unidentified-targets posts.

We denote by d the hidden dimensionality of the encoder and by d_c the dimensionality of the projected T1 summary vector used in CTF. We use $\odot$ for element-wise multiplication and $\oplus$ for vector concatenation. Standard multi-class cross-entropy and token-wise binary cross-entropy losses are denoted by $\text{CE}(\cdot)$ and $\text{BCE}(\cdot)$ respectively.

4.2 Base Multi-Task Architecture

Shared encoder. We encode s with a multilingual Transformer encoder f_θ, obtaining token representations and a pooled sentence vector:

$$\begin{aligned} H &= f_\theta(s) \in \mathbb{R}^{L\times d}, \\ h_{\text{cls}} &= \text{Pool}(H) \in \mathbb{R}^d, \end{aligned} \tag{1}$$

where $H = [h_1, \ldots, h_L]$ and $\text{Pool}(\cdot)$ is mean pooling of the sentence representation. We truncate or pad inputs to $L = 128$ tokens and maintain offset mappings between tokens and characters to supervise spans.

Target detection head (T1). T1 is a standard 3-way classifier over $\mathcal{Y}^{\text{tgt}}$:

$$\begin{aligned} z^{\text{tgt}} &= W^{\text{tgt}} h_{\text{cls}} + b^{\text{tgt}}, \\ p^{\text{tgt}} &= \text{softmax}(z^{\text{tgt}}), \end{aligned} \tag{2}$$

with parameters $(W^{\text{tgt}}, b^{\text{tgt}})$. We optimize the cross-entropy loss

$$\mathcal{L}_{\text{tgt}} = \text{CE}\big(p^{\text{tgt}}, y^{\text{tgt}}\big) \tag{3}$$

on all instances.

Fine-grained type head (T2). T2 predicts a distribution over 12 abuse types for posts that are *targeted-abusive* according to T1. Let h^{type} denote the input representation to the T2 classifier. In the base model (without conditioning), $h^{\text{type}} = h_{\text{cls}}$. Conditional variants replace h^{type} as described in Section 4.3. We compute

$$\begin{aligned} z^{\text{type}} &= W^{\text{type}} h^{\text{type}} + b^{\text{type}}, \\ p^{\text{type}} &= \text{softmax}(z^{\text{type}}), \end{aligned} \tag{4}$$

and train with a class-weighted cross-entropy loss $\mathcal{L}_{\text{type}}$, where type-specific weights mitigate label imbalance. Only examples with $y^{\text{tgt}} = \textit{targeted-abusive}$ contribute to $\mathcal{L}_{\text{type}}$.

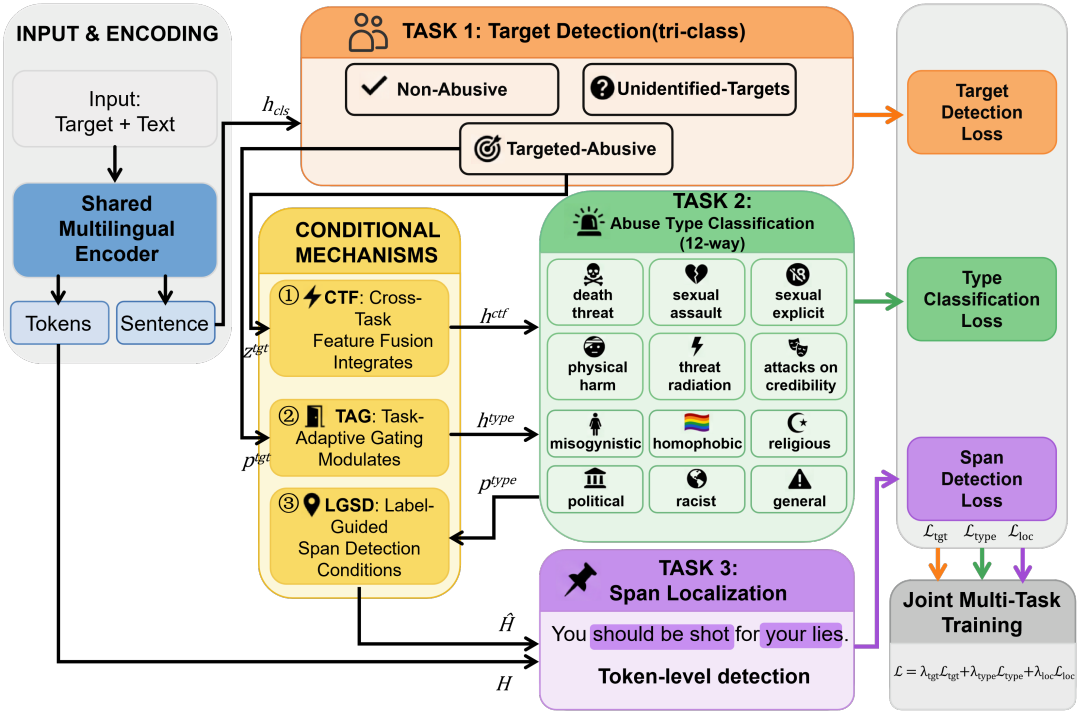


Figure 4: The overall framework of Cascaded-MTL.

