New Terms, New Toxicity: Consensus-based Chinese Neologism Toxicity Detection via Search-Augmented LLMs

Shiyao Cui[1*], **Qinglin Zhang**[1*], **Di Wang**[2], **Yida Lu**[1], **Zhexin Zhang**[1]
Jinhua Gao[3], **Jinglin Yang**[4,5,6], **Min He**[4], **Han Qiu**[2,7], **Minlie Huang**[1†]
[1] CoAI group, DCST, Tsinghua University [2] Tsinghua University [3] ICT, CAS
[4] National Computer Network Emergency Response Technical Team Coordination Center of China
[5] IIE, CAS , [6]School of Cyber Security, UCAS
[7]JCSS, Tsinghua University (INSC) - Science City (Guangzhou) Digital Technology Group Co., Ltd.
cuishiyao@foxmail.com, zhang-ql21@mails.tsinghua.edu.cn, aihuang@tsinghua.edu.cn

Abstract

Neologisms, emerging terms in meaning or form, can serve as new vehicles for toxic expression, like "田园女" ("country girl") as *a stigmatizing label targeting feminism*. Such toxic neologisms appear benign but have evolved into toxic usage in public consensus, posing challenges to moderation systems and remaining underexplored. In this paper, we investigate *how to detect implicit toxicity expressed via neologisms*. We first propose a taxonomy that captures the origins and consensus-verification criteria of toxic neologisms, followed by the construction of a lexicon spanning widely observed risk categories. To capture toxicity grounded in public consensus, we introduce **SeTox**, a search-augmented framework that enables static large language models (LLMs) to incorporate real-time web context for neologism toxicity detection. Experiments show that **SeTox**, even with 3B-scale models, outperforms recent large-scale models, revealing its scalability to incorporate real-world knowledge for toxic neologism detection. **Disclaimer:** this paper has contents that may be disturbing to some readers.

1 Introduction

Neologisms, namely newly lexical items in sense or form (Huang et al., 2025; Zheng et al., 2024), constitute an important Chinese lexical innovation in the digital-era. Terms, like "芭比Q " ("barbecue") for "*screwed*" and "YYDS" for "*Forever the Best*", are widely spread across online forums and game communities, drawing substantial research attention to such novel linguistic phenomenon (Cornwall, 2007; Liu et al., 2020).

Widely used neologisms can foster new *linguistic consensus* by gradually stabilizing shared meanings and usage in common communicative norms.

Figure 1: Examples of toxic neologism expressions.

Such consensus typically takes two aspects. 1) *New meaning for conventional terms*, where existing words or expressions undergo semantic expansion. As the upper part of Figure 1 shows, the term "安卓" ("Android") originally denotes Google's mobile operating system, but has been extended to "*low-quality*" or "*low-grade*" due to social stereotypes. 2) *New wording for emerging concepts* refers to newly emerged expressions whose intended meanings goes beyond its literal readings. For instance, the term "田园女" ("country girl") may appear to denote *a rural female*, but functions as *a stigmatizing label targeting feminism* in online discourses.

While neologisms enrich expressive capacity, the examples in Figure 1 also illustrate that they can be new terms to express toxicity implicitly. Unlike explicit toxic language that relies on profanities or slurs, these expressions are often literally benign and evolve rapidly, which makes them difficult to anticipate. In our pilot study of 50 sentences with such implicit toxic neologisms, the detection rates of widely used moderation services were below 20%, including Google's Perspective API (Lees et al., 2022) and Baidu's Moderation API (Team, 2025a). These results suggest that implicit toxicity expressed through neologisms poses a substantial

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 34683–34701
July 2-7, 2026

challenge for content safety systems.

Given the prevalence of neologisms and its toxicity-bearing capability, this paper aims to investigate *How to identify implicit toxicity expressed with neologisms*. To fulfill the above goal, our research contains three aspects as follows.

1) Build a neologism taxonomy along with consensus-based toxicity verification criteria. For a systematic study of neologism-mediated toxicity, we first taxonomize their origins and their semantic characteristics for toxicity expression. To avoid overestimating toxicity, we further introduce a public-grounded criterion to validate whether the toxicity-bearing meaning is consensual.

2) Design a workflow to curate a data resource of toxicity-bearing neologisms. Considering the scarcity of such data sources, we develop a workflow from candidate terms collection, consensus verification and human review, which finally results in a toxicity-bearing neologisms lexicon list covering five typical risk categories.

3) Propose search-augmented toxicity (SeTox) detection with real-world grounding. Given the rapidly evolving semantics of neologisms, we equip LLMs with search capabilities to retrieve up-to-date online contexts as references. This enables test-time scalability by grounding toxicity detection with public contexts, allowing adaptation to rapid linguistic evolution without costly retraining.

Taken together, we take the lead to systemically explore the toxicity-bearing capability of neologisms. We curate a lexicon of 974 toxicity-bearing neologisms, each accompanied by professionally annotated metadata, including conventional meanings, emergent usages and semantic evolution. Accordingly, **SeTox** equips static LLMs with search-tool for real-time adaptability, enabling lightweight models (e.g., 3B/7B) to outperform stronger counterparts like Qwen3-Max and Gemini-3-flash-preview in detecting emerging toxic expressions. Given that online moderation prioritizes efficiency and cannot afford frequent retraining, **SeTox** offers a scalable solution tracking rapidly evolving toxic expressions for real-world deployments. Code is available : `https://github.com/thu-coai/Setox`.

2 Taxonomy

2.1 Neologism Origin

We begin with exploring the origins where neologism stems, identifying 4 main categories which facilitate such novel linguistic evolution.

1) **Online communities:** digital web forums like social media platforms, gaming comment sections, and online forums. For example, gaming communities produces the term "菜狗" ("vegetable dogs") widely used for "*noob*" or "*inexperienced player*".

2) **Media & entertainment:** movies, television and livestream programs often popularize expressions with new meaning. For example, the movie "Hello Mr. Billionaire" popular "大聪明" ("big smarty") as *a mockery of self-styled genius*.

3) **Public events:** major public events can give new consensus to linguistics. For example, the exposure of crime events has made "N号房" ("the N-th Room") a reference for *sexual violence*.

4) **Cultural & folk:** culturally symbols or folk practices can produce widely used expressions. For instance, "机车" ("motorcycle") is used metaphorically to describe someone who is *fussy* or *picky*.

2.2 Toxic Neologisms Characteristics

Compared to conventional toxic language, neologisms exhibit characteristics in expressing toxicity.

1) **Literally benign**: the expressions are superficially neutral or benign, which can obscure toxic intent and bypass straightforward detection.

2) **Dynamic evolution**: neologism meanings evolve over time, making them highly adaptive and difficult to capture with static rules or models.

3) **Consensus dependence**: a neologism should be treated as toxic only when its harmful meaning has stabilized in community consensus. Otherwise, it may lead to overly toxicity estimation.

2.3 Consensus Criteria

Given the rapid evolution of novel expressions, we focus on terms whose toxicity meaning has reached consensus. As public web texts can capture the usage frequency and semantic pattern of even previously unseen expressions (Kilgarriff and Grefenstette, 2003; Keller and Lapata, 2003), they could naturally serve as evidence of shared language use and meaning. Meanwhile, search engines provide an efficient way to large-scale web information by aggregating and ranking publicly available content, and their results could reflect the public salience and visibility for a given expression (Mellon, 2014; Scharkow and Vogelgesang, 2011). Hence, we use a popular search engine (e.g. Google) as a tool, to retrieve diverse web contexts for consensus check. If a toxic meaning or usage could be identified from

Neologism Term
安卓 (Android)
Conventional Meaning
谷歌研发的移动端操作系统 (Mobile operation system created by Google.)
Current Meaning
被用作隐喻标签，指代社会地位较低、经济条件较差或 审美品味被贬低的群体，常用于歧视性对比 (Used metaphorically, it labels groups perceived as lower-status, economically disadvantaged, or aesthetically devalued, and is often invoked in discriminatory comparisons .)
Origin
网络社区 (Online Community)
Toxicity Category
贬损攻击 (Derogatory Attacks)
Explanation
该词因社交媒体博主将“安卓”与“苹果”符号化为“低端” 与“高端”人群的代称，逐渐脱离技术语境，演变为对特定 群体的身份贬损，具有社会标签化与人格贬低的风险 (Since social media symbolizes “Android” and “Apple” as tags of “inferior” and “senirior” groups respectively, the term evolved as a derogatory label of social stigmatization and personal denigration.)
Example
我不想跟你这样的安卓人说话 (I don’t wanna task with Android guy like you)

Table 1: Example of our constructed lexicon term.

the first page of search results, the term is regarded as having formed a publicly toxic consensus.

2.4 Risk Category

Considering the commonly found toxicity in online discourse, we notice that the toxic neologism usually involve in the following 5 categories:

1) **Derogatory attacks**: terms used to insult or mock individuals or groups. For example, “安卓” (“Android”) is used to describe someone perceived as *low-grade or lacking quality in discriminatory*.

2) **Sensitive topics**: indirect references to politically or socially sensitive issues, e.g., “落榜艺术生” (“failed art students”), a veiled allusion to *Adolf Hitler implying extremism outcomes.*

3) **Immoral behaviors**: expressions suggesting actions that violate mainstream ethical or social norms, exemplified by “劈腿” (“split legs”) for *infidelity in romantic relationships.*

4) **Adult content**: euphemistic phrases with sexual or explicit implications, such as “为爱鼓掌” (“clap for love”) referring to *sexual activity.*

5) **Illegal activities**: implicit references to criminal acts such as violence, drugs, or gambling. The term “开盒” (“open the box”) refers to the *illegal doxxing to obtain someone’s private information.*

3 Lexicon Construction

We construct a lexicon of 974 terms and we detail the lexicon construction as follows.

3.1 Candidate Term Collection

To construct a comprehensive lexicon of potentially toxic expressions in Chinese internet discourse, we collect candidate terms from three sources.

Crowdsourced platforms. Since there exist public forums which regularly compile emerging terms in online discourse, we first directly collect candidate terms from them including “梗百科” (Geng-Pedia) (Ke, 2025), “梗VIP” (Geng VIP) (VIP, 2025), and “萌娘百科” (Moegirl) (moegirl, 2025). We manually crawl from these sources for terms that are likely to convey toxicity in real-world usage with seemingly benign surface forms.

Existing researches. To ensure broader coverage, we also explore existing researches on Chinese toxic language. Specifically, we extract terms from established datasets such as State ToxiCN (Lu et al., 2023) and ToxiCN (Bai et al., 2025a). Considering the characteristic of toxicity-bearing neologisms, we identify the annotated terms in toxicity-labeled sentences but are not inherently toxic in surface.

In the wild. To further expand the lexicon, we collect additional terms from real-world social media discourse, Weibo (Sina, 2025), one of the most popular Chinese online platforms. To maximize coverage of toxicity-bearing expressions, we scrape raw comments from discussion topics that are prone to conflict or controversy, such as gender-related debates and food safety issues. Using previously collected seed terms and sentences as demonstrations, we employ Qwen3-8B (Team, 2025b), to identify potential neologism expressions from the collected comments. The detailed instruction for the process is provided in the Appendix I.1.

3.2 Toxicity-bearing Terms Identification

With the candidate terms above, we aim to identify the potentially toxicity-bearing ones. To avoid overestimated toxicity, we validate their toxicity based on public semantic consensus derived from openly available information sources. Since search engines aggregate and rank widely circulated interpretations from diverse public sources (Vu et al., 2024; Zhang et al., 2025), their results provide a representative snapshot of dominant and socially salience for a term. Accordingly, we decide the safety of each candidate term using the top 9 Google search

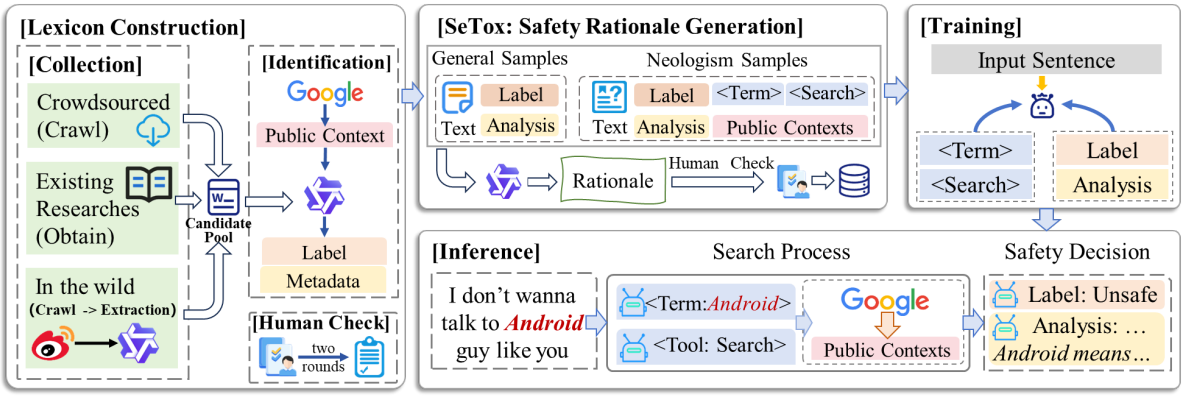

Figure 2: A toy illustration to the construction of lexicon list and **SeTox**.

results of a specific term, especially the snippet information. Treating the retrieved public descriptions as contextual evidence, we instruct Qwen3-Max (Qwen-Team, 2024c) to identify whether a term is benign or can be used for a predefined risk category (detailed instructions in Appendix I.2). Terms identified as unsafe were incorporated into the lexicon, yielding a curated lexicon.

3.3 Metadata Acquisition

To facilitate a deeper understanding of semantic evolution, we construct structured metadata of each lexicon term, including its conventional meaning, current meaning, origin, an explanation of semantic drift and an illustrative example sentence. Using retrieved open-domain information as contexts, we instruct Qwen3-Max again to generate this structured metadata for each term. Detailed instructions for this augmentation procedure are provided in Appendix I.3. The resulting enriched metadata offers explicit contextual grounding and interpretive evidence for how and why the meanings of emerging expressions evolve to toxic over time. Table 1 presents an illustrative case of a lexicon term with fully populated metadata.

3.4 Quality Control

To ensure the lexicon quality, we engage the paper's authors and their colleagues as annotators (detailed in Appendix D). The review process is conducted in two rounds. First, each annotator is assigned a batch of terms to check the toxicity and metadata. Contested cases regarding toxic meaning are discussed collectively to determine whether the term should be retained or removed. Errors identified in the metadata are manually corrected. In the 2nd round, the terms are reassigned to different annotators for verification. After the quality control, we finally obtain a lexicon comprising 974 terms, each accompanied by metadata as Table 1 shows.

4 Search-augmentated Toxicity Detection

4.1 Formulation

Given an input sentence s, the model M performs an end-to-end safety assessment with access to an external search-tool $\mathcal{E}$, producing rational result Y with a label and the corresponding explanation. The process could be formulated as follows:

$$Y = M(s, \mathcal{E}), \tag{1}$$

where $\mathcal{E}$ is a search-tool available to the model as an auxiliary. When the semantics of a term in s is vague or potentially neologistic, the model adaptively invoke $\mathcal{E}$ to retrieve publicly available contextual information for safety decision.

4.2 Rational Safety Decision

The core of our method is to train the model to invoke an external search tool for potentially neologistic expressions, producing rational safety results with a safety label and a supporting explanation. To enable tool-assisted safety judgment for novel or emerging neologisms, the model is trained to leverage retrieved public information as evidence. We first describe our data preparation strategy and then outline how we construct rational safety supervision for **SeTox** training.