Span localization head (T3). T3 is a token-level binary classifier. Given token representations $\hat{H} \in \mathbb{R}^{L\times d}$. In the base model (without conditioning), $\hat{H} = H$. Conditional variants replace $\hat{H}$ as described in Section 4.3. We compute a score for each token:

$$S = \hat{H}\, w^{\text{loc}} + b^{\text{loc}} \in \mathbb{R}^{L},$$
$$p^{\text{loc}} = \sigma(S) \in \mathbb{R}^{L}, \tag{5}$$

where $w^{\text{loc}} \in \mathbb{R}^{d}$ and $b^{\text{loc}} \in \mathbb{R}$, and σ is the element-wise sigmoid. The span loss is a token-wise binary cross-entropy between p^{loc} and y^{loc}, computed over all posts, with padded positions masked out. For posts with $y^{\text{tgt}} \neq$ *targeted-abusive*, we treat y^{loc} as an empty span (all-zero token labels) to provide additional negative supervision and reduce spurious span predictions.

4.3 Conditional Mechanisms for Task Interaction

The three tasks in TAMA describe the same underlying phenomenon from complementary views: T1 establishes whether abuse is on-target, T2 characterizes the semantic intent of targeted abuse, and T3 identifies the textual evidence that supports these judgments.

Naive multi-task learning shares parameters but leaves the interaction implicit, which can lead to gradient interference (e.g., global abusiveness cues dominating rare type cues) and brittle token-level evidence. We therefore introduce three lightweight conditioning operators that explicitly shape representation learning.

4.3.1 Cross-Task Feature Fusion (CTF)

CTF introduces an explicit cross-task channel by compressing the upstream decision evidence into a compact vector and fusing it with the shared sentence representation:

$$c^{\text{tgt}} = W_{\text{proj}} z^{\text{tgt}} \in \mathbb{R}^{d_c}, \tag{6}$$

where $W_{\text{proj}} \in \mathbb{R}^{d_c\times 3}$ and d_c is the dimensionality of the projected T1 summary vector. The fused representation is

$$h^{\text{ctf}} = h_{\text{cls}} \oplus c^{\text{tgt}} \in \mathbb{R}^{d+d_c}. \tag{7}$$

4.3.2 Task-Adaptive Gating (TAG)

TAG uses T1 probabilities as a *soft control signal* to adapt the contribution of h_{cls}:

$$g = \sigma\big(W_{\text{gate}}\, p^{\text{tgt}}\big) \in \mathbb{R}^{d}, \tag{8}$$

where $W_{\text{gate}} \in \mathbb{R}^{d\times 3}$. We obtain a gated sentence representation:

$$\tilde{h}_{\text{cls}} = h_{\text{cls}} \odot g \in \mathbb{R}^{d}. \tag{9}$$

The final T2 input is then

$$h^{\text{type}} = \tilde{h}_{\text{cls}} \in \mathbb{R}^d. \tag{10}$$

4.3.3 Label-Guided Span Detection (LGSD)

Different abuse types rely on distinct lexical and syntactic cues (e.g., explicit threat verbs vs. identity slurs). LGSD lets T2 guide T3 by conditioning token-level span predictions on the inferred abuse-type distribution.

Given $p^{\text{type}} = \text{softmax}(z^{\text{type}})$, we compute a type-aware vector:

$$u = \tanh\big(W^{\text{mod}}\, p^{\text{type}}\big) \in \mathbb{R}^d, \tag{11}$$

with $W^{\text{mod}} \in \mathbb{R}^{d\times 12}$. We then broadcast u across tokens and obtain

$$\hat{H} = H \odot u \in \mathbb{R}^{L\times d}, \tag{12}$$

which is fed into the span head described above. For example, if p^{type} places high probs on *death_threat*, LGSD encourages the span classifier to focus on tokens associated with threats and violent language, while down-weighting generic profanity that might be less diagnostic for this type. Because LGSD uses the full distribution rather than an argmax label, it remains robust when T2 is uncertain or multiple abuse types are partially activated.

4.4 Training Objective and Inference

Joint objective. The overall loss for a mini-batch is a weighted sum of the three task-specific losses:

$$\mathcal{L} = \lambda_{\text{tgt}}\,\mathcal{L}_{\text{tgt}} + \lambda_{\text{type}}\,\mathcal{L}_{\text{type}} + \lambda_{\text{loc}}\,\mathcal{L}_{\text{loc}}, \tag{13}$$

where $\lambda_{\text{tgt}}, \lambda_{\text{type}}, \lambda_{\text{loc}} \geq 0$ control the relative importance of each task. Unless otherwise stated, we set $\lambda_{\text{tgt}} = 1.2$ and $\lambda_{\text{type}} = \lambda_{\text{loc}} = 1.0$, chosen on the development set to balance T1 robustness with downstream type and span accuracy. All parameters of the encoder and heads, including CTF, TAG, and LGSD modules, are optimized jointly.

Inference. At inference time, the model operates in a cascaded manner:

1. Run the encoder and T1 head to obtain p^{tgt} and predict $\hat{y}^{\text{tgt}} = \arg\max p^{\text{tgt}}$.
2. If $\hat{y}^{\text{tgt}} \neq$ *targeted-abusive*, we output *non-abusive* or *unidentified-targets* and skip T2/T3.
3. If $\hat{y}^{\text{tgt}} =$ *targeted-abusive*, we compute T2 and T3 predictions using the conditioned representations: $\hat{y}^{\text{type}} = \arg\max p^{\text{type}}$ and span scores p^{loc}. Spans are extracted by thresholding p^{loc} and merging contiguous positive tokens into character-level phrases.

5 Results

Table 2 reports the overall performance of Cascaded-MTL across three multilingual encoders, including both single-task and multi-task baselines as well as conditional variants. Our findings suggest that Cascaded-MTL yields consistent gains in overall task performance across backbones, as reflected by `Avg_F1`.