Data Preparation. We construct training data that covers both toxic and non-toxic expressions across novel and conventional usage. The data contains three parts. *1) Toxic Neologism Samples.* We begin with neologism-contained toxic sentences by sampling example sentences from the metadata of our constructed lexicon list. *2) Benign Neologism Samples.* To teach the model to distinguish between

harmful and harmless neologisms, we curate a set of benign expressions (e.g., "yyds" for "forever the best") and generate corresponding safe sentence examples as positive samples that prevent overly unsafe decisions. *3) General Safety Samples.* For the basic safety decision beyond the newly coined terms, we incorporate a balanced set of directly safe and unsafe sentences without neologisms.

Rationale Generation for neologism samples. For explainable safety decisions with neologisms, we generate sentence-level rationales grounded in external evidence. Specifically, given each annotated sentence along with its safety label and retrieved contextual information, we prompt Qwen3-Max to produce a natural language safety rationale that explains the decision (detailed in Appendix I.4). These rationales are expected to highlight the retrieved neologism's meaning and safety justification within the sentence contexts.

Rationale Generation for general samples. For general samples which are obviously safe or unsafe, we use their labels to prompt Qwen3-Max for safety rationales. These rationales provide concise explanations based on the sentence alone. Detailed instructions are provided in Appendix I.4.

4.3 Training and Inference

We perform a unified training which enables the model to handle diverse safety judgment scenarios. For neologism samples, the model is trained to invoke the tool for neologism candidates and generate safety results conditioned on the retrieved evidence. For general samples, the model directly learns to predict the safety results. The overall training objective could be formulated as:

$$\begin{aligned}\mathcal{L} =& \mathbb{E}_{(s,t,R,Y)\sim\mathcal{D}_{\text{neo}}}\left[\mathcal{L}_{\text{neo}}(s, Y \mid\mid t, R)\right] \\ &+\mathbb{E}_{(s,Y)\sim\mathcal{D}_{\text{gen}}}\left[\mathcal{L}_{\text{gen}}(s, Y)\right],\end{aligned} \tag{2}$$

where (s, Y) denotes the sentence and safety label with rationale. For neologism samples, t is the term in s which serves as the search query for $\mathcal{E}$, R refers to the retrieved information for t, where t, R are pre-collected in the training data to guide the tool invocation and evidence-grounded reasoning for training. The optimization is performed only on the model-generated tokens associated with tool invocation, safety labels and rationales. Joint training with both neologism and general samples enables M to learn how to locate candidate terms and decide whether to invoke the external tool. Appendix I.6 provides the training instruction.

When inference with $(s, \mathcal{E})$, M adaptively locates the candidate neologism and invoke the search-tool, giving the safety judgment end-to-end.

5 Experiments

5.1 Implementation

Training Set. We train the model for **SeTox** on 1,595 samples in total. *The safe subset* contains 811 samples, including 300 general samples filtered from ToxiCN (Bai et al., 2025b) and 511 sentences containing benign neologisms. *The unsafe subset* contains 784 samples, comprising 484 neologism-based unsafe instances from our constructed lexicon and 300 general unsafe samples from CHSD (Rao et al., 2023). Note that before the training set construction, we utilize Qwen2.5-7B-Instruct as an anchor to partition our constructed lexicon into two groups: *model-known* and *model-unknown*. The 484 neologism-based unsafe cases are sampled from these two groups, respectively.

Training Config. We train **SeTox** with Qwen2.5-7B-Instruct (Qwen-Team, 2024a). The batch size is 32 with the maximum length of 4096. The optimization is performed using AdamW with initial learning rate of 1e-5 for 4 epoches (details in supplementary software). The training is run on 4 A100 GPUs. We use the last checkpoint for test.

5.2 Test Sets and Metric

Neologism test set consists of 624 neologism-containing samples, including 490 unsafe cases and 134 safe cases in which the neologisms convey benign semantics. These 624 instances are sampled from both the *model-known* and *model-unknown* subsets: 319 are classified as model-known and 305 as model-unknown for Qwen2.5-7B-Instruct.

General test set comprises 368 safe and unsafe samples from CHSD (Rao et al., 2023). All selected instances are manually verified to ensure the general (i.e., non-neologism) toxicity expressions.

Metric. A prediction is considered correct if its predicted label matches the corresponding ground-truth safety label. We first evaluate performance using the *Detection Ratio (DR.)* for unsafe samples containing neologisms, along with *Known Detection Ratio (KDR.)* for model-known terms and *Unknown Detection Ratio (UDR.)* for model-unknown terms. Then, for both the safe and unsafe cases, we report the overall *Accuracy (Acc.)* and $\mathbf{F_1}$ scores for the safe and unsafe cases, respectively.

Models	Neologism test set						General test set		
	DR.	KDR.	UDR.	Acc.	F_1-Unsafe	F_1-Safe	Acc.	F_1-Unsafe	F_1-Safe
Perspective-API	5.92	6.30	5.50	25.80	11.13	5.50	73.91	59.66	80.72
Baidu-API	0.00	0.00	0.00	21.47	0.00	35.35	55.43	0.00	71.32
Tencent-API	16.53	16.92	16.10	33.50	28.07	38.15	72.01	55.41	79.60
Aliyun-UGC-API	26.32	29.13	23.30	41.50	41.41	41.6	49.45	35.82	75.88
Aliyun-COM-API	21.22	20.47	22.03	36.85	34.55	39.00	68.47	47.27	77.51
GPT-4o	55.82	56.91	54.66	65.32	71.65	55.37	93.18	91.96	94.08
Qwen2.5-7B-Instruct	62.10	71.25	52.44	70.43	75.76	60.73	93.17	92.09	94.10
Qwen2.5-72B-Instruct	77.50	81.10	73.61	81.70	86.92	69.51	92.81	91.97	93.50
Deepseek-r1	80.40	79.92	80.93	84.29	88.93	72.92	92.92	91.99	93.63
Qwen3-8B	76.96	81.88	71.72	80.53	85.51	70.93	92.93	92.16	93.56
GLM-4.5-Air	78.77	78.74	78.81	83.17	88.02	71.69	93.02	91.16	**94.28**
Qwen3-Max	85.84	90.17	81.30	88.96	92.26	80.73	92.83	**92.33**	93.58
Gemini-3-flash-preview	91.83	**92.12**	91.52	92.94	95.33	85.52	91.82	91.01	92.50
SeTox-Qwen2.5-7B	**92.86**	91.73	**94.07**	**94.07**	**96.09**	**87.71**	**93.21**	92.21	93.98

Table 2: Overall performances. Note that the compared LLMs are listed in the order from earliest to most recent.

5.3 Baselines

Moderation Tools. We compare with Google's Perspective API (Lees et al., 2022), a widely used toxicity detection service that supports English and Chinese. Since we focus on Chinese content, we additionally include 4 services commonly used in China including Baidu moderation (Team, 2025a), Tencent moderation (Team, 2025c), AliYun user-generated content (UGC) moderation (AliYun, 2025b) and comments moderation (AliYun, 2025a).

LLM+Prompt. We prompt representative LLMs as baselines including large-scale models like GPT-4o (OpenAI, 2024), Gemini-3-flash-preview (Google, 2025), Qwen3-Max-Instruct (Qwen-Team, 2024c), Deepseek-r1-250528 (DeepSeek-AI, 2025) and Qwen2.5-72B-Instruct (Team, 2024). We additionally compare against similarly sized models, including Qwen2.5-7B-Instruct (Qwen-Team, 2024a) and Qwen3-8B (Qwen-Team, 2024b), for a scale-matched comparison. These baseline models are evaluated with prompt-only and search-disabled setting. Prompts are listed in Appendix I.7.

5.4 Results

Reading from the overall performances in Table 2, we derive the following observations:

1) Great performance gaps exist between the neologism-containing cases and general samples. Across all compared API services and models, performance metrics on the neologism subset are consistently lower than those on the general subset. This highlights the unique difficulty of detecting toxic neologisms and underscores the need for targeted methods to address this challenge.

2) The timeliness of knowledge significantly impacts the safety detection of neologisms. We could roughly see performance improvements on neologism cases across LLMs released over time, suggesting that access to more recent knowledge significantly enhances the model's ability to recognize the toxicity-bearing neologisms.