On `multilingual-e5-base` backbone, multi-task learning already improves substantially over independent fine-tuning: `MTL-Base` increases `Avg_F1` from 63.00 to 65.98. Incorporating conditional modules further strengthens the joint solution, and the full configuration `+CTF+TAG+LGSD` achieves the best overall score (`Avg_F1` $= 68.25$), with strong gains on type classification and span localization.

The same trend holds when switching to other encoders. On `mDeBERTa-v3-base`, `MTL-Base` slightly improves over `Single_Task` (`Avg_F1`: $62.57 \rightarrow 62.89$), while Cascaded-MTL variants provide larger improvements, reaching 65.07 with `+CTF+TAG+LGSD`. On `XLM-RoBERTa-base`, Cascaded-MTL again improves the overall average from 60.00 (`Single_Task`) and 62.84 (`MTL-Base`) to 65.09 (`+CTF+TAG+LGSD`). These consistent improvements across heterogeneous pretraining objectives and model families support the model-agnostic nature of Cascaded-MTL.

6 Analysis

Cascaded-MTL is tailored to TAMA's hierarchical supervision, where target awareness (T1), abuse intent (T2), and textual evidence (T3) describe the same phenomenon at different granularities.

Cascaded-MTL consistently improves the joint metric across backbones, indicating a portable inductive bias rather than encoder-specific gains. In Table 2, the full configuration `+CTF+TAG+LGSD` yields the best `Avg_F1` on all three encoders, improving over `MTL-Base` by $+2.27$ (`multilingual-e5-base`: $65.98 \rightarrow 68.25$), $+2.18$ (`mDeBERTa-v3-base`: $62.89 \rightarrow 65.07$), and $+2.25$ (`XLM-RoBERTa-base`: $62.84 \rightarrow 65.09$). The

Method	T1_Acc	T1_F1	T2_F1	T3_F1	Avg_F1
multilingual-e5-base					
Single_Task	83.29	69.72	48.74	70.55	63.00
MTL-Base	85.35	71.90	53.86	72.19	65.98
+ CTF	85.35	73.46	51.52	75.11	66.70
+ CTF + TAG	84.45	73.15	50.21	78.10	<u>67.15</u>
+ CTF + TAG + LGSD	83.80	70.90	55.88	77.97	**68.25**
mDeBERTa-v3-base					
Single_Task	85.08	70.19	44.55	72.96	62.57
MTL-Base	83.16	69.67	48.08	70.93	62.89
+ CTF	84.06	70.44	45.15	76.91	64.17
+ CTF + TAG	84.32	70.39	46.63	76.78	<u>64.60</u>
+ CTF + TAG + LGSD	83.93	72.42	46.40	76.40	**65.07**
XLM-RoBERTa-base					
Single_Task	81.49	70.55	39.52	69.92	60.00
MTL-Base	81.88	68.85	49.06	70.62	62.84
+ CTF	84.19	70.45	45.16	75.98	63.86
+ CTF + TAG	82.26	68.39	47.24	76.95	<u>64.19</u>
+ CTF + TAG + LGSD	83.55	68.59	50.44	76.24	**65.09**

Table 2: Performance comparison (%) and ablation between different backbones with Cascaded-MTL. 'Single_Task' denotes the model is fine-tuned on each task independently without task interaction. 'MTL-Base' is the Cascaded-MTL framework without any conditional mechanisms. `Avg_F1` is the mean of `T1_F1`, `T2_F1`, and `T3_F1`. **Bold** and <u>underline</u> denote the best and the second best average F1, respectively.

per-task changes vary by backbone, but the overall improvement is stable, suggesting the method mainly reshapes how the three supervision signals cooperate.

The largest and most reliable gains appear on span localization, which is the most structurally constrained objective in TAMA. Across backbones, conditioning substantially increases `T3_F1` compared to `MTL-Base` (e.g., $72.19 \rightarrow 77.97$, $70.93 \rightarrow 76.40$, $70.62 \rightarrow 76.24$). This pattern is consistent with T3 being sensitive to higher-level semantics: without explicit coordination, token tagging can drift toward generic negativity, while cascading encourages spans to align with an on-target hypothesis and a coherent type interpretation.

Relative to standard multi-task learning, Cascaded-MTL makes cross-task interaction explicit, which stabilizes learning under sparsity and imbalance. `MTL-Base` already improves over `Single_Task` on `Avg_F1`, but the gain can be modest (notably on `mDeBERTa-v3-base`: $62.57 \rightarrow 62.89$), reflecting that implicit parameter sharing does not always translate into better coordination between dense T1 supervision, long-tailed T2 labels, and token-level T3 evidence. By conditioning downstream representations on model beliefs (p^{tgt} and p^{type}), Cascaded-MTL reduces over-confident downstream activation when upstream signals are ambiguous and strengthens downstream signals when upstream confidence is high, improving the overall operating point.

Ablations show complementary contributions: each link improves a different part of the triple, and the full cascade yields the best trade-off. On `multilingual-e5-base`, `+CTF` improves `T1_F1` and `T3_F1`, `+CTF+TAG` yields the strongest `T3_F1` (78.10), and adding `LGSD` is where `T2_F1` increases to its best value (55.88) while maintaining high `T3_F1`, producing the best `Avg_F1`. Similar "best overall from composition" behavior holds on the other backbones, supporting the view that Cascaded-MTL improves performance by enforcing cross-task consistency rather than optimizing tasks in isolation.