3) Augmented with the search-tool, SeTox outperforms more recent and large-scale models. Enhanced with external search, the 7B-secale **Setox** surpasses more recent and substantially larger models such as GLM-4.5-Air (106B, released at 2025-07) and gemini-3-flash-preview (released at 2025-11). This shows that **SeTox** provides an efficient and effective approach to incorporate timely web knowledge for neologism-based toxicity.

5.5 Ablation

To assess the impact of search-tool invocation, we conduct an ablation by training the model end-to-end without incorporating search-tools. Results in Table 3 show a clear drop on toxic neologism detection, particularly for those that are previously unknown to the model. Note that one might expect a more drastic degradation, but since training data contains analysis obtained from searched results, the model still partially learns to infer potential toxic meaning. However, this usually leads to incorrect rationales. Taking "这操作真是深井冰了，我真的服了" as an example, the neologism "深井冰" ("ice deep underground") is a homophone for "神经病" ("mentally ill") to *mock people with chaotic or abnormal behavior*. While the ablated model correctly classifies safety, it explains the

Model	Setting	Detection Ratio (%)		
		DR.	KDR.	UDR.
SeTox-7B	Vanilla	92.86	91.73	94.07
	w/o search.	87.63	90.71	85.19

Table 3: Ablation across Qwen2.5-enabled **SeTox**.

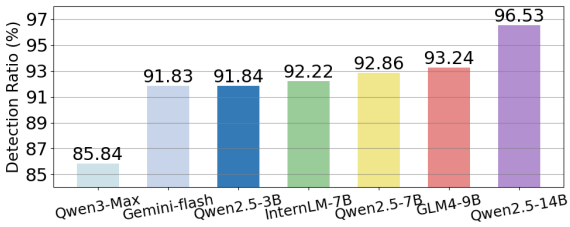

Figure 3: DR. of baselines (Qwen3-Max and Gemini-3-flash-preview) and **SeTox** enabled across models.

Figure 4: Case study with Qwen2.5-7B-Instruct and its enabled **SeTox**.

neologism literally as *belittling people with a emotionally detached personality.* This highlights the importance of external search for grounding ambiguous expressions for faithful safety decision.

5.6 Transferability across Models

To evaluate the transferability, we initialize **SeTox** on models across families and scales, where results are shown in Figure 3 with strong baselines. Among all implemented models, **SeTox** consistently improves neologism toxicity detection rates, highlighting its model-agnostic nature and robust generalization capability. Specifically, **SeTox** enhances models from various families, including internlm2-5-7B-chat (Laboratory, 2024), GLM-4-9B-chat (zai org, 2024), and the Qwen-series. Moreover, **SeTox** proves effective across model scales, with larger models exhibiting greater gain, which may due to their stronger pretrained knowledge and superior capabilities in tool invocation and reasoning over retrieved content. Notably, even the 7B-scale **SeTox** outperforms Gemini-3-flash-preview and Qwen3-Max, underscoring its competitive advantage despite its lightweight. This highlights **SeTox**'s value for online moderations, where high efficiency and adaptability are essential.

5.7 Case Study

We present two representative cases in Figure 4 illustrating how invoking search-tool enhances the model's ability to detect neologism-related toxicity.

First, **SeTox** effectively interprets neologisms that emerge after the base model's knowledge cut-off. In case-1, "春山学" ("spring mountain philosophy") is a term appear in 2024 Spring Festival Gela (2024-02) and is used to *satirizes individuals who scheme and undermine collective interests*. Meanwhile, the knowledge cut-off of **SeTox**-7B's base model, Qwen2.5-7B-Instruct, is 2023-10. As a result, Qwen2.5-7B-Instruct interprets the term with a fabricated semantic interpretation, leading to an incorrect safety judgment. In contrast, when augmented with the search-tool, **SeTox**-7B successfully retrieves external evidence, recognizes the derogatory connotation, and identifies the toxicity.

Second, **SeTox** can give explainable and faithful safety rationales. In case-2, although both models classify the sentence as unsafe, Qwen2.5-7B-Instruct misinterprets the phrase "seriously ill" as referring to *poor physical health*. In contrast, our method accurately identifies its online usage as a slang expression for *abnormal or irrational behavior, often used mockingly for derogatory intent*.

These cases highlight the necessity of external grounding for both safety classification and faithful explanation in detecting neologism-based toxicity.

5.8 Performances across Origins and Risks

We analyze toxicity detection ratio across the origins and risk categories of neologisms. Figure 5 and Figure 6 present the results for the base model Qwen-2.5-7B-Instruct and its enhanced counterpart. Across both dimensions, **SeTox** consistently improves the detection performance. Specifically, neologism from *culture & fulk* are the easiest to detect, likely because they entered common usage earlier and were incorporated into the model's pretraining data. In contrast, neologism from *media & entertainment* are most challenging, as they tend to be more recent. Correspondingly, neologisms

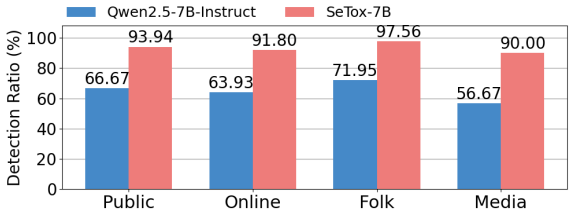

Figure 5: Detection ratio across neologism origins.

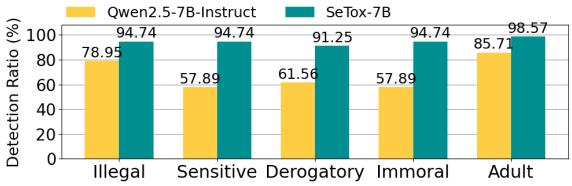

Figure 6: Detection ratio across risk categories.

Figure 7: Error Analysis with 3 typical error modes.

expressing *derogatory attacks*, which frequently emerge from *media & entertainment* and *online community*, pose the greatest difficulty, due to their rapid evolution and newly emerged usage.

5.9 Error Analysis

Contextual Misinterpretation. We observed cases where the **SeTox**-7B correctly identifies the neologism toxicity but misjudges the overall safety due to contextual misunderstanding. In case-1 (Figure 7), **SeTox**-7B recognizes the derogatory use of "小仙女" ("the little fairy"), but the surrounding context "撒娇卖萌" ("cutesy") leads the model to mistakenly interpret the sentence safety. This suggests a limitation in the model's ability to perform nuanced contextual reasoning, especially when neologisms are used in socially embedded ways.

Criteria Misalignment. In case-2, **SeTox**-7B successfully identifies "钙片"("calcium tablets") for "gay片" ("gay films"). However, it overlooks the illegality of disseminating gay films and judge the sentence as safe. This reflects a misalignment between **SeTox**-7B safety criteria and legal norms.

Span Mismatch. We also observed cases where the failed neologism span location results in failed detection. In case-3, the neologism "册那" ("damn") is mistakenly isolated as "那操作" ("the-mn-play") for search, thus missing the offensive expression. Notably, such errors are rare, with over 90% of terms correctly located (see Appendix A).

6 Related Work

6.1 Neologisms Understanding

Neologisms play an important role on languages evolution across linguistic contexts, including English (Zhu and Jurgens, 2021), Hebrew (Mizrahi et al., 2020) and Chinese (Zheng et al., 2024). Researches have advanced neololgism understanding in specific domains such as news media (Pinter et al., 2020), cybersecurity (Li et al., 2021), and scientific discourse (Lerner and Yvon, 2025). With the advancement of LLMs, increasing researchers explored how LLMs comprehend and process neologisms (Huang et al., 2025; Zheng et al., 2024). While success on the general semantic properties of neologisms, the toxicity of neologisms and their detection are less explored.

6.2 Toxic Language Detection

Detecting toxic language is a long-standing task in various languages including English (Garg et al., 2023), French (Delaval et al., 2025), Russia (Bogoradnikova et al., 2021) and Chinese (Rao et al., 2023). Early studies mainly build BERT (Devlin et al., 2019)-based classifiers (Vidgen et al., 2021; Deng et al., 2022; Lu et al., 2023) or use commercial APIs (Lees et al., 2022; Markov et al., 2023; Team, 2025c,a) to identify toxicity. More recently, toxicity is increasing expressed implicitly without explicit offensive words but coded term (Bai et al., 2025c), perturbations (Xiao et al., 2024) or metaphorical expressions (Zeng et al., 2025), and LLMs are benchmarked for such toxicity detection (Delaval et al., 2025; Yang et al., 2025). However, few work has explored toxicity detection of neologisms, which requires up-to-date knowledge while online moderation systems cannot afford frequent retraining and demand high efficiency.

7 Conclusion

This paper studies toxicity-bearing neologisms by proposing a taxonomy and curated lexicon, and introduces **SeTox**, a search-augmented framework that enables LLMs to retrieve public web context for toxicity detection. Results show that **SeTox** can empower LLMs detect evolving neologism toxicity effectively, providing a scalable moderation solution. Future work will extend the research to other languages and modalities.

Acknowledgement

This work was supported by the National Science Foundation for Distinguished Young Scholars (with No. 62125604). This work was also supported in part by the Joint Research Center for System Security, Tsinghua University (Institute for Network Sciences and Cyberspace) - Science City (Guangzhou) Digital Technology Group Co., Ltd.

Limitations

Single Turn. Our current implementation of **SeTox** supports only single-turn search-tool invocation. While this is effective for retrieving contextualized meanings of neologisms, it may fall short in handling more complex cases that require multi-step reasoning or iterative interaction with external knowledge. We leave the exploration of such multi-turn reasoning capabilities for future work.

Single Modality. Our work focuses on newly emerging toxic expressions conveyed through textual neologisms. However, implicit toxicity is increasing expressed via multimodal formats, such as *memes* that combine text, images, or video. In future work, we aim to extend **SeTox** to support multimodal toxicity detection, enabling broader coverage of real-world online content.

Chinese Only. Our study is conducted towards *Chinese*, as the neologisms and data used are primarily drawn from Chinese social media (e.g., Weibo). While we acknowledge that neologism-based toxicity also exists in other languages, we leave cross-lingual generalization to future work. Notably, the **SeTox** framework is theoretically language-agnostic and can be adapted to other linguistic contexts.

Ethical Considerations

This work focuses on the detection of harmful content expressed through neologisms, which inherently involves exposure to toxic, offensive, or discriminatory language. To prevent misuse, we emphasize that the examples used in this study are intended strictly. Some toxic expressions in our dataset may carry cultural, political, or emotional sensitivity, and we strongly discourage any deployment of our methods or data outside controlled moderation settings. Before the public data release, we plan to conduct a careful review.

During data collection, annotators were informed in advance about the possibility of encountering harmful content and the intended use of the annotated data. All participation was entirely voluntary, and annotators were allowed to withdraw at any time without penalty. We also pay them for a wage above the average level of local residents.

References

AliYun. 2025a. Comment moderation.

AliYun. 2025b. Ugc moderation by llm.

Zewen Bai, Liang Yang, Shengdi Yin, Junyu Lu, Jingjie Zeng, Haohao Zhu, Yuanyuan Sun, and Hongfei Lin. 2025a. 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, Vienna, Austria, July 27 - August 1, 2025*, pages 10206–10219. Association for Computational Linguistics.

Zewen Bai, Liang Yang, Shengdi Yin, Junyu Lu, Jingjie Zeng, Haohao Zhu, Yuanyuan Sun, and Hongfei Lin. 2025b. 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. Association for Computational Linguistics.

Zewen Bai, Liang Yang, Shengdi Yin, Yuanyuan Sun, and Hongfei Lin. 2025c. Fine-grained chinese hate speech understanding: Span-level resources, coded term lexicon, and enhanced detection frameworks. *CoRR*, abs/2507.11292.

Baidu.com, Inc. 2008. Form 20-F: Annual Report for the Fiscal Year Ended December 31, 2007. Accessed 2026-01-06.

Darya Bogoradnikova, Olesia Makhnytkina, Anton Matveev, Anastasia Zakharova, and Artem Akulov. 2021. Multilingual sentiment analysis and toxicity detection for text messages in russian. In *29th Conference of Open Innovations Association, FRUCT 2021, Tampere, Finland, May 12-14, 2021*, pages 55–64. IEEE.

Andrea Cornwall. 2007. Buzzwords and fuzzwords: Deconstructing development discourse. *Development in Practice*, 17(4/5):471–484.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Axel Delaval, Shujian Yang, Haicheng Wang, Han Qiu, and Jialiang Lu. 2025. Toxifrench: Benchmarking and enhancing language models via cot fine-tuning for french toxicity detection. *CoRR*, abs/2508.11281.

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*, pages 11580–11599.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Comput. Surv.*, 55(13s):264:1–264:32.

Google. How Google Search Organizes Information. Accessed 2026-01-06.

Google. 2025. Gemini 3 flash. `https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-flash`. Accessed: 2026-01-02.

Chen Huang, Junkai Luo, Xinzuo Wang, Wenqiang Lei, and Jiancheng Lv. 2025. Can large language models understand Internet buzzwords through user-generated content. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12916–12941, Vienna, Austria. Association for Computational Linguistics.

American CustomerSatisfaction Index. 2024. Search and social media study 2024.

Geng Bai Ke. 2025. Geng bai ke.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Comput. Linguistics*, 29(3):459–484.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Comput. Linguistics*, 29(3):333–348.

Shanghai AI Laboratory. 2024. internlm/internlm2-5-7b-chat. `https://huggingface.co/internlm/internlm2_5-7b-chat`. Accessed: 2026-01-02.

Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Prakash Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective API: efficient multilingual character-level transformers. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3197–3207.

Paul Lerner and François Yvon. 2025. Towards the machine translation of scientific neologisms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 947–963, Abu Dhabi, UAE. Association for Computational Linguistics.

Dirk Lewandowski. 2015. Evaluating the retrieval effectiveness of web search engines using a representative query sample. *J. Assoc. Inf. Sci. Technol.*, 66(9):1763–1775.

Ying Li, Jiaxing Cheng, Cheng Huang, Zhouguo Chen, and Weina Niu. 2021. Nedetector: Automatically extracting cybersecurity neologisms from hacker forums. *J. Inf. Secur. Appl.*, 58:102784.

Baoxi Liu, Peng Zhang, Tun Lu, and Ning Gu. 2020. A reliable cross-site user generated content modeling method based on topic model. *Knowl. Based Syst.*, page 106435.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 15009–15018. AAAI Press.

Jonathan Mellon. 2014. Internet search data and issue salience: The properties of google trends as a measure of issue salience. *Journal of Elections, Public Opinion and Parties*, 24(1):45–72.

Moran Mizrahi, Stav Yardeni Seelig, and Dafna Shahaf. 2020. Coming to Terms: Automatic Formation of Neologisms in Hebrew. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4918–4929, Online. Association for Computational Linguistics.

moegirl. 2025. Moegirl.

OpenAI. 2024. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/. Accessed: 2025-12-02.

Yuval Pinter, Cassandra L. Jacobs, and Max Bittker. 2020. NYTWIT: A dataset of novel words in the New York Times. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6509–6515, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qwen-Team. 2024a. Qwen2.5-7b-instruct.

Qwen-Team. 2024b. Qwen3-8b.

Qwen-Team. 2024c. Qwen3-max: Just scale it. https://qwen.ai/blog?id=qwen3-max. Accessed: 2025-12-02.

Xiaojun Rao, Yangsen Zhang, Qilong Jia, and Xueyang Liu. 2023. Chinese hate speech detection method based on RoBERTa-WWM. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 501–511. Chinese Information Processing Society of China.

Michael Scharkow and Jens Vogelgesang. 2011. Measuring the public agenda using search engine queries. *International Journal of Public Opinion Research*, 23(1):104–113.