7 Conclusion

We introduced **TAMA**, a target-aware multilingual abuse benchmark for public figures with aligned supervision for target detection, abuse type classification, and phrase-level span localization. TAMA moves beyond post-level abuse detection and provides a more realistic benchmark for studying who is targeted, what type of abuse is expressed, and which textual evidence supports the decision. We also proposed **Cascaded-MTL**, a lightweight conditional multi-task framework that exploits these task dependencies, and show consistent improvements across three multilingual encoders.

Limitations

We acknowledge the following limitations in this paper: (i) TAMA is constructed from X posts directed at public figures. While this focus matches high-impact moderation scenarios, the data may not fully reflect other platforms (e.g., long-form comments, multimodal posts) or richer conversational context (threads, reply chains), where targeting and intent can depend on discourse history. (ii) The benchmark centers on five prominent public figures. Although they span different regions and languages, the target set is small relative to the diversity of public figures and sociopolitical contexts worldwide. Models trained on TAMA may therefore underperform when transferred to unseen targets, emerging events, or culturally specific abuse phenomena not represented in the dataset. (iii) The fine-grained type distribution is highly imbalanced, with a few common categories dominating and several rare types having very limited support. This makes reliable estimation for low-frequency categories difficult and may bias both learning and evaluation toward frequent abuse patterns. (iv) Cascaded-MTL explicitly links tasks via conditional mechanisms. Although we observe consistent improvements, errors or miscalibration in upstream predictions can still affect downstream behavior (e.g., type or span quality under ambiguous targeting). Improving calibration and robustness under upstream uncertainty remains an important direction.

Ethical Considerations

This work studies abusive language and may include offensive content; we limit verbatim examples and encourage cautious use. Although our benchmark is built from publicly available posts, they may still be sensitive, so we follow data-minimization and recommend platform-compliant release (e.g., sharing post IDs rather than redistributing full text where applicable) and honoring deletions/takedowns. Because abuse labeling is subjective and multilingual norms vary, we encourage annotator safeguards and transparent reporting of limitations and bias. Finally, target-aware abuse detection has safety benefits but also dual-use risks (harassment, surveillance, censorship); we recommend human-in-the-loop deployment and clear documentation consistent with the ACL Code of Ethics.

Acknowledgments

This paper is funded by the Natural Science Foundation of Shandong Province under grant ZR2023QF151 and the Natural Science Foundation of China under grant 12303103.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Singh Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio C'esar Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, and 70 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv*, abs/2404.14219.

Muhammad Ahmad, Muhammad Waqas, Ameer Hamza, Sardar Usman, Ildar Batyrshin, and Grigori Sidorov. 2025. Ua-hsd-2025: Multi-lingual hate speech detection from tweets using pre-trained transformers. *Computers*, 14(6):239.

AI@Meta. 2024. Llama 3 model card.

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.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515, Montreal, Canada.

Charlotte Galpin and Patrick Vernon. 2024. Post-truth politics as discursive violence: Online abuse, the

public sphere and the figure of 'the expert'. *The British Journal of Politics and International Relations*, 26(2):423–443.

Genevieve Gorrell, Mark Greenwood, Ian Roberts, Diana Maynard, and Kalina Bontcheva. 2018. Twits, twats and twaddle: Trends in online abuse towards uk politicians. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schaefer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, and Amit Kumar Jaiswal. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages. *Preprint*, arXiv:2112.09301.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Paul Röttger, Abigail Oppong, Andiswa Bukula, Chiamaka Ijeoma Chukwuneke, Ebrahim Chekol Jibril, Elyas Abdi Ismail, Esubalew Alemneh, Hagos Tesfahun Gebremichael, Lukman Jibril Aliyu, Meriem Beloucif, Oumaima Hourrane, Rooweither Mabuya, Salomey Osei, and 8 others. 2025. AfriHate: A multilingual collection of hate speech and abusive language datasets for African languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1854–1871, Albuquerque, New Mexico. Association for Computational Linguistics.

Usman Naseem, Shuvam Shiwakoti, Siddhant Bikram Shah, Surendrabikram Thapa, and Qi Zhang. 2025. GameTox: A comprehensive dataset and analysis for enhanced toxicity detection in online gaming communities. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 440–447, Albuquerque, New Mexico. Association for Computational Linguistics.

Pia Pachinger, Janis Goldzycher, Anna Planitzer, Wojciech Kusa, Allan Hanbury, and Julia Neidhardt. 2024. AustroTox: A dataset for target-based Austrian German offensive language detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11990–12001, Bangkok, Thailand. Association for Computational Linguistics.

Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, Giulio Zizzo, Muhammad Zaid Hameed, Mark Purcell, Michael Desmond, Qian Pan, Inge Vejsbjerg, Elizabeth M. Daly, Michael Hind, Werner Geyer, and 3 others. 2025. Granite guardian: Comprehensive LLM safeguarding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 607–615, Albuquerque, New Mexico. Association for Computational Linguistics.

Julie Posetti, Diana Maynard, and Nabeelah Shabbir. 2023. *Guidelines for Monitoring Online Violence Against Female Journalists*. Office of the Organization for Security and Co-operation in Europe (OSCE).

Julie Posetti, Nabeelah Shabbir, Diana Maynard, Kalina Bontcheva, and Nermine Aboulez. 2021. The Chilling: Global trends in online violence against women journalists. Research discussion paper, UNESCO.