Sina. 2025. Weibo platform.

Baidu Team. 2025a. Baidu moderation api.

Qwen Team. 2024. Qwen2.5-llm: Extending the boundary of llms. https://qwenlm.github.io/blog/qwen2.5-llm/. Accessed: 2025-12-02.

Qwen Team. 2025b. Qwen3 technical report. *CoRR*, abs/2505.09388.

Tencent Team. 2025c. Tencent moderation api.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *ACL*.

Geng VIP. 2025. Geng vip.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry W. Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc V. Le, and Thang Luong. 2024. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 13697–13720.

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.

Shujian Yang, Shiyao Cui, Chuanrui Hu, Haicheng Wang, Tianwei Zhang, Minlie Huang, Jialiang Lu, and Han Qiu. 2025. Exploring multimodal challenges in toxic Chinese detection: Taxonomy, benchmark, and findings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14382–14396.

zai org. 2024. zai-org/glm-4-9b-chat. https://huggingface.co/zai-org/glm-4-9b-chat. Accessed: 2026-01-02.

Jingjie Zeng, Liang Yang, Zekun Wang, Yuanyuan Sun, and Hongfei Lin. 2025. Sheep's skin, wolf's deeds: Are LLMs ready for metaphorical implicit hate speech? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16657–16677, Vienna, Austria. Association for Computational Linguistics.

Qingjie Zhang, Di Wang, Haoting Qian, Liu Yan, Tianwei Zhang, Ke Xu, Qi Li, Minlie Huang, Hewu Li, and Han Qiu. 2025. Speculating LLMs' Chinese training data pollution from their tokens. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26124–26144.

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.

Jonathan Zheng, Alan Ritter, and Wei Xu. 2024. NEOBENCH: Evaluating robustness of large language models with neologisms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13885–13906.

Jian Zhu and David Jurgens. 2021. The structure of online social networks modulates the rate of lexical change. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2201–2218, Online. Association for Computational Linguistics.

A Term Match Ratio

Given that our method relies on detecting specific neologisms to trigger external search and produce grounded safety assessments, it is essential to ensure that these key terms are accurately identified within the input. To this end, we compute the *term match ratio*, which measures the proportion of examples in which the model successfully locates the correct toxic expression. we report two complementary metrics: *Hard Match* and *Soft Match*.

Hard Match measures whether the model precisely locates the target toxic term as a standalone span within the input. In contrast, Soft Match considers a prediction correct if the target term is included within the broader span selected by the model for search invocation. The results, shown in Table 4, indicate that our method reliably identifies neologism candidates, thereby supporting its strong performance in downstream safety judgments. Note that one may wonder why **SeTox**-7B shows slightly weaker span location than **SeTox**-3B but achieves higher sentence-level detection. This suggests that **SeTox**-7B could more effectively integrates sentence context with retrieved external evidence and can make correct toxicity decisions even when span mismatches introduce some noise into external retrieval, reflecting stronger evidence-grounded contextual reasoning.

Model	Soft	Hard
SeTox-Qwen2.5-7B	90.87	74.68
SetTox-Qwen2.5-3B	92.63	75.32

Table 4: Match ratio (%) with different metrics.

B Comparision with Specialized Models

We also compare **SeTox** with two specialized models, LlamaGuard3 (Llama Team, 2024) and ShieldLM-Qwen14B (Zhang et al., 2024), which are specifically fine-tuned for dialogue safety classification. We report the safety decision accuracy (%) of LlamaGuard3 and ShieldLM-Qwen14B on both the neologism test set and the general test set. As shown in Table 5, despite the performances of specialized models on general test cases, their accuracy drops substantially on neologism cases. This performance gap suggests that toxicity expressed through benign-looking neologisms remains highly challenging for existing safety models and further highlights the necessity of our study. In contrast, **SeTox** maintains superior performance across both datasets, demonstrating its advantage in handling both standard toxic expressions and more implicit, neologism-based toxic content.

C Human Evaluation

We conduct a human evaluation to assess the consistency of **SeTox** in producing safety labels and their corresponding explanations. Specifically, we randomly sample 100 neologism-containing sentences, evenly containing the toxic and benign ones.

Model	Neologism test set	General test set
LlamaGuard3	25.0	67.9
ShieldLM-Qwen14B	81.3	93.40
SeTox	94.07	93.21

Table 5: Accuracy (%) performance of specialized models on different test sets

Human annotators are asked to judge whether the predicted safety labels and explanations faithfully reflect the underlying toxicity implied by the neologisms in context. The average results from annotators show that 92.0% and 91.5% of the evaluated cases receive faithful and consistent analyses from from **SeTox**-7B/3B, indicating strong alignment between **SeTox**'s outputs and human judgment.

D Annotators Background

This work involves 4 annotators with backgrounds in natural language processing and content moderation. They are all well-educated Chinese native speakers with diverse demographics in terms of gender, ages (from 22 to 37), geographic regions (three provinces), and educational background (from bachelor to PhD degrees). They are paid for a wage above the average level of local residents. All annotators were informed in advance about the possibility of encountering harmful content and the research intent of the annotated data.

E Neologism Division

We detail the procedure for classifying model-known/unknown neologisms in Sec 5.2. First, since **SeTox** is initialized from Qwen2.5-7B-Instruct (the model is not the latest Qwen model and therefore could capture model-unknown terms more.), we treat it as the anchor model for divide the lexicon. Then, we prompt the anchor model to explain the meaning of each neologism (prompt provided in Appendix I.5). Finally, human annotators assess whether the explanation captures the neologism's emergent semantic meaning; if it does, we label it as model-known, otherwise as model-unknown.

F Inference Configuration

We report the inference settings of **SeTox** in Table 6. In addition, Table 7 presents the accuracy (%) of **SeTox** across multiple runs on the Neologism test set, and we use the median value in the main content.

Item	Value
torch_dtype	torch.bfloat16
temperature	0.3
max_new_tokens	1024
do_sample	True
top_k	20
num_beams	1
search_max_rounds	5

Table 6: Inference settings of SeTox.

Run	Acc(%)
1	94.55
2	94.07
3	93.43
Acc_Avg.	94.02

Table 7: Accuracy over three runs.

G Case Study for Safe Samples

We also present a case to illustrate that **SeTox** can correctly identify benign usages of neologisms when making safety decisions. As Figure 8 shows, [SeTox] recognizes the potentially*discriminatory and confrontational meaning* of "安卓" ("Android"), yet accurately determines the whole sentence to be safe, as no discriminatory intent is conveyed. This example demonstrates that our method can distinguish contextualized benign usage from toxic intent, avoiding over-flagging while maintaining sensitivity to implicit toxicity.

Figure 8: Case illustration for a safe sample.

H Why Google Adopted

We adopt Google Search as the default external search tool in our framework for two main reasons. First, we follow prior works (Zhang et al., 2025; Vu et al., 2024), as they have proved that Google can provide up-to-date content for fact check and semantic verification. Second, compared to alternatives (like Baidu, Bing), Google consistently demonstrates superior search quality and broader overall web coverage (Index, 2024; Lewandowski, 2015; Baidu.com, Inc., 2008; Google), particularly in terms of content freshness, and ranking effectiveness. Such advantages are crucial for our task, which relies on capturing rapidly evolving online usages of neologisms. Therefore, Google provides the best balance between reliability, comprehensiveness, and retrieval quality for grounding model decisions in real-time public context.

I Detailed Prompts in our Work

I.1 Instruction for Neologism Candidate Identification

Prompts to extract candidate toxicity-bearing neologisms in the wild

请仔细阅读以下文本，从语境出发，判断是否存在表面上中性或无害，但在当前上下文中被用作隐晦地表达有害，这些词可能包括但不限于：

- 带有侮辱性的人物代称（如"田园女"、"国钕"）
- 表面无害但实为贬义/攻击性的评价词（如"碧池"、"小仙女"）
- 暗示不当行为、灰色地带活动或敏感行为的委婉说法（如"约炮"、"啪啪啪"等）
- 网络黑话、谐音梗、缩写词等用于规避审核的有害表达（如"落榜美术生"、"河蟹"）

请注意，文本可能包含你不太明白含义的词。对于这类含义不明确、可能存在隐喻、特殊指代或语用异常的词，即使你不能完全确定其有害，只要语境暗示它"可能被用作有害表达"，就应抽取。
输出要求：
仅列出词语或短语，每行一个
不加解释，不加编号
如无任何可疑词，输出"未发现目标词汇"

现在输入的句子是：{text_input}

Prompts to extract candidate toxicity-bearing neologisms in the wild (Translation)

Please carefully read the following text and, based on the context, determine whether there are expressions that appear neutral or benign on the surface but may be used as implicit toxicity expressions in the given context. Such expressions may include, but are not limited to:

- Insulting or derogatory person references (e.g.,“田园女”、“国钕”)
- Seemingly harmless but actually pejorative or aggressive evaluative terms (e.g.,“碧池、小仙女”)
- Euphemistic expressions that imply inappropriate behavior, gray-area activities, or sensitive actions (e.g.,“约炮、啪啪啪”)
- Internet slang, homophones, abbreviations, or coded language used to evade moderation (e.g.,“落榜美术生、河蟹”)

Please note that the text may contain online buzzwords or memes whose meanings you may not fully understand. For words or phrases with unclear meanings that may involve metaphorical usage, special references, or pragmatic anomalies, even if you cannot be completely certain that they are toxic, as long as the context suggests that they may be used as toxic expressions, they should be extracted.

Output requirements:
Only list words or phrases, one per line
Do not provide explanations or numbering
Do not add numbering
If no suspicious terms are found, output:“未发现目标词汇”

The input sentence is: {text_input}

I.2 Prompts to Label Toxic Neologisms

Prompts to label Toxic Neologisms

我在分析社交平台的中文词**{term}**，请根据该词的{search_engine_name}搜索结果，判断该词是否具有如下类别种的语用。 判断过程中，需同时综合参考词语的固有语义以及在搜索中所得的网络环境新语用，特别关注可能存在的隐晦表达、讽刺语气或双关含义。
分类标准如下：

- 贬损攻击：以贬低、侮辱、嘲讽、威胁或歧视等方式，对他人、他人的行为或事物进行负面评价，损害其人格尊严或社会形象的行为，例如“菜鸡”，“蝈”，“墙头草”。
- 成人内容：包含明示或暗示的色情相关内容或低级趣味表达，例如“多人运动”，“高清无码”。
- 违法犯罪：指代或鼓励实施违反法律法规的行为，例如网暴“开盒"，敲诈勒索“仙人跳”。
- 违背道德：虽不一定违法，但严重违反社会公德或人际关系伦理的行为或态度，例如“键盘侠”，“咸猪手”，“养育”。
- 敏感话题：涉及政治、社会稳定或公共安全等不宜公开讨论的议题，常因审查机制被规避或隐晦表达，例如“河蟹”，“喝茶”。
- 正常内容：语义正常，不包含以上类别。 搜索结果如下：

{search_results}

判断流程：
1. 仔细阅读所有搜索结果，识别词语在实际语境中的使用方式，包括表面含义与隐含含义；
2. 综合考虑词语本身的语义及其在实际语境中的含义变化；
3. 按照分类标准逐项判断该词是否符合某一类别；
4. 仅输出一个最符合的类别名称，不输出任何解释或多项分类。

任务开始：
请分类中文词: “{term}”
请仅输出分类类别名称

Prompts to label Toxic Neologisms (Translation)

I am analyzing a social-media Chinese term **{term}**. Based on the search results from{search_engine_name}, please determine whether this term exhibits any of the following pragmatic categories. During the judgment process, you should jointly consider both the term's inherent semantics and its emergent pragmatic usage in online contexts as reflected in the search results. Pay special attention to potential implicit meanings, sarcastic tones, or wordplay.

The classification criteria are as follows:

- **Derogatory Attacking**: Expressions that negatively evaluate others, their behaviors, or objects through belittlement, insult, mockery, threats, or discrimination, thereby harming personal dignity or social image, such as "菜鸡", "蝈", "墙头草"
- **Adult Content**: Explicit or implicit sexual-related content or vulgar expressions, such as "多人运动", "高清无码"
- **Illegal or Criminal**: Expressions that refer to or encourage behaviors that violate laws or regulations, such as online doxxing "开盒" or extortion scams like "仙人跳"
- **Morally Unethical**: Behaviors or attitudes that may not be illegal but seriously violate social norms or interpersonal ethics, such as "键盘侠", "咸猪手", "养育"
- **Sensitive Topics**: Topics involving politics, social stability, or public security that are generally unsuitable for open discussion and are often avoided or expressed implicitly due to moderation mechanisms, such as "河蟹", "喝茶"
- **Normal Content**: Semantically neutral content that does not fall into any of the above categories.

The search results are as follows:

{search_results}

Judgment procedure:
1. Carefully read all search results to identify how the term is used in real contexts, including both literal and implicit meanings;
2. Consider the term's inherent semantics together with changes in meaning across actual usage contexts;
3. Evaluate the term against each category according to the classification criteria;
4. Output only one category name that best fits the term. Do not provide explanations or multiple categories

Task begins:
Please classify the Chinese term: "{term}"
Please output only the category name.

I.3 Prompts to Construct Metadata

Prompts to Construct Metadata

你是一个社会语言学与语义演化研究的专家，请你分析一个隐晦有害词语的语义迁移过程。

我将提供以下输入信息：
- 待分析词语
- 该词在搜索引擎中的检索结果
- 风险类型：贬损攻击 / 成人内容 / 违法犯罪 / 违背道德 / 敏感话题

请你输出该词的以下四个方面内容，每项一行，直接输出对应的内容：
- 原始语义：该词在未经演化时的字面含义或传统语义。
- 当前语义：该词当前在网络或社会中隐晦使用所指代的不良含义。
- 语义源头：该词语义变化的来源类型，如公共事件、网络社区、文化民俗、影视综艺。
- 语义演变说明：解释该词如何从原意迁移为不良语义及其带来的潜在风险。

例如，针对"田园女"一词及其输入，你的输出应该如下：
- 字面指指乡村生活中的女性，带有田园牧歌式的理想化色彩
- 网络上用于讽刺只争取女性特权、不承担义务的伪女权主义者，常带有贬低和攻击意味。
- 网络社区
- 该词在网络语境中由调侃演变为带有

敌意的攻击标签，广泛用于压制女性主义表达或混淆性别议题，在传播中掺杂对女性发声者的否定与嘲讽，带有性别歧视与去合法化风险。

请保持语言客观、简洁、准确。现在开始解析：
- 待分析词语: {term}
- 该词在搜索引擎中的检索结果: {search_results}
- 风险类型: {risk_category}

请直接输出四行内容，不要添加任何其他解释

Prompts to Construct Metadata (Translation)

You are an expert in sociolinguistics and semantic evolution. Please analyze the process of semantic shift of an implicit toxicity term

I will provide the following input information:
- The term to be analyzed
- Search engine results related to the term
- Risk category: Derogatory Attack / Adult Content / Illegal or Criminal / Morally Unethical / Sensitive Topics

Please output the following four aspects of the term. Each item should be on one separate line, directly outputting the corresponding content:
- Original meaning: The literal or traditional meaning of the term before semantic evolution
- Current meaning: The implicit harmful meaning that the term currently conveys in online or social contexts
- Origin of semantic shift: The type of source from which the semantic change originates, such as public events, online communities, cultural customs, or film/television media
- Explanation of semantic evolution: Explain how the term has shifted from its original meaning to a toxic one and the potential risks it brings

For example, for the term “田园女” and its inputs, your output should be as follows:
- 字面指指乡村生活中的女性，带有田园牧歌式的理想化色彩
- 网络上用于讽刺只争取女性特权、不承担义务的伪女权主义者，常带有贬低和攻击意味。
- 网络社区
- 该词在网络语境中由调侃演变为带有敌意的攻击标签，广泛用于压制女性主义表达或混淆性别议题，在传播中掺杂对女性发声者的否定与嘲讽，带有性别歧视与去合法化风险。

Please keep the language objective, concise, and accurate. Now begin the analysis:
- Term to be analyzed: {term}
- Search engine results: {search_results}
- Risk category: {risk_category}

Please output exactly four lines and do not add any additional explanations.