Isadora Salles, Francielle Vargas, and Fabrício Benevenuto. 2025. HateBRXplain: A benchmark dataset with human-annotated rationales for explainable hate speech detection in Brazilian Portuguese. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6659–6669, Abu Dhabi, UAE. Association for Computational Linguistics.

Hasan Kerem Şeker, Gökçe Uludoğan, Pelin Önal, and Arzucan Özgür. 2025. HATECAT-TR: A hate speech span detection and categorization dataset for Turkish. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25568–25579, Suzhou, China. Association for Computational Linguistics.

Sarah Sobieraj. 2020. *Credible threat: Attacks against women online and the future of democracy*. Oxford University Press.

Jens Steffek. 2010. Public accountability and the public sphere of international governance. *Ethics & International Affairs*, 24(1):45–67.

Christoph Tillmann, Aashka Trivedi, Sara Rosenthal, Santosh Borse, Rong Zhang, Avirup Sil, and Bishwaranjan Bhattacharjee. 2023. Muted: Multilingual targeted offensive speech identification and visualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 229–236, Singapore. Association for Computational Linguistics.

Manuel Tonneau, Diyi Liu, Niyati Malhotra, Scott A. Hale, Samuel Fraiberger, Victor Orozco-Olvera, and Paul Röttger. 2025. HateDay: Insights from a global hate speech dataset representative of a day on Twitter. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2297–2321, Vienna, Austria. Association for Computational Linguistics.

Emily A. Vogels. 2021. The state of online harassment. https://www.

pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/. Survey of 10,093 U.S. adults on online harassment experiences.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Qingpo Yang, Yakai Chen, Zihui Xu, Yu-ming Shang, Sanchuan Guo, and Xi Zhang. 2025. SCCD: A session-based dataset for Chinese cyberbullying detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9533–9545, Abu Dhabi, UAE. Association for Computational Linguistics.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. 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)*, pages 762–770, Toronto, Canada. Association for Computational Linguistics.

A Experimental Setup

Tasks and metrics. For T1 target detection, we report overall accuracy and macro-averaged F1 over the tri-class label set (`T1_Acc` and `T1_F1`). For T2 fine-grained abuse type classification, we report macro F1 over the 12 abuse types (`T2_F1`), computed only on *targeted-abusive* posts. For T3 phrase-level span localization, character-level spans are mapped to token-level labels and we report macro F1 over span labels (`T3_F1`). All results are reported on the held-out test split of the dataset. We construct the train/dev/test partitions with an approximately 8:1:1 ratio using per-label stratification, so that each label has coverage in all three splits. For very low-frequency labels that cannot be perfectly divided, we allow minor deviations while keeping the overall partitioning as close as possible to this ratio.

Backbones. In order to evaluate the robustness of Cascaded-MTL, three widely used encoder-only Transformer classifiers are evaluated, including: `multilingual-e5-base` (Wang et al., 2024), `XLM-RoBERTa-base` (Conneau et al., 2020) and `mDeBERTa-v3-base` (He et al., 2023).

Configurations. Inputs are truncated/padded to a maximum length of 128. We train with learning rate 2×10^{-5}, batch size 16, and 30 epochs. At the end of each epoch, we run validation and update the saved best checkpoint when development performance improves; final results are reported on the test set using the best development checkpoint.

Target	Round 1	Round 2	Final
Theresa May	0.2161	0.5201	0.7631
Liz Truss	0.3548	0.4372	0.6699
Maria Ressa	0.3769	0.5938	0.7277
Yulia Navalnaya	0.2933	0.6350	0.6637
Vladimir Soloviev	0.1343	0.5537	0.7235
Average	0.2751	0.5480	0.7096

Table 3: Inter-annotator agreement (Cohen's kappa) by target and annotation rounds.

B Cohen's kappa by target and annotation rounds

Table 3 shows a monotonic and substantial improvement in inter-annotator agreement across rounds (average κ: 0.2751 $\rightarrow$ 0.5480 $\rightarrow$ 0.7096), indicating that iterative calibration and guideline refinement effectively reduced ambiguity and stabilized labeling decisions. We also observe large target-dependent variance in Round 1 (0.1343–0.3769), which shrinks markedly in the final round (0.6637–0.7631), suggesting that the finalized guidelines generalize across different public figures and language contexts.

The relatively low agreement in Round 1 mainly reflects three factors. First, although the initial guidelines were already detailed, annotators

adopted different internal thresholds for borderline cases, especially when a post was clearly abusive but the most appropriate T2 label could be interpreted either as *general* abuse or *attacks_on_credibility*. Second, some posts plausibly matched multiple fine-grained abuse types. In the pilot stages, annotators were allowed to surface multiple candidate labels with confidence in order to expose uncertainty patterns, which naturally lowered κ; after calibration, we adopted a single-label rule requiring annotators to choose the primary, most confident label for the released dataset. Third, a small subset of cases remained intrinsically difficult because of implicit abuse, sarcasm, multilingual pragmatic cues, and borderline cases between policy criticism and personalized harassment.

To address these challenges, after each round we conducted disagreement analysis, refined the guidelines with boundary-case clarifications, held calibration discussions, and used expert arbitration for persistent conflicts. Typical recurring confusions included distinguishing *general* insults from *attacks_on_credibility*, and distinguishing *targeted-abusive* from *unidentified-targets* when the abusive expression was directed at another entity rather than the designated public figure. The monotonic increase in κ across rounds indicates that these refinements substantially reduced subjective bias and improved annotation consistency.

C T2 per-class performance

To reveal model behavior on rare abuse categories, Table 4 reports per-class F1 scores for T2 together with test-set support, using multilingual-e5-base. The model performs strongly on high-support categories such as *general* (F1 = 74.02, n = 143) and *attacks_on_credibility* (F1 = 70.64, n = 103). It also achieves competitive performance on several mid- and lower-support categories, including *misogynistic* (72.00, n = 14) and *racist* (69.57, n = 13). For very rare categories, performance is naturally more volatile due to extremely limited support, for example *religious* (40.00, n = 3), *sexual_assault* (50.00, n = 3), and *physical_harm* (40.00, n = 3). This breakdown shows that T2 is relatively reliable on well-supported abuse types, while minority categories remain the main challenge under realistic class imbalance.

T2 label	F1	#nums
death_threat	52.17	13
sexual_assault	50.00	3
sexual_explicit	45.45	13
physical_harm	40.00	3
radiation_of_threats	40.00	3
attacks_on_credibility	70.64	103
misogynistic	72.00	14
homophobic	66.67	4
religious	40.00	3
political_sectarian	50.00	10
racist	69.57	13
general	74.02	143

Table 4: Per-class F1 scores and test-set support for T2 on multilingual-e5-base.

D Annotation guidelines and definitions for T1

The detailed annotation guidelines and label definitions for T1 are shown in Table 5.

E Annotation guidelines and definitions for T2

We followed the abuse definitions from previous work by Posetti et al. (2023). The detailed annotation guidelines and label definitions for T2 are shown in Table 6.

F Benchmarking LLMs on TAMA

F.1 zero-shot prompting LLMs

We first evaluate zero-shot prompting to assess how well instruction-tuned LLMs handle TAMA without task-specific training. We test `Qwen2-7B-Instruct` (Yang et al., 2024), `Phi-3-mini-128k-instruct` (Abdin et al., 2024), and `Meta-Llama3-8B-Instruct` (AI@Meta, 2024) on the TAMA test set. For cross-lingual consistency, we use a single English prompt template while keeping the original (potentially multilingual) post text as input.

Configurations. All subtasks are formulated as closed-set generation with strict outputs and deterministic decoding (greedy; temperature $= 0$). For T1, the input includes the post text and the designated target handle (with an optional target profile for disambiguation), and the model outputs label only. The detailed prompt template of T1 is shown in Table 8. For T2 and T3, we follow the dataset dependency and evaluate only on instances whose gold T1 label is *targeted-abusive*: T2 outputs a single label, and T3 outputs only a JSON

Label	Definition	Guidelines
non-abusive	The post does not contain abusive language.	Includes neutral/supportive content or legitimate criticism without abusive expressions.
targeted-abusive	The post contains abusive language **directed at the designated target** (the public figure of the post–target pair).	The abusive expression is aimed at the target (e.g., insults, threats, degrading language), regardless of whether other entities are mentioned. This label triggers T2 (abuse type) and T3 (span localization).
unidentified-targets	The post contains abusive language, but the abuse is **not directed at the designated target**; instead it targets other entities or the intended target cannot be reliably identified as the designated public figure.	Use this label when abuse is aimed at another person/group/institution, or when the post contains abuse but does not clearly target the designated figure.

Table 5: T1 label definitions for targeted abuse detection.

Label	Definition
death_threat	A threat suggesting the target will be killed or should die.
sexual_assault	A threat suggesting the target will be, or deserves to be, sexually assaulted.
sexual_explicit	Sexually explicit language or content involving sexual acts (including references to anatomy or suggestive sexual content).
physical_harm	A threat suggesting the target will suffer other forms of physical harm, or language that incites violence against the target.
radiation_of_threats	Threats or abuse that extend beyond the target and endanger people close to them (e.g., children, parents, partners, siblings).
attacks_on_credibility	Language implying the target is unfit for their role, cannot be trusted, or insulting their intelligence/mental capacity with the aim of damaging professional reputation.
misogynistic	Belittling or degrading language toward women, or language that incites hatred toward women.
homophobic	Language expressing homophobia.
religious	Language attacking a person's religion or derogating their faith (including cases involving false assumptions about faith).
political_sectarian	Language attacking perceived political affiliations or political/sectarian philosophies.
racist	Language expressing racism.
general	Other personal insults (e.g., mild swearing or slurs) that do not fall into the more specific categories above.

Table 6: T2 label definitions for fine-grained abuse type classification.

list of abusive phrases copied verbatim from the input. The detailed prompt templates of T2 and T3 are shown in Table 9 and Table 10, respectively. To reduce format drift, we constrain decoding for T1/T2 to valid label ID outputs.

Results. Table 7 reports zero-shot prompting results on three instruction-tuned LLMs and a Qwen2-7B-Instruct zero-shot result under the HARE (Yang et al., 2023) prompting strategy. A clear pattern emerges: while LLMs can localize salient offensive phrases reasonably well (T3_F1 ranges from 64.86–71.12), they struggle to produce the structured, target-grounded decisions required by TAMA, especially for fine-grained type attribution (T2_F1 ranges from 5.48–18.79). This mismatch reflects the hierarchical nature of TAMA (target $\rightarrow$ type $\rightarrow$ span): span extraction can often be approximated by highlighting locally toxic expressions, but reliable moderation requires first resolving whether the abuse is directed at the designated public figure (T1) and then mapping the intent to a strict 12-way taxonomy (T2), which is sensitive to label semantics, multilingual variation, and long-tailed type frequencies.

We further assess prompting effects with an additional zero-shot HARE setting on Qwen2-7B-Instruct. As shown in Table 7, HARE improves T1 target detection (T1_Acc 48.88$\rightarrow$55.20, T1_F1 41.95$\rightarrow$46.11), but brings no gain on T2 (T2_F1 18.34$\rightarrow$17.96) and is nearly unchanged on T3 (T3_F1 66.51$\rightarrow$66.12), resulting in only a modest Avg_F1 increase (42.27$\rightarrow$43.40). These results suggest that appropriate prompt engineering can indeed bring slight but effective improvements to the overall detection performance. At the same time, they also indicate that most of the performance gains on TAMA mainly come from supervised modeling of task dependencies, rather than by prompt engineering alone. In future work, we will try more prompting strategies, to further examine their impact on detection performance.

Model	T1_Acc	T1_F1	T2_F1	T3_F1	Avg_F1
Qwen2	48.88	41.95	18.34	66.51	42.27
Phi3	52.98	40.23	18.79	64.86	41.29
Llama3	74.99	49.79	5.48	71.12	42.13
HARE	55.20	46.11	17.96	66.12	43.40

Table 7: Zero-shot performance between different LLMs with TAMA.

Consequently, the overall Avg_F1 of zero-shot LLMs clusters around 41–42 despite their comparatively strong T3, indicating that free-form generation does not translate into globally consistent tri-task outputs. In contrast, Cascaded-MTL explicitly aligns learning with this hierarchy by converting upstream beliefs into continuous conditioning signals (CTF/TAG/LGSD), thereby coupling global target awareness and semantic type hypotheses with evidence extraction. This dependency-aware inductive bias yields substantially higher and more stable joint performance under fine-tuning across backbones (Table 2), demonstrating that explicit task-dependency modeling is crucial for target-aware multilingual abuse moderation beyond what zero-shot prompting can offer.

F.2 Fine-tuning LLMs

We also fine-tune the above three LLMs on TAMA to evaluate their performance by using the same train/dev/test split as described in Section A. All tasks are trained with answer-only supervision, and the loss is computed only on the answer tokens by masking the prompt tokens. T1 is trained on the full split and predicts a single ID, the detailed fine-tune prompt template of T1 is shown in Table 13. Following the dataset dependency, T2 and T3 are trained and evaluated only on instances whose gold T1 label is *targeted-abusive*: T2 predicts one label ID, and T3 predicts a JSON array of abusive phrases constructed from gold span annotations. The detailed fine-tune prompt templates of T2 and T3 are shown in Table 14 and Table 15.

Model	T1_Acc	T1_F1	T2_F1	T3_F1	Avg_F1
Qwen2	86.11	73.95	45.32	79.63	66.30
Phi3	78.79	53.51	22.19	56.80	44.17
Llama3	87.40	76.58	45.47	77.99	66.68

Table 11: Performance comparison between fine-tuning different LLMs with TAMA.

Configurations. We use LoRA with r=32, α=16, and dropout 0.05 on the projection modules. We train for 10 epochs with learning rate 2×10^{-5}, per-device batch size 1 and gradient accumulation 8. The maximum sequence length is 256 for T1/T2 and 512 for T3. We select the best checkpoint by dev macro F1 at epoch end and evaluate it on the test set. At inference time, we set `max_new_tokens`=8 for T1/T2 (label ID outputs) and 256 for T3.

Results. Table 11 shows that supervised fine-tuning substantially strengthens LLMs on TAMA, but the gains are uneven across tasks, revealing a clear gap between learning what to flag and learning why/what type it is. After fine-tuning, Qwen2 and Llama3 achieve strong targeted-abuse detection and evidence extraction (e.g., T1_F1 = 73.95/76.58 and T3_F1 = 79.63/77.99), suggesting that with sufficient labeled supervision, LLMs can internalize the anchor-target setup and recover salient abusive phrases. However, fine-grained type classification remains the bottleneck: T2_F1 stays markedly lower (45.32/45.47 for Qwen2/Llama3 and only 22.19 for Phi3), which in turn limits overall Avg_F1—especially for smaller models. This pattern is consistent with TAMA's long-tailed, multilingual type space, where reliable type discrimination requires separating subtle semantic cues (e.g., credibility attacks vs. general abuse) rather than merely detecting negativity.

In contrast, Cascaded-MTL is explicitly designed to exploit TAMA's hierarchy by turning upstream beliefs into conditioning signals that constrain downstream predictions, promoting cross-task consistency between target awareness, type semantics, and span evidence. As a result, Cascaded-MTL is less prone to the common fine-tuned-LLM failure mode of producing plausible spans that are weakly aligned with the predicted type, and it offers a more reliable joint operating point for moderation scenarios that require coherent target–type–evidence outputs rather than strong single-task performance in isolation.

System

```
You are a content review expert. Given a text and a specific @Twitter account (the target), determine
whether the text contains hate speech/abusive language and whether it is directed at that target.
You will also be given a brief profile of the target account. Use the target profile ONLY to:
resolve coreference (e.g., "she/he/they", "your father", role/title),
disambiguate which person/account is being referred to,
understand context about the target (public role, identity, etc.).
Do NOT use the profile to assume abuse when the text itself is not abusive.
Choose exactly ONE of the following three classes, and output ONLY its ID:
0 = unidentified-targets (hate speech/abusive language present but NOT directed at the given target)
1 = targeted-abusive (hate speech/abusive language directed at the given target)
2 = non-abusive (no hate speech/abusive language)
Important:
Even if multiple accounts are mentioned, judge targeting ONLY with respect to the given target
account.
Output ONLY one character: 0 or 1 or 2. No explanation.
```

User

```
Target account: {TARGET_HANDLE}
Target profile: {TARGET_PROFILE}
Text:"{POST_TEXT}"
```

Table 8: Zero-shot prompt template for T1. {TARGET_HANDLE} denotes the designated target account, {TARGET_PROFILE} denotes the optional target profile (when available), and {POST_TEXT} denotes the input post text.

System

```
You are a classification expert.
The input text is already confirmed to be targeted-abusive.
Your job is to choose exactly ONE fine-grained hate/abuse category ID.
Choose exactly ONE ID from this closed set:
0 = death_threat
1 = sexual_assault
2 = sexual_explicit
3 = physical_harm
4 = radiation_of_threats
5 = attacks_on_credibility
6 = misogynistic
7 = homophobic
8 = religious
9 = political_sectarian
10 = racist
11 = general
Output ONLY the ID (an integer from 0 to 11). No explanation, no extra text.
```

User

```
Text:
"{POST_TEXT}"
```

Table 9: Zero-shot prompt template for T2. {POST_TEXT} denotes the input post text.

System

```
You are a content review expert.
Task:
Extract all hateful/abusive phrases from the given text.
Output format:
Return ONLY a JSON array (list) of strings, e.g.:
["phrase1", "phrase2"]
Rules:
Each phrase MUST appear verbatim in the input text (copy exact substrings).
If there is no hateful phrase, output [].
Output ONLY the JSON array. No explanation, no extra text.
```

User

```
Text:
"{POST_TEXT}"
```

Table 10: Zero-shot prompt template for T3. {POST_TEXT} denotes the input post text.

Partial example of the zero-shot HARE prompt for T1

```
...
Follow these steps internally (do NOT output the steps):
1) Abuse detection: identify whether the text contains hate speech/abusive language (explicit insults, slurs, dehumanization, threats, derogatory phrases).
2) Target identification: identify who the abusive content is directed at; use the target profile only for coreference/disambiguation.
3) Target grounding: decide whether the abuse is directed at the given target account (even if multiple accounts are mentioned).
4) Label decision: map the decision to exactly one ID below.
...
```

Partial example of the zero-shot HARE prompt for T2

```
...
The input text is already confirmed to be targeted-abusive.
Your job is to choose exactly ONE fine-grained hate/abuse category ID from a closed set.
Follow these steps internally (do NOT output the steps):
1) Identify the abusive expression(s) in the text (e.g., slurs, insults, threats, demeaning phrases).
2) Decide the abuse type by matching the meaning of the abusive expression(s) to the most appropriate category.
3) If multiple abuse types appear, choose the single BEST category that most directly reflects the core abusive intent.
4) Output the final ID.
...
```

Partial example of the zero-shot HARE prompt for T3

```
...
Follow these steps internally (do NOT output the steps):
1) Identify candidate hateful/abusive expressions (e.g., slurs, insults, threats, dehumanization, derogatory phrases).
2) Verbatim check: keep only candidates that appear as exact substrings in the input text (copy exactly).
3) Span tightening: keep only the minimal phrase that directly expresses hate/abuse; avoid adding extra surrounding context words.
4) Deduplicate: if the same phrase appears multiple times, include it only once.
...
```

Table 12: Partial examples of the zero-shot HARE prompts for T1–T3. Since most parts of these prompts are identical to the corresponding base zero-shot prompts, this table shows only the HARE-specific additions and modified instructions. Shared content is replaced with ellipses. User inputs, output formats, and other unchanged prompt components are omitted because they remain the same as in the corresponding base zero-shot prompt templates.

System

```
You are a content review expert.
Task: Given a target and a text, output ONLY the tri-class ID:
0 = unidentified-targets
1 = targeted-abusive
2 = non-abusive
Output ONLY one character: 0 or 1 or 2. No explanation.
```

User

```
Text: {POST_TEXT}
Target: {TARGET}
Answer:{TRI_ID}
```

Table 13: Fine-tuning prompt template for T1. {TARGET} denotes the designated target account, and {POST_TEXT} denotes the input post text. {TRI_ID} is the tri-class ID.

System

```
You are a content review expert.
Task: Given a text confirmed to be targeted-abusive. Output ONLY the fine-grained category ID.
Choose exactly ONE ID from this closed set:
0 = death_threat
1 = sexual_assault
2 = sexual_explicit
3 = physical_harm
4 = radiation_of_threats
5 = attacks_on_credibility
6 = misogynistic
7 = homophobic
8 = religious
9 = political_sectarian
10 = racist
11 = general
Output ONLY the ID (0-11). No explanation.
```

User

```
Text: {POST_TEXT}
Answer:{FINE_ID}
```

Table 14: Fine-tuning prompt template for T2. {POST_TEXT} denotes the input post text. {FINE_ID} is the fine-grained category ID.

System

```
You are a content review expert.
Task: Extract all hateful/abusive phrases from the text.
Output ONLY a JSON array (list) of strings, e.g.:
["phrase1", "phrase2"]
Rules:
Each phrase MUST appear verbatim in the text.
If none, output [].
```

User

```
Text: {POST_TEXT}
Answer:
{JSON_PHRASE_LIST}
```

Table 15: Fine-tuning prompt template for T3. {POST_TEXT} denotes the input post text. {JSON_PHRASE_LIST} is the output JSON list including all abuse spans.