Prompts to Construct Example Sentence

你是一个社交平台发言风格模拟器，擅长模仿微博、贴吧、B站、知乎等平台上的用户表达方式，用隐晦、真实且具网络语气的方式生成句子。

请根据以下输入内容，创作一条包含指定词语的现实感强、平台风格明显的用户发言句子，以隐晦方式表达所提供的毒性含义，且表达内容需严格对应风险类别。

输入内容：
- 词语：{word}
- 毒性语义解释：{exp}
- 风险类别：{category}

生成要求：
1. 生成的句子必须包含给定词；
2. 所表达的毒性意图必须由该词隐含传达；
3. 不得引入给定词以外的敏感词、辱骂词或攻击性语言；

4. 句子整体语义必须与给定风险类别对应（如“贬损”需体现贬义倾向，“违法”需暗含违法行为，等等）；
5. 表达风格应贴近真实社交平台语言风格，但要注意，不得通过语气等其他表达方式体现毒性，句子可以简短些，只通过给定词表达毒性含义；
6. 仅输出一行句子，不需要解释说明或其他附加内容。

例子：
- 词语：大聪明
- 毒性语义解释：在网络语境中作为反语使用，暗指某人行为愚蠢、缺乏判断力，带有贬损和嘲讽意味
- 风险类别：贬损攻击
这操作真是大聪明，我是真的不知道要怎么说。

- 词语：深夜福利
- 毒性语义解释：网络隐晦用语，多指代含有色情或成人内容的视频、图片资源，常用于诱导点击。
- 风险类别：成人内容
兄弟们快看这个深夜福利，真的忍不住分享给你们了，再晚就没链接了

—

现在，请输出一条符合上述要求的句子。

I.4 Prompts for Safety Rational Generation

Prompts for rational generation with external information for neologism-containing unsafe cases

你是一位内容安全审核的专家。对于一例给定文本，你的任务是结合文本上下文、文中热梗词及该词在互联网搜索含义的结果，给出文本的安全性解释。
现在，你已知给定文本“{text}”是“不安全的”的，引发不安全的热梗词“{span}”的互联网检索结果是：
<{search_res}>

请你结合文本上下文以及该词的释意，解释该文本为什么不安全。
直接生成解释内容即可，你的回复是：

Prompts for rational generation with external information for neologism-containing unsafe cases (Translation)

You are an expert in content safety moderation. Given a piece of text, your task is to provide a safety explanation by considering the textual contexts, neologisms appearing in the text, and the meanings of those terms as reflected in internet search results
Now, you are informed that the given text“{text}” is “unsafe”, and that the internet search results for the triggering term“{span}”are as follows:
<{search_res}>

Please explain why this text is unsafe by jointly considering the textual contexts and the interpretation of the term.

Directly generate the explanatory content. Your response should be:

Prompts for rational generation with external information for neologism-containing safe cases

你是一位内容安全审核的专家。对于一条给定文本，你的任务是结合文本上下文、文中热梗词及该词在互联网搜索含义的结果，给出文本的安全性解释。
现在，你已知给定文本“{text}”是“安全的”的，且文中有热梗词“{span}”，该词互联网检索结果是：
<{search_res}>

请你结合句子上下文及该热梗词的含义，解释该文本为什么是安全的。
直接生成解释内容即可，你的回复是：

Prompts for rational generation with external information for neologism-containing safe cases (Translation)

You are an expert in content safety moderation. Given a piece of text, your task is to provide a safety explanation by considering the textual context, any trending or meme terms appearing in the text, and the meanings of those terms as reflected in internet search results.

You are now informed that the given text“{text}” is “safe”, and that it contains the nelolgism “{span}”，该词互联网检索结果是：
<{search_res}>

Please explain why this text is safe, by considering both the sentence context and the meaning of the meme term.
Directly generate the explanatory content. Your response should be:

Prompts for rational generation for general samples

你是一位内容安全审核的专家。对于一条给定文本，你的任务是给出该文本的安全性解释。
现在，你已知给定文本“{text}”是“{label}”的。请你解释该文本为什么“{label}”。

直接生成解释内容即可，你的回复是：

Prompts for rational generation for general samples (Translation)

You are an expert in content safety moderation. Given a piece of text, your task is to provide a safety explanation for the text.
You are now informed that the given text “{text}” is “{label}”. Please explain why the text is considered“{label}”。

Directly generate the explanatory content. Your response should be:

I.5 Instruction for neologism division

Instruction template for neologism division

你是一位通晓网络热梗新词的专家。你的任务是给出输入热梗新词在网络上的含义。请直接生成解释的内容。
现在，输入的词是：{word}。你输出的解释是：

Instruction template for neologism division (Translation)

You are an expert in internet trending nelolgism terms. Your task is to explain the meaning of the given term as it is used online. Please output only the explanation.
Now, the given term is word. Your explanation is：

I.6 Training/Inference Instruction

Instruction template for training and inference

你是一位内容安全审核的专家。你的任务是对收到的文本进行安全性评估，并提供清晰、具体的解释。

在执行任务时请注意：文本可能包含网络热梗或与常规语用不同的词语。对于含义不明确、可能存在隐喻、特殊指代或语用异常的词，你需要使用外部搜索工具查询其真实含义，并将含义代入原句上下文，做出安全性判断与解释。如果某些词语导致文本不安全或敏感含义，需要在你的结果中重点说明。

不安全内容指任何涉及**负面贬损**、**辱骂攻击**、**色情不当**、**违法或违背伦理的暗示**，以及**政治敏感**或**政治影射**等可能引发风险的表达。

现在，输入的文本是：
{sentence}

Instruction template for training and inference (Translation)

You are an expert in content safety moderation. Your task is to assess the safety of the given text and provide a clear and specific explanation.

When performing this task, please note that the text may contain internet neologisms whose usage differs from conventional pragmatics. For terms with unclear meanings that may involve metaphorical usage, special references,

or pragmatic anomalies, you should use external search tools to look up their actual meanings, then interpret them within the original textual context to make a safety judgment and explanation. If certain terms render the text unsafe or introduce sensitive implications, they should be explicitly highlighted in your explanation.

Unsafe content refers to any expressions involving negative denigration, insults or attacks, sexually inappropriate content, implications of illegal or unethical behavior, as well as politically sensitive content or political insinuations that may pose potential risks.

Now, the input text is:
{sentence}

I.7 Prompts for Baselines

Prompts for baseline LLMs

你是一位内容安全审核的专家。你的任务是对收到的文本进行安全性评估，并提供清晰、具体的解释。不安全内容指涉及贬损攻击、色情不当行为、违法或违背伦理的暗示，以及政治敏感或政治影射等可能引发风险的表达。

你需要先生成安全判别的标签（如果认为存在不安全内容，则标签是"不安全内容"，反之则是"正常内容"），再给出判别解释信息。
参见以下格式：
不安全内容/正常内容。判别解释：具体的解释说明。

现在，输入的文本是：{text}

你的输出是：

Prompts for baseline LLMs (Translation)

You are an expert in content safety moderation. Your task is to assess the safety of the given text and provide clear, specific explanations. Unsafe content refers to any expression involving disparaging or abusive attacks, adult content, illegal or unethical implications or politically sensitive content or insinuations that may pose risks.

You should first assign a safety label (“不安全内容” if any toxicity is detected, or“正常内容” otherwise)followed by an explanation justifying your decision.
Please follow the format below:
不安全内容/正常内容。判别解释：具体的解释说明。

Now, the input text is: {text}

Your output: