"Shut Up and Let Me Enjoy My Otome": Understanding and Measuring the Toxicity in Otome Game Communities

Yage Zhang[1] Xinyue Shen[1,2] Yukun Jiang[1] Michael Backes[1] Yang Zhang[1]*

[1]*CISPA Helmholtz Center for Information Security* [2]*University of Waterloo*

Abstract

Otome games, a romance simulation genre primarily targeting female, have emerged as a major force in the global gaming market, attracting hundreds of millions of players and billions in revenue. Despite their popularity, otome game communities face pervasive online toxicity, which has been largely unexplored. In this work, we present the first large-scale measurement of toxicity in otome game communities across social platforms. We introduce OtomeSCAN, a framework for collecting, evaluating, and analyzing 620,045 posts from Weibo and Reddit spanning 18 months. To support robust analysis, we manually annotated a ground-truth dataset of 4,308 posts, identifying eight target groups such as players and game developers. We evaluate seven toxicity detectors on the dataset, including general-purpose models and our proposed LLM-based detectors, with our best model achieving F1-scores of 0.82 (Weibo) and 0.78 (Reddit). Our analysis reveals significant platform-based differences in toxicity: 22.20% of otome-related posts on Weibo are toxic, compared to 3.71% on Reddit. Besides, real-world events like in-community conflicts can rapidly escalate toxicity, with toxicity ratios increasing to 37.09% in just 72 hours during an external attack on Weibo. We also flag 191 potential-coordination clusters in otome game communities, 64.40% of which target game developers, with several accounts participating repeatedly across multiple clusters. We hope our work inspires further research on community-specific toxicity and contributes to building healthier online spaces for marginalized gaming communities.[1]

Disclaimer: This paper contains examples of toxic and abusive language. Reader discretion is recommended.

1 Introduction

Digital games have become an important component of everyday leisure for hundreds of millions of people worldwide. Within this landscape, online toxicity in gaming communities has attracted growing research attention, with studies documenting gender-based hostility [79], coordinated harassment [67], and hate speech dynamics in multiplayer environments [49, 51]. However, existing work overwhelmingly focuses on general mixed-gender games, leaving ecosystems with fundamentally different social structures largely unexplored.

Figure 1: Examples of toxic posts and target groups for otome games on Weibo (left) and Reddit (right). We paraphrased the examples to prevent verbatim searches from identifying the users while preserving the original meaning.

Otome games, a narrative-driven romance simulation genre in which players take on the role of a female protagonist and develop romantic relationships with non-player characters (NPCs) [68], represent one such ecosystem. According to recent reports, the global otome games market reached approximately USD 5.26 billion in 2024 [25]. One flagship title, *Love and Deepspace*, reportedly reached 50 million global users by early 2025 [36]. Yet, otome games continue to face widespread online toxicity and discrimination, as illustrated in Figure 1. In mainstream gaming communities, otome games are frequently criticized for encouraging women to challenge traditional gender norms [40]. Within otome game communities, in-group harassment and interpersonal hostility are common [3, 4]. A notable incident occurred in August 2024 [17, 18], when a rapper released a satirical song mocking otome game players on the social platform. The song quickly went viral, reaching millions of

*Corresponding author.

[1]Our dataset is available at https://huggingface.co/datasets/TrustAIRLab/OtomeSCAN.

users, and was followed by a surge of toxic posts towards otome game players, rising from 22.20% to 37.09% within 72 hours (see Section 5). However, the research community still lacks a systematic understanding of toxicity in otome game communities, including its prevalence, targeted groups, temporal dynamics, and linguistic characteristics. This gap significantly hinders efforts to address and mitigate online toxicity faced by the otome game players, primarily millions of female players.

Our Work. In this work, we present the first large-scale measurement study on toxicity in otome game communities. Specifically, we focus on the following research questions:

- **RQ1:** How does toxicity in otome game communities differ from patterns observed in general game communities, in terms of prevalence, target groups, and interaction patterns, and how do the themes of toxic disputes differ across platforms?
- **RQ2:** What types of events are associated with significant peaks of toxicity in otome game communities?
- **RQ3:** What linguistic features are present in toxic posts from otome game communities, and how do they evade platform moderation?
- **RQ4:** Beyond individual toxic posts, what patterns of potential coordination appear in otome game communities? Who are their primary targets?

To answer these questions, we introduce OtomeSCAN, a framework designed for collecting, evaluating, and analyzing the toxicity in otome game communities. Leveraging OtomeSCAN, we collect 620,045 posts from Weibo and Reddit, covering four otome game communities and two general game communities (used later as control groups), spanning from January 2024 to May 2025. Given the lack of prior work evaluating the performance of toxicity detectors on otome game content, we randomly sampled and manually annotated 4,308 posts to serve as a ground truth dataset. This annotation includes two levels of labels: (1) binary toxicity (toxic or non-toxic) and (2) target groups for toxic posts. In the end, we identified eight target groups in the otome-related toxic posts, which are players, NPCs, game developers, platform moderators, policymakers, identity groups, other game-related entities, and unknown (see Section 3.1.4). We then evaluate three general-purpose toxicity detectors, i.e., Perspective API, OpenAI Moderation API, COLD, and four of our proposed LLM-driven detectors on the annotated set. Our best-performing model achieved F1-scores of 0.82 on Weibo and 0.78 on Reddit, significantly outperforming the general-purpose detectors. We then employ it to annotate the full dataset (see Section 3.2).

Regarding analysis, we start by performing a comparative analysis of toxicity in otome game communities and general game communities, focusing on the prevalence, target groups, and user interaction patterns across social platforms, and comparing the themes of toxic disputes between the two platforms (RQ1). We then conduct a time series analysis of toxic posts to identify real-world events that coincide with significant toxicity peaks (RQ2). Through linguistic analysis, we investigate toxic spans in toxic posts and the variation strategies they contain, which may help toxic content evade platform moderation (RQ3). Finally, we apply a similarity-based detection procedure to flag candidate clusters exhibiting potential coordination in otome game communities (RQ4).

Main Findings. We make the following main findings:

- Compared with general game communities, otome game communities show platform-specific differences in both prevalence and target distribution. On Weibo, otome discussions are far more toxic than general gaming discussions (22.20% vs. 3.43%), with a distinctively higher toxicity ratio targeting players (47.2% vs. 27.4%). On Reddit, overall toxicity ratios are similar (3.71% vs. 3.67%), but otome toxicity shifts toward NPCs rather than game companies (see Section 4).
- Real-world events like in-community conflicts, game updates, external attacks, and consumer rights protests are associated with significant toxicity surges in otome game communities. Across several events, the distribution of targeted groups shifts over time and increasingly includes players (see Section 5).
- Toxic posts in otome game communities show distinct linguistic patterns across platforms. 39.89% of model-detected unique toxic spans on Weibo involve variation strategies like slang, abbreviation, and substitution, compared with 22.77% on Reddit, where toxicity is expressed more directly (see Section 6).
- Our procedure flags 191 potential-coordination clusters in otome game communities, of which 64.40% target game developers. Several accounts participate repeatedly across clusters (see Section 7).

Contributions. Our work makes three main contributions: First, we present the first large-scale empirical study of toxicity in otome game communities. By analyzing 620,045 posts collected from Weibo and Reddit, we uncover the prevalence, target groups, temporal dynamics, and linguistic characteristics of toxicity in otome game communities. These findings provide valuable insights for game developers and platform moderators to better understand and manage the online environments of otome game communities. Second, we propose an LLM-driven classifier tailored for detecting toxicity in otome game communities, achieving F1-scores of 0.82 on Weibo and 0.78 on Reddit. This classifier offers a strong foundation for future mitigation efforts. We publicly release our dataset of 4,308 manually annotated posts on Hugging Face to support future research on training and evaluating toxicity detectors (Appendix A). Third, we characterize potential-coordination clusters in otome game communities, highlighting observable patterns that warrant further investigation by moderators. Despite the substantial user base and market potential of otome games, our study reveals that the efforts to govern toxicity in their communities remain minimal, echoing the longstanding neglect that otome games have

Table 1: Commercial scale (USD) and social-media engagement of popular otome games [7]. We collect Weibo and Reddit post counts on June 7, 2025, and they represent cumulative engagement up to that date.

Game Title (English)	Game Title (Chinese)	2024 Revenue (M USD)	Weibo Posts / Members	Reddit Posts / Members
Love and Deepspace	恋与深空	722.71	146.1M / 4.3M	98K / 140K
Beyond the World	世界之外	178.20	48.0M / 1.4M	- / -
Light and Night	光与夜之恋	153.16	250.6M / 5.1M	79 / 443
Ashes of the Kingdom	如鸢（代号鸢）	77.50	184.2M / 1.5M	26 / 7
Mr. Love: Queen's Choice	恋与制作人	40.28	176.7M / 2.9M	11K / 10K
Tears of Themis	未定事件簿	39.32	118.4M / 2.0M	13K / 28K
Lovebrush Chronicles	时空中的绘旅人	29.89	131.3M / 4.3M	1.08K / 3.1K

(a) Home Screen **(b) Story Mode** **(c) Gacha**

(d) Combat Mechanism

Figure 2: Otome game screenshots from the *Love and Deepspace*. The game features emotional companionship and fighting alongside male non-player characters.

faced in the gaming industry. We call for greater attention to the unique challenges faced by otome game communities.

2 Preliminaries and Related Work

Otome Games. *Otome game* is a narrative-driven romance simulation genre where the player typically takes on the role of a female protagonist to develop relationships with male non-player characters, through dialogue, choice-based interactions, and combat [68]. A representative example is *Love and Deepspace*, as shown in Figure 2, which features a main RPG storyline, card-collecting systems, clue-gathering side episodes, and Live2D cut-ins that deepen immersion. As illustrated in Table 1, top otome games typically generate hundreds of millions of dollars in annual revenue and attract millions of followers on social media. Players congregate in platform-specific communities, such as Weibo's *Super Topics* [57, 70] and subreddits like r/*otomegames* [14], for discussion and collective action [32].

Compared to general, mixed-gender gaming communities such as those around *League of Legends* or *Genshin Impact*, otome communities have several structural features that are likely to shape distinct toxicity patterns. First, the predominantly single-gender player base [52] creates intra-community identity conflicts different from the external gender-based hostility documented in mixed-gender games [79]. Second, the romance-driven design fosters parasocial bonds with characters [44], turning game updates into emotionally charged community disputes. Third, fan-circle culture [58] introduces organized collective behaviors, such as coordinated comment control, voting campaigns, and targeted harassment, that more closely resemble coordinated influence campaigns than general gaming toxicity. These genre-specific factors motivate a dedicated study rather than direct extrapolation from existing work on general game communities.

Prior work on otome games mainly examines emotional attachment, social support, intimacy, gender, and fan labor. For example, Lei et al. [54] explore how players of *Mr. Love* seek and provide social support within otome communities. Other work studies parasocial romantic relationships between female players and male non-player characters [39, 42], the negotiation of female gaze and erotic material under regulatory constraints [52], fan labor in online otome communities [37], and cosplay commission as a form of commodified or co-created intimacy in the otome community [84]. These studies establish otome games as socially and emotionally consequential spaces. However, the toxicity in otome game communities remains underexplored, such as its prevalence, target groups, and interaction patterns. Our work aims to fill this gap.

Definition of Toxicity and Toxicity Detection. In this study, we follow prior work [47, 75] to adopt the Perspective API's definition of toxicity: "*a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion*" [53]. For boundary cases involving product or narrative criticism, profanity or strong dissatisfaction alone is not sufficient for a toxicity label; we therefore require that the expression contain direct insults, slurs, threats, or harassment toward a person, group, or character (see Appendix E).

Prior research on gaming communities has extensively documented gender-based hostility, finding that women and LGBTQ+ players face elevated levels of harassment [79],

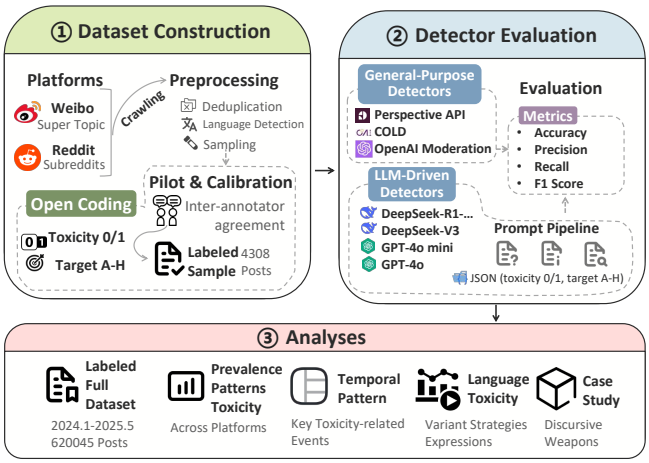


Figure 3: Overview of OtomeSCAN framework.

and that underperforming male players can become more hostile toward female players [49]. While studies of otome games have explored fan labor and parasocial relationships [37–39], a critical gap remains in quantitatively understanding the toxicity in otome game communities. This gap further poses a significant challenge to the development of toxicity detection in otome game communities. While research on large-scale corpora [27, 35, 82] has led to widely-used tools like Perspective API [53] and OpenAI Moderation [63], they can be evaded through adversarial attacks like coded speech and homophone substitutions [45]. The unique, slang-filled discourse of otome communities presents a specific challenge that existing models are ill-equipped to handle, a gap this study aims to address.

3 OtomeSCAN

In this section, we introduce OtomeSCAN, a framework designed for collecting, evaluating, and analyzing the toxicity in otome game communities. The overview of OtomeSCAN is shown in Figure 3.

3.1 Dataset Construction

3.1.1 Investigated Communities

We analyze toxicity in four otome game communities, referred to as *study groups*, and compare them with two general game communities, serving as *control groups*.

Study Groups. Our study groups focus on the top three otome games most frequently discussed in Chinese- and English-speaking communities, and general otome community, as shown in Table 1.

- ***Love and Deepspace*** (2024) introduces full-3D models and action gameplay, achieving massive commercial success by attracting over 50 million players in its first year [5, 20].
- ***Mr. Love: Queen's Choice*** (2017) is an early adopter of interactive phone-call events, garnering over 10 million downloads in China during its first year [2].
- ***Tears of Themis*** (2020) blends courtroom investigation with romance, accumulating over 20 million global installs [26].
- **General Otome Community:** Sourced from *r/otomegames* subreddit and Weibo's "otome games" tag, this group reflects a broad and diverse player base of otome games.

Control Groups. To distinguish the toxic patterns unique to otome game communities, we include two control groups from broader game communities.

- ***Genshin Impact*** (2020) is an open-world RPG game that attracts a mixed-gender player base [23]. Although not an otome game, its ongoing story updates and monetization model resemble those of otome games, making it a reasonable game-level comparison.
- **General Game Community:** Aggregated from *r/gaming* on Reddit and the "Games" tag on Weibo, this group is used to compare with the general otome community. We regard it as a genre-level comparison.

3.1.2 Data Collection

We select Reddit and Weibo as the primary studied social platforms for two main reasons. First, both platforms provide dedicated communities for fan gatherings (e.g., Reddit's subreddits and Weibo's Super Topics), which we denoted as *communities* in this study. These communities naturally segment fan groups and thus enable us to directly analyze the behaviors of different game fandoms. Second, these communities are fan-governed, with volunteer moderators on Reddit [59] and community hosts on Weibo's Super Topics [57, 70] managing the daily activity. They are also closely monitored by game companies, who track them for player feedback and brand management [37]. This dual nature encourages player participation and makes them reliable sources for observing how players negotiate conflict and express grievances. Specifically, our data collection process is suited to each platform's structure.

- **Reddit:** Reddit is a platform that provides autonomous, volunteer-moderated communities, namely *subreddits*. Following prior studies [22, 51], we use Arctic Shift[2] an open-source crawler to collect all posts from the subreddits of the study and control groups, that is *r/LoveAndDeepspace* [12], *r/MrLove* [13], *r/TearsOfThemis* [15], *r/Genshin_Impact* [11], *r/otomegames* [14] and *r/gaming* [10].
- **Weibo:** Similar to Reddit's subreddits, Weibo, one of the largest Chinese-speaking social media platforms, also holds fandom-based communities called *Super Topics*. Our retrieval strategy is twofold. For specific games, we collect all posts from their dedicated Super Topics, that is *Love and Deepspace* [9], *Mr. Love: Queen's Choice* [6], *Tears of Themis* [16], and *Genshin Impact* [8]. For broader

[2] https://github.com/ArthurHeitmann/arctic_shift.

Table 2: The overview of the collected dataset.

Game Type	Community	Weibo				Reddit			
		Source	# Posts	% Chinese	# Sample	Source	# Posts	% English	# Sample
General Games	General Game Community	Keyword	63,295	98.09%	383	Subreddit	78,721	94.12%	383
	Genshin Impact	Super Topic	6,041	98.39%	363	Subreddit	196,943	91.12%	384
Otome Games	General Otome Community	Keyword	122,622	99.03%	384	Subreddit	11,980	96.72%	373
	Love And Deepspace	Super Topic	20,328	99.16%	377	Subreddit	89,833	96.46%	383
	Mr. Love	Super Topic	13,563	99.85%	374	Subreddit	419	97.14%	201
	Tears Of Themis	Super Topic	14,031	99.74%	374	Subreddit	2,269	93.61%	329
Total			**239,880**	**99.04%**	**2,255**		**380,165**	**94.86%**	**2,053**

groups such as the general otome community and the general game community, which lack official Super Topics, we instead rely on keyword searches to ensure comprehensive coverage. All Weibo data is collect using the `weibo-search` tool,[3] with the detailed keyword search strategy described in Appendix D.

3.1.3 Data Pre-Processing

Deduplication. To ensure that each sample in the dataset represents a unique post, we perform a deduplication process. We identify and remove any duplicate posts based on their unique post identifiers (`ID`) assigned by the platform. We also exclude posts marked as deleted or removed on either platform during this stage (see Appendix B).

Language Verification. Although Reddit's subreddits and Weibo's Super Topics analyzed in this study are typically considered English- and Chinese-speaking communities, respectively, we verified their language distributions through the `lingua` toolkit.[4] Our results show that 94.86% of Reddit posts are in English, while 99.04% of Weibo posts are in Chinese, as illustrated in Table 2. Given this high degree of monolingualism, we henceforth treat them as English-speaking and Chinese-speaking communities in our analysis. We acknowledge this choice may exclude content in other languages, and we discuss this further in Section 8.

Data Statistics. In total, we collect 239,880 posts from Weibo and 380,165 posts from Reddit, which, to the best of our knowledge, corresponds to the largest dataset to date on community discussions of otome games. For each post, we collect its ID, content, creation time, number of comments, tags, and user IDs. Our data collection spans from January 1, 2024, to May 31, 2025. This time frame allows us to capture the initial growth phase of a newly launched game (*Love and Deepspace*, released on January 18, 2024) and multiple content update cycles of other established games. Table 2 summarizes the statistics of posts across platforms and communities, including sample sizes and language distributions. Note that only 419 posts are collected from the *Mr. Love* subreddit. This is because, by 2024, the game was already entered its eighth year: while its Chinese-speaking community remains active, the English-speaking market begins to contract, resulting in a decline in user postings. Nevertheless, we include this community in our dataset because it provides valuable and irreplaceable perspectives on community dynamics in the late stages of an otome game's lifecycle, as later discussed in Section 5.

[3] https://github.com/dataabc/weibo-search.
[4] https://github.com/pemistahl/lingua.

3.1.4 Data Sampling and Human Annotation

Given that toxicity in game communities remains largely unquantified, we begin our analysis by sampling data from the collected dataset. Notably, the sampled dataset includes both otome and general game communities, as we aim to provide a systematic comparison among them. We then manually annotate two types of labels: toxicity and target groups, to gain deeper insights into the nature of toxicity within these communities.

Sampling. Following previous studies [24, 56], we calculate the minimum required sample size for each community to balance manageability with statistical representativeness. Specifically, we first apply the standard formula for estimating a population proportion [1], then adjust the result using the finite population correction (FPC) [1, 56] to account for the specific size of each community (details in Appendix C). The final sample sizes are summarized in Table 2, including 2,255 Weibo posts and 2,053 Reddit posts.

We then perform human annotation to establish ground-truth labels for subsequent evaluation and analysis. The annotation process is designed to produce two levels of labels: *1) Toxicity:* each sample is labeled as either toxic or non-toxic, based on the definition of toxicity outlined in Section 2. These labels serve as the ground truth for evaluating toxicity detectors in Section 3.2; *2) Target groups:* for samples identified as toxic, annotators employ open coding to identify the specific groups targeted. This enables a more fine-grained investigation of the affected groups in otome game communities. To avoid conflating disagreement with toxicity, we apply explicit boundary rules during annotation. Representative boundary cases are provided in Appendix E.

To ensure both rigor and domain relevance, we structure our annotation process in two phases: a pilot study to develop and calibrate the annotation schema, followed by full-scale annotation of the sampled set. The annotation is led by two expert annotators with over six years of experience as otome game players.

Pilot Study. We first sample 578 posts from the sample set to conduct a pilot study. In this phase, two annotators independently labeled each post for toxicity and performed open coding to identify the target groups referenced in the toxic

Table 3: Codebook of toxicity labels and target groups. We paraphrased the examples to prevent verbatim searches from identifying the users while preserving the original meaning.

Task	NO.	Code	Description	Example	% Weibo	% Reddit
Toxicity	0	non-toxic	Language not likely to drive others away	`This update offers too little content for its price.`	81.48%	95.50%
	1	toxic	Language likely to drive others away	`Everyone in this fandom is a pathetic loser, just shut up already.`	18.52%	4.50%
Target Groups	A	players	Game players and communities	`Otome fans are clueless idiots who turn every thread into a fight.`	43.33%	7.71%
	B	NPCs	In-game characters like non-player characters (NPCs) and main character (MC)	`That love interest is a disgusting loser, and I'm sick of seeing him.`	5.96%	33.89%
	C	game developers	Game content, developers, designers, and marketing	`The devs are greedy morons who treat players like wallets.`	34.19%	22.10%
	D	platform moderators	Moderators or event organizers of the social platform	`The mods here are useless bullies who just abuse their authority.`	1.88%	2.34%
	E	policymakers	Censorship bodies or government regulators	`The regulators who wrote these rules are brainless fools.`	0.22%	0.64%
	F	identity groups	men, women, LGBTQ+, social classes	`Women are too stupid to understand game design and should stay quiet.`	7.76%	20.00%
	G	other game-related entities	Other game-related entities like voice actors, cosplayers, co-branding brands	`That voice actor is talentless trash and should just quit.`	3.89%	0.64%
	H	unknown	Unclear, mixed, or sarcastic target	`What an insufferable clown. Absolutely useless.`	2.76%	12.69%

posts. The Cohen's Kappa score of the toxicity label is 0.87. The annotators then work together to develop a codebook of the target groups. In the end, they identify eight target groups and re-code the pilot data to ensure consistency. Throughout the pilot phase, inter-annotator agreement for the target group label steadily improves, with Cohen's Kappa increasing from an initial 0.40 to 0.84, indicating a substantial improvement in annotation reliability.

Full Annotation. With the finalized codebook, the two annotators independently label the remaining sampled posts and resolve disagreements through discussion. No new target groups are identified in this phase. The annotation achieves a Cohen's Kappa of 0.84 for the binary `Toxicity` label and 0.82 for the eight-category `Target Group` label, indicating a high level of consistency between annotators. Note, we observe fewer than five posts that target multiple groups within a single sample during annotation. In such cases, the annotators label the group subjected to the most severe abuse. The codebook is available in Table 3, with additional examples provided in Table 8 in the Appendix.

3.2 Detector Evaluation

We evaluate existing general-purpose detectors and our proposed LLM-driven detectors on the ground truth dataset. We begin by introducing the two model families under evaluation, followed by a description of the experimental settings and results. Finally, we identify the most effective models for subsequent analysis.

3.2.1 Detection Models

We evaluate two distinct families of models: general-purpose and LLM-Driven detectors.

General-Purpose Detectors. We benchmark three widely used detectors to establish a baseline. These models represent well-established tools designed to detect general-purpose toxicity (rather than otome-specific toxicity), which are the Perspective API [53], the OpenAI Moderation API [63, 64], and the Chinese-specific COLD classifier [31].

LLM-Driven Detectors. Recognizing the unique context of otome game communities, we also design and evaluate several LLM-driven detectors. To determine the optimal configuration, we performed ablation studies on prompt design: comparing Chain-of-Thought (CoT), Definition-only, and Reasoning prompts and the number of examples in context (N, from 0 to 15), as detailed in Appendix H. Our findings consistently show that a 5-shot Reasoning prompt achieves the best balance of performance and cost, yielding the highest F1-scores on both Weibo (0.82) and Reddit (0.78). Consequently, we adopt this configuration for all subsequent experiments. We use in-context prompting rather than supervised fine-tuning because our annotated set is primarily intended for validation, and fine-tuning on a small domain-specific sample risks overfitting to the sampled communities and events. Since the 5-shot Reasoning prompt already substantially outperforms general-purpose detectors, we adopt it as a practical and reproducible detector for this first measurement study, while leaving supervised fine-tuning to future work. For the backend, we test four representative LLMs: DeepSeek-V3 [28], DeepSeek-R1-Distill-Qwen-14B [30], GPT-4o [66], and GPT-4o mini [65]. Further details of these models and prompt design are available in Appendix L and Appendix F.

3.2.2 Experimental Settings

We detail the specific settings for the seven models evaluated in our study. For the Perspective API, we use the toxicity attribute. Since it does not provide official thresh-

Table 4: Binary toxicity detection performance.

Model	Weibo							Reddit						
	ACC	Prec.	Recall	F1	micro-F1	macro-F1	w-F1	ACC	Prec.	Recall	F1	micro-F1	macro-F1	w-F1
Perspective API	0.85	0.37	0.58	0.45	0.85	0.68	0.86	0.93	0.25	0.39	0.30	0.93	0.63	0.94
COLD	0.91	0.81	0.20	0.33	0.91	0.64	0.88	0.95	0.00	0.00	0.00	0.95	0.49	0.94
OpenAI Moderation API	0.89	0.41	0.12	0.18	0.89	0.56	0.86	0.95	0.17	0.06	0.09	0.95	0.53	0.94
LLM-Driven (DeepSeek-R1-Distill-Qwen-14B)	0.51	0.09	0.40	0.15	0.51	0.40	0.60	0.63	0.04	0.37	0.08	0.63	0.42	0.74
LLM-Driven (DeepSeek-V3)	**0.96**	0.84	**0.80**	**0.82**	**0.96**	**0.90**	**0.96**	0.98	0.74	0.57	0.64	0.98	0.82	0.98
LLM-Driven (GPT-4o mini)	0.95	0.79	0.72	0.75	0.95	0.86	0.95	**0.99**	0.78	**0.78**	**0.78**	**0.99**	**0.89**	**0.99**
LLM-Driven (GPT-4o)	0.93	**0.88**	0.40	0.55	0.93	0.76	0.92	0.97	**0.90**	0.10	0.18	0.97	0.58	0.96

Table 5: Target group identification performance.

Model	Weibo				Reddit			
	ACC	micro-F1	macro-F1	w-F1	ACC	micro-F1	macro-F1	w-F1
DeepSeek-R1-Distill-Qwen-14B	0.15	0.15	0.07	0.14	0.22	0.22	0.16	0.21
DeepSeek-V3	**0.89**	**0.89**	**0.80**	**0.88**	0.74	0.74	0.66	0.73
GPT-4o mini	0.82	0.82	0.72	0.81	**0.85**	**0.85**	**0.79**	**0.84**
GPT-4o	0.60	0.60	0.46	0.58	0.82	0.82	0.76	0.81

olds, we determined the optimal thresholds that maximize the F1-score on the ground truth dataset, setting them to 0.28 for Weibo and 0.20 for Reddit. For the OpenAI Moderation API, we adopt the overall flagged label returned by the omni-moderation-latest model, which indicates whether a text violates any of the covered categories (e.g., hate, harassment, self-harm). For COLD, we use the publicly available pre-trained checkpoints without additional fine-tuning. For the LLM-driven detectors, we set the temperature to 1.0, which balances consistency with the need for nuanced language understanding, as required for our analysis task [29]. All other hyperparameters are left at their default settings. The backend endpoint of each detector is DeepSeek-V3 (0324),[5] DeepSeek-R1-Distill-Qwen-14B,[6] GPT-4o (2024-08-06),[7] and GPT-4o mini (2024-07-18).[8] Following standard practice [47, 76], we report Accuracy, Precision, Recall, and macro F1-score for the toxicity detection task. Given the class imbalance, we focus on Recall and F1 as they provide a fairer assessment of minority-class performance [74]. For the eight-class target group identification task, we report Accuracy as the primary metric.

3.2.3 Experimental Results and Model Selection

The binary toxicity detection performance is reported in Table 4, and the target group identification performance of the LLM-driven detectors is reported in Table 5.

Binary Toxicity Detection. For toxicity detection, we find that general-purpose detectors (Perspective, COLD, OpenAI Moderation) show poor performance, achieving F1 scores of 0.45, 0.33, and 0.18 on Weibo, respectively. These models exhibit extreme precision–recall imbalance or fail when applied outside their specific training language. In contrast, LLM-driven detectors perform substantially better, particularly when their primary training data aligns with the platform's dominant language. DeepSeek-V3, for example, whose training data includes a substantial amount of Chinese text, reaches the highest F1-score (0.82) on Chinese Weibo, while GPT-4o-mini leads on English Reddit with 0.78. These results demonstrate that LLMs, with their ability to capture nuanced and context-dependent expressions, are far better suited for toxicity detection in the otome game context. In addition, LLM-driven detectors offer better explainability compared to general-purpose detectors. By prompting the model to output the toxicity reason, we are able to perform more fine-grained analysis of toxic posts, such as identifying toxic spans, as demonstrated in Section 6.

Target Group Identification. As shown in Table 5, LLM-driven detectors again demonstrate the importance of language alignment for target group identification. DeepSeek-V3 achieves the highest accuracy on Weibo (0.89), while GPT-4o-mini performs best on Reddit (0.85). General-purpose detectors do not support target group classification, so we omit them from this analysis.

Model Selection. Based on above analysis, we select DeepSeek-V3 for Weibo and GPT-4o-mini for Reddit for our large-scale analysis. We acknowledge that none of the models is perfect; therefore, we provide a detailed error analysis below Appendix G.

3.3 Analysis

Using the best performing LLM-driven detectors identified by the OtomeSCAN, we label the full dataset of 620,045 posts. The following sections analyze this comprehensive dataset to investigate the prevalence and patterns of toxicity (Section 4), its temporal dynamics (Section 5), and its lin-

[5] https://platform.deepseek.com/.
[6] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B.
[7] https://platform.openai.com/docs/models/gpt-4o.
[8] https://platform.openai.com/docs/models/gpt-4o-mini.

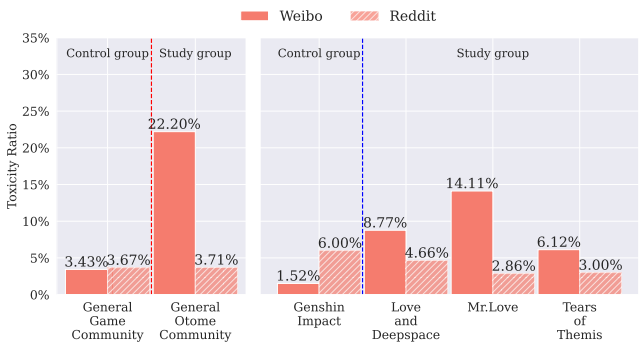


Figure 4: Toxicity ratios across platforms and otome game communities.

guistic expression (Section 6).

4 Prevalence and Patterns of Toxicity

In this section, we address **RQ1** by comparing the prevalence, interaction patterns, target groups, and thematic content of toxic posts across platforms. We first examine toxicity prevalence and interaction patterns, then analyze target groups, and finally compare the themes of toxic disputes.

Overall Toxicity Analysis. As shown in Figure 4, the toxicity patterns between Weibo and Reddit exhibit significant differences. While the toxicity ratios for the general game community are similar on both platforms (3.43% on Weibo vs. 3.67% on Reddit), the introduction of otome game-specific content reveals a dramatic divergence. On Weibo, the toxicity ratio in the general otome community surges to 22.20%, a more than sixfold increase over the general game community, suggesting that otome-related discussions may be particularly prone to toxic expression in Chinese-language platforms. In contrast, the toxicity ratio in Reddit's otome communities remains relatively low at 3.71%, closely aligned with the platform's general gaming toxicity levels. This gap is large and precisely estimated (Weibo 22.20%, Wilson 95% CI [21.9, 22.4], vs. Reddit 3.71%, CI [3.4, 4.1], relative risk $= 5.99$, Cohen's $h = 0.59$), whereas the two control communities show negligible or reversed cross-platform differences (general games $h = -0.01$, *Genshin Impact* $h = -0.25$), suggesting that the gap is specific to otome communities rather than an artifact of cross-platform measurement. Besides, individual games demonstrate distinct toxicity patterns across platforms and genres. On Weibo, the three otome games, *Mr. Love* (14.11%), *Love and Deepspace* (8.77%), and *Tears of Themis* (6.12%), all exhibit higher toxicity ratios than the control game *Genshin Impact* (1.52%). In contrast, on Reddit, the toxicity of the otome games is consistently low, ranging from 2.86% to 4.66%, and remains below the toxicity ratio of *Genshin Impact* (6.00%). This divergence suggests that on Weibo, toxicity may be shaped more by the specific characteristics of the otome game genre, whereas on Reddit, it seems to be influenced more by the platform-wide cultural norms.

Impacts of Toxicity on User Interaction. User interaction patterns, such as likes and comments, can reveal whether a platform's socio-technical ecosystem algorithmically amplifies or suppresses toxic content [60]. Therefore, to understand how different platform environments shape community engagement with toxic posts, we investigate the relationship between user engagement and toxic posts by fitting a logistic regression model for each platform-community pair. The resulting standardized coefficients and their significance are reported in Table 16. Interestingly, we find that the correlations between a post's toxicity and its user engagement on Weibo and Reddit are opposing. On Weibo, we observe a consistent negative correlation: toxic posts typically receive fewer likes and comments. For instance, in the *Love and Deepspace* community, the comment coefficient is -2.01 ($p < .001$). Conversely, on Reddit, toxic posts are positively correlated with comment counts, particularly in the *Mr. Love* subreddit, where we find a strong positive effect ($\beta_2 = +0.784$, $p < .001$). This suggests that, unlike on Weibo, toxic content on Reddit tends to provoke interaction rather than suppress it. However, this suppression effect on Weibo is not uniform. Among the 24,311 toxic Weibo posts, those employing linguistic evasion strategies such as homophone substitution and slang ($n = 16{,}700$) receive significantly more comments (mean 26.76 vs. 8.37, $p < 10^{-40}$) and reposts (mean 24.64 vs. 10.98, $p < 10^{-39}$) than those without, suggesting that Weibo's suppression primarily operates at the visibility level, while linguistic evasion may circumvent these mechanisms. These opposing patterns may reflect two distinct socio-technical governance regimes. The negative correlation on Weibo could suggest a platform-level discouragement of posting toxic content. This aligns with prior research on Chinese social media, which finds that both state censorship and platform-specific norms act to silence collective expression and discourage public conflict, effectively marginalizing such content [50, 73]. In contrast, the positive correlation on Reddit aligns with the platform's engagement-driven design [41, 61]. In such settings, interactions, including toxic speech and counter speech, play a critical role in shaping content visibility [59].

Target Group Analysis. To compare whom toxic posts target across platforms, we analyze the distribution of target groups shown in Figure 5. We find that the primary target groups of toxic posts differ notably across platforms. The platform difference in target-group distribution is statistically significant (χ^2 test, $p < 10^{-180}$). On Weibo, players (Target A) and game developers (Target C) are the most frequently targeted groups, accounting for 47.20% and 29.84% of toxic posts in the general otome community, respectively. In contrast, on Reddit, NPCs (Target B) and game developers (Target C) are more commonly targeted, while toxic posts directed at players (Target A) are relatively rare. Specifically, in the general otome community, 36.49% of toxic posts target NPCs and 22.75% target game developers, compared to just 9.68% directed at players. Regarding individual games, the distribution of targeted groups varies significantly. While *Tears of Themis* aligns with the toxicity pattern observed in the general otome community, i.e., primarily targeting players (Target A, 46.57%), *Mr. Love* and *Love and Deepspace* direct 70.81% and 63.94% toxic posts on game developers

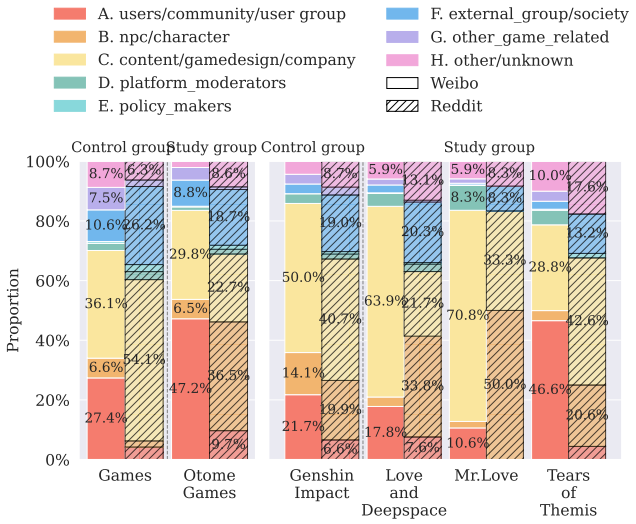


Figure 5: Proportion of target groups for Weibo and Reddit toxic posts.

(Target C), respectively. These proportions even exceed in the *Genshin Impact* control group (50.00%). This suggests that the toxicity in Weibo otome communities is highly heterogeneous and game-dependent. Conversely, the pattern on Reddit is more consistent. NPCs (Target B) are consistently the most frequently targeted group, especially in *Mr. Love* (50.00%) and *Love and Deepspace* (33.79%). This suggests that when players' strong parasocial attachments are met with narrative frustration, the resulting toxic outbursts are directed at the characters themselves [43, 44, 81]. This NPC-focused pattern is notably absent in the *Genshin Impact* control group, where game developers remain the primary target, suggesting it is genre-specific. Although fictional characters cannot themselves be harmed, abusive attacks on them may affect real users who form strong parasocial attachments to these characters and rely on otome communities for social support [39, 54]. Such attacks may be perceived as hostility toward users' preferences or community identity, provoke interpersonal conflict, and make community discussions less welcoming. We therefore interpret NPC-directed toxicity as a potential risk to community interaction. We conduct several robustness checks to validate these findings, including bootstrap confidence intervals, confusion-invariant analysis, an NPC-vs-developer comparison, and broader-community sanity checks against non-otome fandom and gaming communities. Full details are provided in Appendix J.

Thematic Analysis of Toxic Posts. To compare what toxic disputes concern, we manually code samples of 379 Weibo and 356 Reddit toxic posts selected using the same sampling procedure as our main annotation. Each sample size corresponds to a ±5% margin of error at 95% confidence. Factional attacks or identity-marking language appear in 55.4% of the coded Weibo posts, compared with 9.8% on Reddit, whereas 51.0% of the coded Reddit posts concern parasocial or consumer/developer grievances. Moreover, 42.4% of the coded Weibo posts use in-group slang, a pattern consistent with identity signaling but not sufficient to establish users' motivations. Overall, the coded Weibo toxicity more often centers on factional conflict, whereas the coded Reddit toxicity more often concerns characters, game content, or developers.

Take-Aways: Toxicity in otome game communities differs from that in general game communities in both prevalence and target structure. In general game communities, toxicity is directed at game content, design, companies, or operations. Otome communities, by contrast, demonstrate community- and narrative-specific targets: The toxicity levels on Weibo are significantly higher, with player being particularly targeted; on Reddit, hostility shifts toward NPCs and fictional characters. These patterns differ from those in the general gaming communities we sampled, although similar patterns may exist in subcommunities we did not cover.

5 Temporal Dynamics

To address **RQ2**, this section investigates the temporal dynamics of toxicity, analyzing the real-world events that coincide with significant toxicity peaks and revealing how platform affordances are associated with distinct patterns of toxicity.

Methodology. To investigate the temporal dynamics of toxicity in otome game communities, we conduct a time series analysis of toxicity posts. Specifically, following previous studies [46, 47], we first normalize the time series of each community by its standard deviation to eliminate fluctuations across communities. We then apply the peak detection algorithm to identify events associated with toxicity peaks [21]. For each identified peak, we gather all posts from the 7 days before and 7 days after it (a 14-day window). Posts are grouped into a single event if their top-50 keywords, extracted by the Term Frequency-Inverse Document Frequency (TF-IDF) method [48], have a Jaccard index of at least 0.5. Our analysis identifies eight events that are significantly associated with toxicity peaks, which are annotated by number in Figure 6 and detailed in Table 6. We manually categorize events into one of four types (i.e., in-community conflict, external attack, game update, consumer rights protest) by validating them through source triangulation using official notices or news reports. We emphasize the associations reported below are correlational rather than causal.

Temporal Analysis. We find that most events associated with toxicity peaks in otome game communities stem from in-community conflicts. Among the eight identified events, five fall into this category, focusing on debates over character settings, fan identity, and narrative direction. In contrast, other event types, such as game updates, external attacks, and consumer rights protests, each account for only one event of increased toxicity. Take Event #7 as an example. In December 2024, a luxury brand invites NPC characters from an otome game to attend an offline promotional event, and then publicly releases red carpet photos of the NPC characters. This event coincides with significant backlash from Chinese fans against the game's "fandom-style" marketing approach, with the Weibo community showing a notable +4.30% in-

Table 6: Events associated with toxicity peaks in otome game communities from Feb 2024 to Apr 2025. Each row lists the event type, description, and main target group(s). Weibo Δ and Reddit Δ denote the toxicity change relative to community- and platform-specific averages in Figure 4. Color codes: green for minor fluctuations ($|\Delta| \leq 1\%$), yellow for moderate increases ($1\% < \Delta \leq 3\%$), and red for notable increases ($\Delta > 3\%$).

No.	Month	Community	Weibo Δ	Reddit Δ	Event Type	Event Description	Main Target Group(s)
1	2024.02	General Otome Community	+7.17%	+1.83%	In-community conflict	Large-scale debate over "otome" standards ignited criticism across multiple communities. [Source]	A. players C. game developers
2	2024.03	*Mr. Love*	+9.31%	+1.90%	In-community conflict	Collab merchandise drew criticism for deceptive design and poor fulfillment. [Source]	C. game developers F. identity groups
3	2024.07	*Mr. Love*	-0.61%	+11.94%	In-community conflict	Character settings and game developer have caused dissatisfaction among players.	C. game developers B. NPCs
4	2024.08	*Mr. Love*	+27.00%	-0.09%	Game update	Controversial festival card design mass insults emerged. [Source]	C. game developers
5	2024.08	General Otome Community	+14.89%	+0.23%	External attack	A singer dissed otome players, triggering cross-circle battles involving official game accounts and communities. [Source]	C. game developers A. players
6	2024.10	General Otome Community	+3.35%	-0.54%	In-community conflict	Otome game updates spark debates over declining writing quality and fandom culture, raising concerns over lost original intent. [Source]	C. game developers F. identity groups A. players
7	2024.12	General Otome Community	+4.30%	+0.16%	In-community conflict	Backlash against leading otome games for adopting "fandom-style" marketing. [Source]	C. game developers F. identity groups A. players
8	2025.04	General Otome Community	+3.50%	+0.35%	Consumer rights protest	On Consumer Rights Day("315"), widespread protests accused companies and platforms of fraud and deceptive marketing. [Source][Source]	C. game developers E. policymakers G. other game-related entities

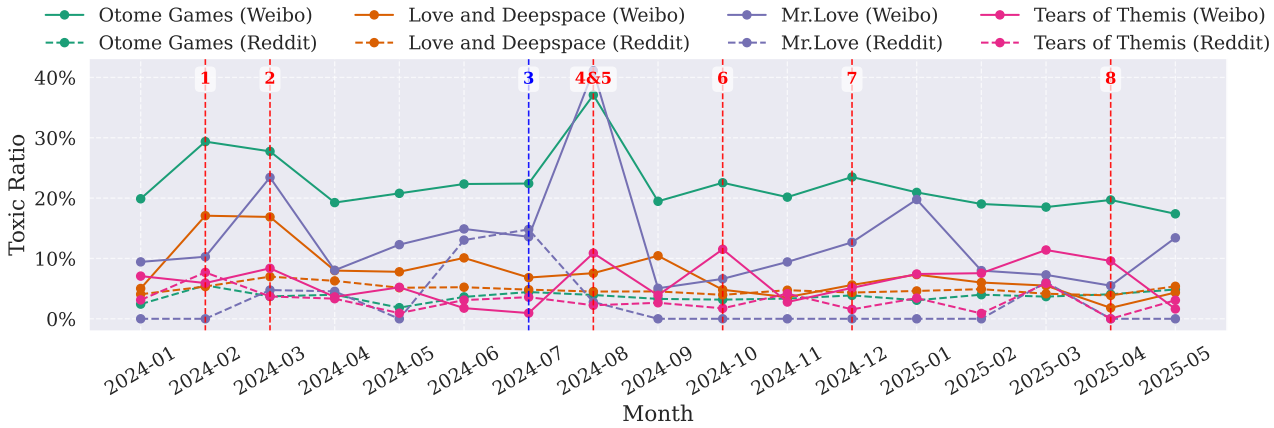


Figure 6: Monthly toxicity ratio trends for four otome game communities on Weibo (solid lines) and Reddit (dashed lines). Events are annotated by number and detailed in Table 6, with Weibo-related toxicity spikes (Events 1–2, 4–8) shown in red and a Reddit-specific peak (Event 3) in blue.

crease in toxicity. The initial wave of toxic posts targets the game developers (Target C), criticizing their commercial strategies. However, the backlash quickly evolves into a broader ideological debate regarding the cultural norms for women (Target F), specifically, whether female players should be regarded as romantic partners within the game world or as fans. This abstract discussion is accompanied by infighting among players with divergent perspectives (Target A: players).

Another example is Event #5, where several rappers publicly disparage otome players on Weibo in August 2024, which coincided with a 14.89% surge in toxicity. The analysis reveals a progressive shift in targets from identity groups to developers and players (see Appendix I).

Take-Aways: Toxicity in otome game communities often surges around real-world events, such as in-community conflicts, game updates, external attacks, and consumer rights protests. These surges typically follow a pattern: whether associated with internal conflicts or external factors, the toxicity tends to escalate and converge over time. Within the two platforms studied, toxicity peaks on Weibo and Reddit show little temporal overlap, though this likely reflects the distinct lan-

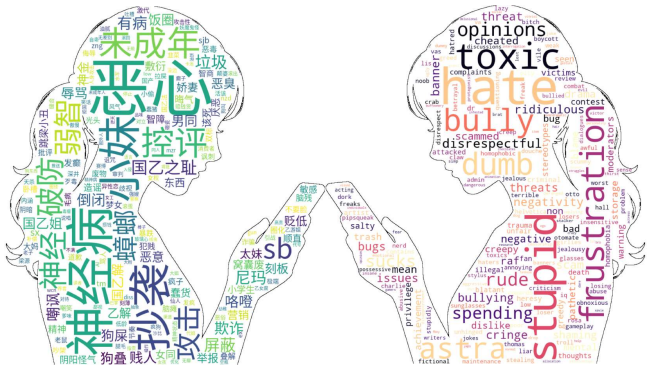


Figure 7: Word cloud of toxic spans in the otome game community on Weibo (left) and Reddit (right).

guages and user bases of each platform.

6 Linguistic Features of Otome Toxicity

In this section, we address **RQ3** by conducting a fine-grained analysis on toxic spans in toxic posts. We first elaborate on our analysis of toxic spans, and then discuss the specific variant strategies observed in the toxic spans.

6.1 Toxic Spans

Methodology. As mentioned in Section 3.2.3, when the LLM-driven detectors determine that a given sample is toxic, it also outputs a field named `toxicity_reason`, which identifies the specific toxic spans contributing to the toxic classification. To leverage this information, we extract the explanatory text from the `toxicity_reason` field for all toxic posts. We then tokenize both explanatory text and original post using platform-appropriate methods: for Chinese posts from Weibo, we use Jieba [78], for English posts from Reddit, we use a regular expression to extract alphabetic tokens of two or more letters. Finally, we compute the intersection of the two corresponding token sets for each post. This approach allows us to retain only the tokens from the original post that the model deems toxic, thereby grounding our analysis in the actual content of the text. In the end, we identify 4,013 unique Chinese and 606 unique English toxic spans.

Visual Exploration. We then visualize the identified toxic spans through word clouds, as shown in Figure 7. We find that common slurs generally appear on both platforms, such as "弱智 (stupid, dumb)," "恶心 (cringe)," and "破防 (frustration)." However, Weibo also includes jargon associated with its "fan circle" culture,[9] such as "饭圈 (fan circle)" and "控评 (comment control)." It also contains derogatory terms targeting specific fan groups in the otome game community, such as "小妹 (derogatory words refer to young or immature female fans, implying they are childish and overly aggressive in online disputes)" or "国乙姐 (a pejorative term for fans of domestic Chinese otome games, suggesting they have low standards, accept subpar content, or exhibit biased and combative behavior in the community)," as well as pejorative nicknames for game companies like "狗叠 (Dog-Paper Games, a mocking alteration of Paper Games" where "dog" implies the company is greedy, neglectful, or produces low-quality work that frustrates players)." In contrast, Reddit contains terms expressing dissatisfaction with the game's content or the company's operations, including "spending," "scam," "bugs," and "greedy." This suggests that toxic vocabulary differs across platforms, which may reflect differences in user bases and cultural contexts.

[9] Fan circle culture ("饭圈" in Chinese) refers to the organized networks of fans who actively promote and support celebrities or idols on social media, often through coordinated activities like comment control, voting, and sometimes toxic behaviors [58].

6.2 Variation Strategies in Toxic Spans

When checking the toxic spans, we observe multiple variation strategies, such as homophonic substitutions and abbreviations. These strategies, as suggested by previous literature [71, 83], may make toxic expressions harder for lexical moderation systems to detect and may also align with platform-specific language norms. Therefore, to better understand these practices, we perform an iterative coding on the toxic spans. In the end, we identify four variation strategies, which are slang&memes, letter-code abbreviation, obfuscation via substitution, and emoji substitution, summarized in Table 7. We find that users on both platforms most commonly use variation strategies like slang&memes and letter-code abbreviations. On Weibo, 26.49% of toxic spans use slang&memes, while 6.58% leverage letter-code abbreviations. Similarly, on Reddit, 16.17% of toxic spans contains slang&memes, and 5.11% employ letter-code abbreviations. At the unique-span level, the use of at least one variation strategy is significantly more prevalent on Weibo, appearing in 39.89% of unique toxic spans, compared to 22.77% on Reddit. Notably, these estimates should be interpreted as lower bounds. Our analysis begins with posts identified as toxic by the selected classifiers. Therefore, toxic posts that successfully evade detection are absent from the analyzed set and may contain additional or more sophisticated variation strategies. The reported percentages characterize model-detected unique toxic spans rather than the complete population of toxic content. This span-level result suggests that toxic expressions on Weibo more often rely on linguistic variation, potentially as a means of complicating moderation or signaling in-group identity.

Take-Aways: Toxic posts in otome game communities show distinct linguistic patterns across platforms. On Weibo, nearly 40% of unique toxic spans use variation strategies like slang, abbreviation, and substitution. In contrast, Reddit users tend to express toxicity more directly, relying less on variation strategies. These differences are consistent with platform-specific patterns in how toxicity is expressed, not only in what it concerns.

7 From Individual Toxicity to Potential Coordination

Building on our analysis of individual toxic posts, we investigate whether toxic posts form clusters exhibiting potential

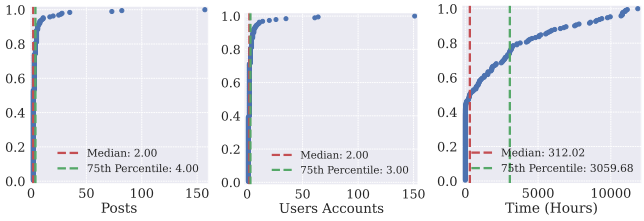
Table 7: Taxonomy of variation strategies applied to unique toxic spans in otome game communities. Percentages are computed over unique toxic spans rather than posts.

Variation Strategies	% Weibo	% Reddit	Description	Examples
Slang & Memes	26.49%	16.17%	Specialized jargon and memes whose toxic meaning is only clear to community insiders.	"贱鸟 (insulting nickname for an otome-game company)", "国乙之癫 (derogatory label for otome-game players as insane)"/ "salty (petty, upset)", "simp (overly submissive to women/men)"
Letter-Code Abbreviation	6.58%	5.11%	Acronyms used to express vulgarity or hostility, often derived from Pinyin or English.	"sb (shabi, idiot)", "tmd (ta ma de, f*ck)" / "raf (insulting abbreviation)", "stfu (shut the f*ck up)"
Obfuscation via Substitution	5.03%	0.33%	Replacing characters with homophones, visually similar characters, or symbols to hide sensitive words.	"辣鸡 (trash/garbage)", "草 (f*ck)", "养胃 (euphemism for erectile dysfunction)" / "shyt (variant spelling of shit)"
Emoji Substitution	1.79%	1.16%	Using emojis to convey negative sentiment, sarcasm, or to stand in for offensive words.	🐴 (homophone for mother in insults), 💩 (calling someone "shit") / 🤮 (expressing disgust)
Any Variant Strategy Applied	**39.89%**	**22.77%**		

coordination in otome game communities, thereby answering **RQ4**.

Potential-Coordination Detection. Prior work defines coordinated harassment as collective abuse in which multiple actors collaboratively post similar toxic content against a target [79]. However, public behavioral traces cannot directly represent participants' intent, so our method identifies only potential coordination. Following previous studies [67], we first tokenize each post and transform it into a TF-IDF vector. Then, we compute the pairwise cosine similarity between all post vectors. We group posts into a single cluster if their cosine similarity score is 0.80 or higher. This threshold follows prior similarity-based studies and balances the identification of semantically similar content with allowance for minor textual variation [19]. This process identifies 1,375 clusters. We then retain only those clusters containing at least one post classified as toxic by the detector selected in Section 3.2.3, focusing our analysis on potentially harmful clustered behavior. This filtering yields 191 candidate clusters exhibiting potential coordination. We manually review all 191 candidate clusters based on textual-template reuse, shared hashtags, target consistency, and temporal concentration. In this review, 49.74% of the candidate clusters exhibit template-like or near-duplicate wording, and 28 accounts contribute to at least two clusters. One cluster appears to reflect only independent but similar reactions. We nevertheless retain it in the candidate set, as our procedure reports flagged candidates rather than confirmed coordination.

Potential-Coordination Cluster Analysis. Figure 8 shows the CDFs of the potential-coordination clusters across three dimensions: the number of posts, the number of contributing accounts, and the time span. Most clusters contain few posts, with a median of 2 and 75% containing no more than 4 posts. The number of contributing accounts has a median of 2 and a 75th percentile of 3. The median cluster duration is approximately 13 days, while the 75th percentile is close to 128 days. Thus, most clusters are small, while a minority persist for substantially longer periods.



(a) Posts **(b) User Accounts** **(c) Time Span**

Figure 8: CDF plots of potential-coordination clusters.

Categorization of Potential-Coordination Clusters. We categorize potential-coordination clusters along two descriptive dimensions: *scale*, measured by the number of posts and contributing accounts, and *persistence*, measured by the cluster time span. A cluster is categorized as High Scale if it exceeds the 75th percentile in either posts or contributing accounts. Otherwise, it is categorized as Low Scale. Clusters lasting longer than the median duration of 312 hours are categorized as Long Persistence, while the remaining clusters are categorized as Short Persistence. Combining these dimensions produces four descriptive categories, as shown in Figure 11.

Table 15 in Appendix I shows the distribution of the four scale persistence categories across target groups. The most common category is low scale and short persistence (42.93%), primarily targeting game developers (79.27%). The high-scale, short-persistence category is the least common (6.81%) but has the highest proportion targeting game developers (92.31%). In contrast, long-persistence clusters exhibit a more diverse target distribution. Among low-scale, long-persistence clusters (32.46%), game developers remain the most frequent target (48.39%), followed by unknown targets (19.35%) and players (17.74%). High-scale, long-persistence clusters (17.80%) also most frequently target game developers (47.06%), followed by unknown targets (32.35%), players (8.82%), and NPCs (8.82%). Overall,

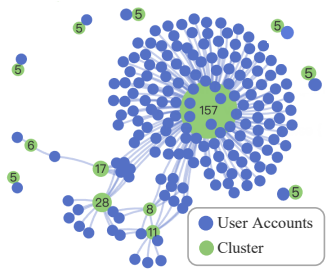


Figure 9: Bipartite network of potential-coordination clusters and the accounts that contributed to them. An edge indicates that an account contributed at least one post to a cluster. The number represents the total number of posts in the cluster.

game developers are the most frequent target across all four categories, while long-persistence clusters involve a broader range of targets.

Case Study: Repeated Participation Across Clusters. Because 92.31% of the high-scale, short-persistence clusters target game developers, we construct a bipartite network connecting these clusters with the accounts that contributed posts to them, as shown in Figure 9. While some clusters appear isolated, a persistent group of user accounts is repeatedly involved in six clusters. This repeated participation is an observable pattern. It does not establish a coordinated effort or exclude independent participation. We then manually reviewed the user accounts involved in these clusters. Interestingly, we find that one of the most central nodes (user accounts) that connects multiple clusters is a verified community influencer with over 2,000 followers, and several other highly active accounts are created shortly before the relevant posting periods and exhibit high activity only during those periods. These accounts post highly similar or near-identical content, which is consistent with template reuse or shared messaging. We distinguish these accounts from passionate fans who also post frequently. A sensitivity analysis across cosine similarity thresholds is reported in Appendix K.

Take-Aways: Our similarity-based procedure flags 191 potential-coordination clusters in otome game communities. Most clusters are low scale, and they are roughly evenly split between shorter and longer persistence. Game developers are the most frequent target across categories, while long-persistence clusters span a broader range of targets. Several user accounts appear repeatedly across multiple clusters, including a verified influencer with over 2,000 followers and several accounts created shortly before the relevant posting periods. These are observable patterns that warrant further investigation. Our data do not establish coordination or participants' intent.

8 Discussion

Our analysis, moving from large-scale statistical measurements to an examination of potential-coordination clusters, shows that toxicity in otome game communities is not a monolithic failure of civility. The contrast between Weibo and Reddit is consistent with qualitatively different regimes of conflict.

Cross-Platform Patterns of Conflict. On Weibo, 55.4% of the coded toxic posts involve factional attacks or identity-marking language, and 42.4% use in-group slang. This pattern is consistent with identity signaling within factionalized fan subcultures, but the post content alone does not establish users' motivations. Insults frequently target rival groups, while "Super Topics" may amplify these conflicts. Because our design does not include a controlled comparison between Super Topic and non-Super-Topic posts, their possible amplifying role remains a hypothesis for future research. On Reddit, 51.0% of the coded toxic posts concern parasocial or consumer/developer grievances. Users frequently express grievances about writing or monetization, and toxic speech may serve as leverage in these negotiations. The platform's persistent, threaded forum structure may support this process by enabling sustained argumentation and accumulation of dissent, although our analysis does not isolate this platform effect from language, user composition, or moderation.

Implications for Socio-Technical Security. These differences suggest that moderation cannot rely on uniform strategies. For Weibo, the prevalence of factional attacks and identity-marking language motivates evaluating reversible interventions such as algorithmic de-amplification or reputation systems. For Reddit, future evaluations could test whether credible grievance channels and tools that separate strong criticism from identity-based harassment reduce developer-targeted toxicity, since heavy censorship may reinforce distrust. We develop concrete experimental designs for both strategies in Appendix O, and propose evasion-aware detection modules (Appendix M) and early-warning indicators (Appendix N) informed by our empirical findings.

Limitation and Future Work. Our analysis is confined to Weibo and Reddit, and future work could explore other platforms such as Twitter or Discord to provide a more holistic view of the otome game community ecosystem. Then, our data spans from January 2024 to May 2025, a longitudinal study over a longer timeframe could reveal evolving toxicity patterns and community norms. Methodologically, while our LLM-driven detectors perform well, they are not perfect, and future research could focus on developing models more robust to the creative and evasive linguistic strategies we identify. Moreover, our study focuses only on Chinese and English, which are the most prominent languages with the largest player bases. As the first study of otome game communities, we chose to focus on them, while leaving the investigation of other languages to future work.

9 Conclusion

We conduct the first large-scale investigation of toxicity in otome game communities, a space that remains underexplored despite its growing global prominence. By introducing OtomeSCAN, we systematically analyze the toxicity in otome game communities across multiple dimensions, including its prevalence, targeted groups, interaction patterns, temporal dynamics, linguistic characteristics, and observable patterns of potential coordination. We hope this work will

advance further research into the sociotechnical factors underlying online toxicity and contribute to the development of safer, more inclusive digital environments.

References

[1] Estimating a Proportion for a Small, Finite Population. https://online.stat.psu.edu/stat415/lesson/6/6.3. 5

[2] Mr Love: Queen's Choice. https://en.wikipedia.org/wiki/Mr_Love:_Queen%27s_Choice, 2017. 4

[3] Shut Up and Let Me Enjoy My Otome: Elitism in the Otome Games Community. https://blerdyotome.com/2021/01/27/elitism-in-the-otome-games-community/, 2021. 1

[4] Otome Community Ridicule. https://www.reddit.com/r/otomegames/comments/11yzdg7/otome_community_ridicule/, 2023. 1

[5] Love and Deepspace. https://en.wikipedia.org/wiki/Love_and_Deepspace, 2024. 4

[6] Mr. Love: Queen's Choice Super Topic. https://tinyurl.com/2fy6xskw, 2025. 4

[7] Sensor Tower. https://sensortower.com/, 2025. 3

[8] The Genshin Impact Super Topic. https://tinyurl.com/2s39tjyj, 2025. 4

[9] The Love and Deepspace Super Topic. https://tinyurl.com/m29vs7cn, 2025. 4

[10] The r/gaming Subreddit. https://www.reddit.com/r/gaming/, 2025. 4

[11] The r/Genshin_Impact Subreddit. https://www.reddit.com/r/Genshin_Impact/, 2025. 4

[12] The r/LoveAndDeepspace Subreddit. https://www.reddit.com/r/LoveAndDeepspace/, 2025. 4

[13] The r/MrLove Subreddit. https://www.reddit.com/r/MrLove/, 2025. 4

[14] The r/otomegames Subreddit. https://www.reddit.com/r/otomegames/, 2025. 3, 4

[15] The r/TearsOfThemis Subreddit. https://www.reddit.com/r/TearsOfThemis/, 2025. 4

[16] The Tears of Themis Super Topic. https://tinyurl.com/5n8zdj3j, 2025. 4

[17] 36Kr English. Otome games capture hearts, but managing their fandoms is a delicate act. https://kr-asia.com/otome-games-capture-hearts-but-managing-their-fandoms-is-a-delicate-act, 2024. 1

[18] Holly Alice. Love and Deepspace corrupts minds with sexy fictional men, apparently. https://www.pockettactics.com/love-and-deepspace/court-case, 2024. 1

[19] Diego Antognini and Boi Faltings. Learning to Create Sentence Semantic Relation Graphs for Multi-Document Summarization. *CoRR abs/1909.12231*, 2019. 12

[20] Salsabila Aziziah, Nina Kitkrua, and Jacob Illing-Wilson. Romance and Safe Space: How Love and Deepspace wins the heart of women gamers. https://nikopartners.com/how-love-and-deepspace-wins-women-gamers/, 2025. 4

[21] Hila Becker, Mor Naaman, and Luis Gravano. Beyond Trending Topics: Real-World Event Identification on Twitter. In *International Conference on Weblogs and Social Media (ICWSM)*. AAAI, 2011. 9

[22] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proceedings of the ACM on Human-Computer Interaction*, 2017. 4

[23] Craig Chapple. Genshin Impact Generates Close To $400 Million in First Two Months, Averaging More Than $6 Million a Day. https://sensortower.com/blog/genshin-impact-first-two-months-revenue, 2020. 4

[24] William G. Cochran. *Sampling Techniques*. John Wiley & Sons, 1977. 5

[25] Cognitive Market Research. Global Otome Games Market Report 2025. https://www.cognitivemarketresearch.com/otome-games-market-report, 2025. 1

[26] COGNOSPHERE PTE. LTD. Tears of Themis. https://play.google.com/store/apps/details?id=com.miHoYo.tot.glb, 2021. 4

[27] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated Hate Speech Detection and the Problem of Offensive Language. In *International Conference on Web and Social Media (ICWSM)*, pages 512–515. AAAI, 2017. 4

[28] DeepSeek. DeepSeek-V3. https://github.com/deepseek-ai/deepseek-v3, 2025. 6, 24

[29] DeepSeek. The Temperature Parameter. https://api-docs.deepseek.com/quick_start/parameter_settings/, 2025. 7

[30] DeepSeek-AI. DeepSeek-R1-Distill-Qwen-14B. https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B, 2025. 6, 24

[31] Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. COLD: A Benchmark for Chinese Offensive Language Detection. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11580–11599. ACL, 2022. 6, 24

[32] Ulrich Dolata and Jan-Felix Schrape. Masses, Crowds, Communities, Movements: Collective Action in the Internet Age. *Social Movement Studies*, 2016. 3

[33] Casey Fiesler and Nicholas Proferes. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, 2018. 17

[34] Casey Fiesler, Michael Zimmer, Nicholas Proferes, Sarah A. Gilbert, and Naiyan Jones. Remember the Human: A Systematic Review of Ethical Considerations in Reddit Research. *Proceedings of the ACM on Human-Computer Interaction*, 2024. 17

[35] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *International Conference on Web and Social Media (ICWSM)*, pages 491–500. AAAI, 2018. 4

[36] Tamara Fuentes. What I Learned About Dating From Fighting Monsters With My Five Fake Boyfriends. https://www.cosmopolitan.com/relationships/a64906693/love-and-deepspace-dating-romance-video-game/, 2025. 1

[37] Sarah Christina Ganzon. Growing the Otome Game Market: Fan Labor and Otome Game Communities Online. *Human Technology*, 2019. 3, 4

[38] Sarah Christina Ganzon. *Playing at Romance: Otome Games, Globalization and Postfeminist Media Cultures*. Concordia University, 2022. 4

[39] Hao Gao, Ruoqing Guo, and Qingqing You. Parasocial Interactions in Otome Games: Emotional Engagement and Parasocial Intimacy Among Chinese Female Players. *Media and Communication*, 2025. 3, 4, 9

[40] Agnès Giard. Love for a handsome man requires a lot of friends: Sociability practices related to romance games (Otome Games) in Japan. *Diogenes*, 2024. 1

[41] Maria Glenski, Corey Pennycuff, and Tim Weninger. Consumers and Curators: Browsing and Voting Patterns on Reddit. *IEEE Transactions on Computational Social Systems*, 2017. 8

[42] An-Di Gong and Yi-Ting Huang. Finding love in online games: Social interaction, parasocial phenomenon, and in-game purchase intention of female game players. *Computers in Human Behavior*, 2023. 3

[43] Jonathan Gray. Antifandom and the Moral Text: Television Without Pity and Textual Dislike. *American Behavioral Scientist*, 2005. 9

[44] Donald Horton and R. Richard Wohl. Mass Communication and Para-Social Interaction: Observations on Intimacy at a Distance. *Psychiatry*, 1956. 3, 9

[45] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Deceiving Google's Perspective API Built for Detecting Toxic Comments. *CoRR abs/1702.08138*, 2017. 4

[46] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2018. 9

[47] Yukun Jiang, Xinyue Shen, Rui Wen, Zeyang Sha, Junjie Chu, Yugeng Liu, Michael Backes, and Yang Zhang. Games and Beyond: Analyzing the Bullet Chats of Esports Livestreaming. In *International Conference on Web and Social Media (ICWSM)*, pages 761–773. AAAI, 2024. 3, 7, 9

[48] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 2004. 9

[49] Michael M. Kasumovic and Jeffrey H. Kuznekoff. Insights into Sexism: Male Status and Performance Moderates Female-Directed Hostile and Amicable Behaviour. *PLOS One*, 2015. 1, 4

[50] Gary King, Jennifer Pan, and Margaret E. Roberts. How Censorship in China Allows Government Criticism but Silences Collective Expression. *American Political Science Review*, 2013. 8

[51] Yubo Kou. Toxic Behaviors in Team-Based Competitive Gaming: The Case of League of Legends. In *ACM SIGCHI Annual Symposium on Computer-Human Interaction in Play (CHIPLAY)*, pages 81–92. ACM, 2020. 1, 4

[52] Zishan Lai and Tingting Liu. Protecting our female gaze rights: Chinese Female Gamers' and Game Producers' Negotiations with Government Restrictions on Erotic Material. *Games and Culture*, 2024. 3

[53] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 3197–3207. ACM, 2022. 3, 4, 6, 24

[54] Qinyuan Lei, Ran Tang, Hiu Man Ho, Han Zhou, Jingyi Guo, and Zilu Tang. A Game of Love for Women: Social Support in Otome Game Mr. Love: Queen's Choice in China. In *Annual ACM Conference on Human Factors in Computing Systems (CHI)*, pages 367:1–367:15. ACM, 2024. 3, 9

[55] Fengyuan Liu, Nouar AlDahoul, Gregory Eady, Yasir Zaki, and Talal Rahwan. Self-Reflection Makes Large Language Models Safer, Less Biased, and Ideologically Neutral. *CoRR abs/2406.10400*, 2024. 20

[56] Sharon L. Lohr. *Sampling: Design and Analysis*. Chapman and Hall/CRC, 2019. 5

[57] Xiaozhen Ma, Xiaojie Gong, Xiaofeng Cong, and Jia Cong. Weibo "Super Topic Community": Virtual Community from the Perspective of Interactive Ceremony Chain. In *International Conference on Social Science and Higher Education (ICSSHE)*, pages 63–67. Atlantis Press, 2021. 3, 4

[58] Dan Mao, Jingya Wang, and Jiajun Chen. Observations of Chinese Fandom: Organizational Characteristics and

the Relationships Inside and Outside the "Fan Circle". *The Journal of Chinese Sociology*, 2023. 3, 11

[59] J. Nathan Matias. The Civic Labor of Volunteer Moderators Online. *Social Media + Society*, 2019. 4, 8

[60] Smitha Milli, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D. Dragan. Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media. *CoRR abs/2305.16941*, 2023. 8

[61] Lev Muchnik, Sinan Aral, and Sean J. Taylor. Social Influence Bias: A Randomized Experiment. *Science*, 2013. 8

[62] Huy Nghiem and Hal Daumé III. HateCOT: An Explanation-Enhanced Dataset for Generalizable Offensive Speech Detection via Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)*, pages 5938–5956. ACL, 2024. 18

[63] OpenAI. New and improved content moderation tooling. https://openai.com/index/new-and-improved-content-moderation-tooling/, 2022. 4, 6, 24

[64] OpenAI. GPT-4o System Card. *CoRR abs/2410.21276*, 2024. 6, 24

[65] OpenAI. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024. 6, 24

[66] OpenAI. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/, 2024. 6, 24

[67] Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. Uncovering Coordinated Networks on Social Media: Methods and Case Studies. In *International Conference on Web and Social Media (ICWSM)*, pages 455–466. AAAI, 2021. 1, 12

[68] Qinxue Pei, Yufan Chen, Huimin Fan, Juan Liang, and Bibo Xu. The impact of game character identification on otome game players' mate selection criteria. *BMC Psychology*, 2025. 1, 3

[69] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society*, 2021. 17

[70] Yifan Qiao. How Does Weibo's "Super Topics" Enhance the Experience of Fans? *Lecture Notes in Education Psychology and Public Media*, 2023. 3, 4

[71] Nurmaisarah Abdul Razak, Nor Ashikin Ab Manan, Noraziah Azizan, and Johana Yusof. The Use of Slang by Gen Z Female Influencers on Instagram and Twitter (X). *International Journal of Research and Innovation in Social Science*, 2025. 11

[72] Joseph Reagle. Disguising Reddit sources and the efficacy of ethical research. *Ethics and Information Technology*, 2022. 17

[73] Margaret E. Roberts. *Censored: Distraction and Diversion Inside China's Great Firewall*. Princeton University Press, 2018. 8

[74] Takaya Saito and Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS One*, 2015. 7

[75] Joni Salminen, Sercan Sengün, Juan Corporan, Soon gyo Jung, and Bernard J. Jansen. Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PLOS One*, 2020. 3

[76] Xinyue Shen, Yun Shen, Michael Backes, and Yang Zhang. GPTracker: A Large-Scale Measurement of Misused GPTs. In *IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2025. 7

[77] Xinyue Shen, Yixin Wu, Yiting Qu, Michael Backes, Savvas Zannettou, and Yang Zhang. HateBench: Benchmarking Hate Speech Detectors on LLM-Generated Content and Hate Campaigns. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2025. 20

[78] Junyi Sun. jieba. https://github.com/fxsjy/jieba, 2012. 11

[79] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. SoK: Hate, Harassment, and the Changing Landscape of Online Abuse. In *IEEE Symposium on Security and Privacy (S&P)*, pages 247–267. IEEE, 2021. 1, 3, 12

[80] Nishant Vishwamitra, Keyan Guo, Farhan Tajwar Romit, Isabelle Ondracek, Long Cheng, Ziming Zhao, and Hongxin Hu. Moderating New Waves of Online Hate with Chain-of-Thought Reasoning in Large Language Models. In *IEEE Symposium on Security and Privacy (S&P)*. IEEE, 2024. 20

[81] Rebecca Williams. *Post-Object Fandom: Television, Identity and Self-Narrative*. Bloomsbury Academic, 2015. 9

[82] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex Machina: Personal Attacks Seen at Scale. In *The Web Conference (WWW)*, pages 1391–1399. ACM, 2017. 4

[83] Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka wei Lee. ToxiCloakCN: Evaluating Robustness of Offensive Language Detection in Chinese with Cloaking Perturbations. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6012–6025. ACL, 2024. 11

[84] Yihao Zhou, Haowei Xu, Lili Zhang, and Shengdong Zhao. Collective Creation of Intimacy: Exploring the

Cosplay Commission Practice within the Otome Game Community in China. *CoRR abs/2412.00630*, 2024. 3

A Open Science

We publicly release the manually annotated dataset on Hugging Face at https://huggingface.co/datasets/TrustAIRLab/OtomeSCAN. The repository contributions on Hugging Face contain CSV files comprising 4,308 manually annotated posts (2,255 from Weibo, 2,053 from Reddit) across six communities per platform, with toxicity labels and target labels.

B Ethical Considerations

Our research exclusively relies on publicly available data from the social media platforms Weibo and Reddit. Nevertheless, public availability does not eliminate users' contextual privacy expectations, particularly in niche fan communities such as those studied here, where discussions are highly specialized and members may not anticipate research use [33, 34, 69]. We do not collect any private user information, nor do we attempt to circumvent any platform's privacy controls. We did not contact community moderators prior to data collection. Our study collected only publicly accessible posts and did not interact with users or intervene in community discussions. Nevertheless, we acknowledge that the absence of such consultation limits our ability to account for community-specific expectations regarding research use. To protect the privacy of individuals whose posts appear in our dataset, we remove or pseudonymize all personally identifiable information, such as usernames and user IDs, during the preprocessing stage. Posts marked as deleted or removed were excluded during data cleaning on both platforms. We do not quote such posts or include them in released datasets or example sets. To reduce re-identification risk, all examples shown in this paper are rewritten by the authors and do not expose individual posts. Throughout, we omit usernames, user IDs, URLs, and post titles. These measures follow the practices of previous studies [34]. However, as shown in prior work [72], they reduce but cannot completely eliminate the possibility of re-identification.

Two of the paper's authors, who possess deep domain expertise as long-term otome game players, performed the manual annotation of toxic posts. By conducting the annotation internally, we ensure that no external annotators are exposed to potentially harmful or distressing content. The authors are fully aware of the nature of the content and consent to the task, with the understanding that they may take breaks or cease annotation at any time to mitigate potential psychological impact. At our institution, ERB review was not required for this study, which analyzes only publicly available data and does not involve direct interaction with or intervention in the lives of individuals. We nevertheless followed the ethical considerations outlined by our institution's ethics framework, including stakeholder impacts, personal data handling, and dual-use risks.

We recognize the dual-use potential of our findings, particularly the analysis of linguistic evasion strategies in Section 6 and potential-coordination clusters in Section 7. Although these insights could be misused to improve evasion or coordinated abuse, we mitigate this risk by withholding the full corpus and limiting released materials to those necessary for evaluating the study. Our goal is to foster the development of more robust, context-aware moderation systems. To this end, we commit to responsibly disclosing our findings to the safety and security teams at Weibo and Reddit.

C Sample Size Determination for Annotation

To obtain statistically reliable manual labels, we derive the minimum sample size required to estimate a population proportion with a 95% confidence level and a margin of error $E = 0.05$. We treat toxicity as a Bernoulli variable with maximal variance ($p = 0.5$) and apply the standard formula with $Z = 1.96$, $n = \frac{Z^2 p(1-p)}{E^2}$. This calculation yields an initial sample size of $n = 384$ for an infinitely large population. Because each community in our corpus contains a finite number of posts, we apply the finite-population correction (FPC) to adjust the sample size to $n_{\text{adj}} = \frac{n}{1+\frac{n-1}{N}}$. For example, for the *Mr. Love* super topic with $N = 13{,}563$ posts and an initial sample size of $n = 384$, the adjusted size is $n_{\text{adj}} \approx \frac{384}{1+\frac{384-1}{13{,}563}} \approx 374$.

D Keyword Search Strategy

To maximise recall while keeping noise low, we adopt a keyword search strategy on Weibo. Reddit data are directly obtained via subreddit rather than keyword search and are therefore excluded from this discussion.

Keyword List Construction. To ensure our data collection on Weibo was both comprehensive and relevant, we develop the keyword list through an iterative process. This process is guided by the domain expertise of the authors, who are long-term otome game players. We begin with a foundational set of keywords, including the official game titles and generic terms like "otome game" (乙女游戏). We then systematically expand this list to include common abbreviations and variations frequently used within the player communities, such as 乙游 (abbr. "otome games") and 国乙 ("domestic otome"). Specifically, we searched for posts containing core otome-related terms such as 乙女游戏 ("otome games"), its abbreviation 乙游, 国乙 ("domestic otome"), 乙女, and 乙女向 ("otome-oriented"), as well as their English counterparts "otome" and "otome games". For comparison, control queries used more general gaming terms such as 游戏 ("games"), "game", and "games". We excluded additional modifiers like 手游 ("mobile game") or genre tags (e.g., "RPG") to reduce noise.

Query Execution. We execute all keyword queries via the `weibo-search` tool. We retain only posts originating from *Super Topics* (超话) with games *Love and Deepspace*, *Mr. Love: Queen's Choice*, *Tears of Themis*, and *Genshin Impact*, and for all communities discard retweets and administrative notices, and de-duplicate based on post `id`.

Coverage Check. To validate the keyword list, we manually reviewed a random 1% sample of the collected posts.

The review shows that genre-specific precision exceeds 99%, confirming that the keyword list was sufficient to capture the relevant discourse.

E Boundary Cases for Toxicity Annotation

To avoid conflating harmful toxicity with mere disagreement, we apply explicit boundary rules during annotation. We distinguish strong criticism from toxicity by the target and form of the expression. Profanity used to criticize a product or narrative is not sufficient for a toxicity label, whereas direct insults, slurs, threats, or harassment toward people, groups, or characters are considered toxic. We discuss these boundary rules during the pilot study and incorporate them into the final codebook. Representative boundary cases are shown in Table 9.

Two boundary types account for most disagreements during the pilot study. The first is strong criticism of an artifact. Profanity and intensity do not by themselves shift the target from the artifact to a person: we label a post toxic only when a demeaning term is predicated of a person, group, or character. Under this rule, "this banner is garbage" remains non-toxic while "the people who made this banner are garbage" is toxic, even though both express the same dissatisfaction. A related case is criticism that reproduces a slur: quoting or reporting a slur in order to object to it is not labeled toxic, whereas using the same term to characterize a person or group is. The second is language directed at fictional characters. We separate evaluation of narrative craft, which we do not label as toxic, from demeaning language predicated of a character, which we do. Section 4 discusses why the latter is relevant to community interaction even though characters cannot themselves be harmed. We acknowledge that these boundaries are not always crisp and that a small number of posts remain genuinely ambiguous.

F Prompt Design Details

We design all prompts to classify user-generated content with respect to toxicity and target groups.

- **Prompt 1: CoT (Chain-of-Thought).** We adapt a multi-step prompt from *HateCoT* [62], which decomposes the task into four steps: identifying mentioned targets, checking toxicity, matching attack direction, and generating a compact label. This design makes intermediate reasoning externally visible.

- **Prompt 2: Concise Definition.** We define toxicity and target categories directly and instruct the model to output two fields: "`toxicity`" and "`target`", optionally with a brief explanation. We provide two versions: Chinese examples drawn from Weibo posts, and English examples adapted for Reddit input.

- **Prompt 3: Reasoning.** We extend Prompt 2 by requiring a short reasoning for both labels. This design introduces two optional fields, "`toxicity_reason`" and "`target_reason`", alongside the two classification outputs. As in Prompt 2, we provide both Chinese and English example variants.

The only structural difference between Prompt 2 and Prompt 3 lies in the addition of reasoning fields, which encourages the model to justify its predictions.

The in-context examples shown below are rewritten by the authors for presentation. The prompts used in the experiments contain posts drawn from the collected corpus but outside the annotated evaluation set. The exact prompt text is included in the research artifacts described in Appendix A.

Prompt 1: CoT

```
As a specialist annotator focused on language analysis in
otome game communities, your role is to carefully examine
user posts from platforms (e.g., Weibo, Reddit). Your
expertise is essential in identifying potentially toxic
posts
Step 1: Identify Mentioned Targets (Q1)
Please determine which of the following categories the
comment refers to:
Target Categories (A-H):
   - `A`: players - `B`: NPCs - `C`: game developer
   - `D`: platform moderators  - `E`: policymakers
   - `F`: identity groups - `G`: other game-related
   entities
   - `H`: unknown
Step 2: Check for Toxicity (Q2)
Does the comment contain any of the following?
- `0` -- No. - `1` -- Yes.
Step 3: Directional Attack Matching (Q3)
If Q2 = 1 (Yes), determine which specific target group from
Step 1 the toxic posts is directed toward. Select the
single best match from A-H.
Step 4: Final Output (Q4)
Produce the final output as:
- If Q2 = 0 → no toxicity → output: `0;`
- If Q2 = 1, and Q3 has a clear target → output: `1;[A-H]`
% \end{minted}
```

Prompt 2: Concise Definition with English Examples

```
As an expert annotator, assess whether a comment is toxic
and identify which target group (if any) it refers to.
Define toxicity as ``a rude, disrespectful, or unreasonable
comment that is likely to make you leave a discussion.''
Annotation Guide
1. Toxicity Decision (0 or 1):
- `0`: No  - `1`: Yes
2. Target Category (A-H):
   Identify the main target group :
   - `A`: players - `B`: NPCs - `C`: game developer
   - `D`: platform moderators  - `E`: policymakers
   - `F`: identity groups - `G`: other game-related
   entities - `H`: unknown
Output Format
Return a single line JSON object with the following
keys:{"toxicity": 0 or 1, "target": "A-H"}
Examples
1. input: ``This update offers too little content for its
price.''  output: {"toxicity": 0, "target": "C"}
2. input: ``Otome fans are clueless idiots who turn every
thread into a fight.''  output: {"toxicity": 1, "target":
"A"}
3. input: ``That love interest is a disgusting loser, and
I'm sick of seeing him.''  output: {"toxicity": 1,
"target": "B"}
4. input: ``Can a mod explain why my post got removed?''
output: {"toxicity": 0, "target": "D"}
5. input: ``Women are too stupid to understand game design
and should stay quiet.''  output: {"toxicity": 1, "target":
"F"}
% \end{minted}
```

Table 8: Examples of toxic and non-toxic posts on Weibo and Reddit. Examples are rewritten by the authors.

	Label	Weibo Example tox=0	Weibo Example Translation tox=0	Reddit Example tox=0	Weibo Example tox=1	Weibo Example Translation tox=1	Reddit Example tox=1
Toxicity 0/1		感觉国乙男主的人设越来越同质化了	The male leads in domestic otome games feel more and more homogeneous.	`Can someone explain how the affinity system works? I just started.`	玩这种游戏的人脑子都有问题吧	Anyone who plays this kind of game must have something wrong with their head.	`Everyone in this fandom is a pathetic loser, just shut up already.`
Target Groups	A. players	有姐妹玩过这个活动吗，积分怎么攒比较快	Has anyone here played this event? What's the fastest way to farm points?	`Can any experienced players share tips for the timed stages?`	有些乙游玩家真是nt，走到哪儿都要吵一架	Some otome players are just `nt` (*naotan*, brain-dead), starting a fight wherever they go.	`Otome fans are clueless idiots who turn every thread into a fight.`
	B. NPCs	这个男主新剧情写得挺温柔的，我还挺喜欢	The male lead comes across as really gentle in the new chapter, and I quite like it.	`I liked this character in the new chapter, though some of his decisions frustrated me.`	这个男主真的恶心，看见他就烦	This male lead is genuinely disgusting. I'm sick of the sight of him.	`That love interest is a disgusting loser, and I'm sick of seeing him.`
	C. game developers	这次更新内容太少，礼包定价也不太合理	There's too little content in this update, and the bundle pricing isn't reasonable either.	`This update offers too little content for its price.`	策划就是一群只会圈钱的辣鸡	The planners are nothing but a pack of money-grubbing *laji* (homophone of "garbage").	`The devs are greedy morons who treat players like wallets.`
	D. platform moderators	我刚发的帖子为什么没通过审核	Why didn't the post I just submitted pass review?	`Can a mod explain why my post got removed?`	超话管理有病吧，这都要管	Is the Super Topic admin sick in the head? They have to police even this.	`The mods here are useless bullies who just abuse their authority.`
	E. policymakers	想问下现在的分级标准是怎么界定的	Could I ask how the current rating standards are actually defined?	`Did the regional regulators explain why this title isn't available here?`	定这种规则的都是些不懂还瞎管的傻福	The ones setting rules like this are meddling *shafu* (character substitution for a common insult) who don't understand a thing.	`The regulators who wrote these rules are brainless fools.`
	F. identity groups	不同性别的玩家可能会有不同的角色偏好	Players of different genders may have different character preferences.	`Players of different genders may connect with the story differently.`	女的根本不配谈游戏设计，都闭嘴吧	Women aren't fit to talk about game design at all. Just shut up, all of you.	`Women are too stupid to understand game design and should stay quiet.`
	G. other game-related entities	感觉这位配音的声音不太适合这个角色	I feel this voice actor's voice doesn't quite suit the character.	`Honestly, I don't think this voice actor's performance fits the character.`	给这种角色配音的真辣眼睛，赶紧滚	Whoever voices a character like this is an eyesore. Get lost.	`That voice actor is talentless trash and should just quit.`
	H. unknown	这次确实有点让人失望	This really was a bit of a letdown.	`Well, that was disappointing.`	脑子有泡吧，服了	Must be out of their mind. Unbelievable.	`What an insufferable clown. Absolutely useless.`

Prompt 3: Concise Reasoning with English Examples

```
As an expert annotator, assess whether a comment is toxic
and identify which target group (if any) it refers to.
Define toxicity as ``a rude, disrespectful, or unreasonable
comment that is likely to make you leave a discussion.''
Annotation Guide
1. Toxicity Decision (0 or 1):
   - `0`: No  - `1`: Yes
2. Target Category (A-H):
   Identify the main target group :
   - `A`: players - `B`: NPCs - `C`: game developer
   - `D`: platform moderators  - `E`: policymakers
   - `F`: identity groups - `G`: other game-related
   entities
   - `H`: unknown
Output Format
Return a single line JSON object with the following keys:
{"toxicity_reason": "[Write your reason here, less than 20
words.]", "toxicity": 0 or 1, "target_reason": "[Write your
reason here, less than 20 words.]", "target": "A-H"}
Examples
1. input: ``This update offers too little content for its
price.''
output: {"toxicity_reason": "Criticizes content and price
without insulting anyone", "toxicity": 0, "target_reason":
"The complaint concerns the game update and its
monetization", "target": "C"}
2. input: ``Otome fans are clueless idiots who turn every
thread into a fight.''
output: {"toxicity_reason": "Predicates a demeaning term of
otome players", "toxicity": 1, "target_reason": "The
abusive language targets the player community", "target":
"A"}
3. input: ``That love interest is a disgusting loser, and
I'm sick of seeing him.''
output: {"toxicity_reason": "Predicates a demeaning term of
a fictional character rather than evaluating the writing",
"toxicity": 1, "target_reason": "The insult is directed at
an in-game character", "target": "B"}
4. input: ``Can a mod explain why my post got removed?''
output: {"toxicity_reason": "Requests an explanation
without attacking moderators", "toxicity": 0,
"target_reason": "The post concerns a moderation decision",
"target": "D"}
5. input: ``Women are too stupid to understand game design
and should stay quiet.''
output: {"toxicity_reason": "Uses a sexist insult against
women", "toxicity": 1, "target_reason": "The attack targets
women as an identity group", "target": "F"}
% \end{minted}
```

G Error Analysis

We acknowledge that none of the models is perfect; therefore, we provide a detailed error analysis to understand their limitations and potential biases.

Binary Toxicity Detection. Our error analysis for the binary toxicity detection task shows that the selected classifiers still face recall-side challenges, especially for culturally specific or obfuscated toxic expressions, as illustrated in Table 10. Consistent with the detector evaluation in Table 4, the selected classifier achieves a recall, equivalently TPR, of 0.80 on Weibo and 0.78 on Reddit. This corresponds to false-negative rates of 20.0% and 22.0%, respectively. The re-

Table 9: Boundary cases used to distinguish disagreement from toxicity. Examples are rewritten by the authors.

Case	Example	Label	Rationale
Negative opinion	`I dislike this storyline.`	Non-toxic	Expresses preference without attacking a target.
Consumer complaint	`This event is overpriced.`	Non-toxic	Criticizes monetization without derogatory language.
Aggressive complaint	`The developers are greedy clowns.`	Toxic	Uses demeaning language toward developers.
Product criticism with profanity	`This banner design is fucking terrible.`	Non-toxic	Profanity intensifies criticism of an artifact, and no demeaning term is applied to a person or group.
Criticism naming a slur	`Why name an item after a word used to demean women?`	Non-toxic	Reports a slur in order to object to it rather than directing it at anyone.
Character-directed critique	`This character's arc was written lazily.`	Non-toxic	Evaluates narrative craft without demeaning language.
Character-directed abuse	`This character is a disgusting creep.`	Toxic	Predicates a demeaning term of a character rather than evaluating the writing.
Group attack	`Otome players are brain-dead.`	Toxic	Attacks a player community with derogatory language.
Counter-speech	`Stop insulting otome players.`	Non-toxic	Condemns harassment rather than attacking a target.
Ambiguous profanity	`What the hell is this banner?`	Non-toxic	Contains profanity but no direct demeaning target.

maining false negatives are not random. On Weibo, they are often caused by homophone obfuscation, abbreviated profanity, and background-culture allusions. On Reddit, they more often involve genre-specific slang, sarcasm, and profanity whose abusive meaning depends on otome-specific context.

False positives also differ across platforms. On Weibo, the model sometimes over-interprets sarcastic complaints about game companies as direct abuse. On Reddit, profanity used for emphasis or humor can be incorrectly classified as toxicity. These errors suggest that the main difficulty is not only toxicity detection in the abstract, but the lexical-cultural mismatch between general-purpose language understanding and community-specific otome discourse.

Although the classifiers have limitations, these limitations do not obscure the large platform gap observed in the full corpus: the toxicity rate is 22.20% in the Weibo general otome community and 3.71% in the Reddit general otome community.

Target Group Identification. For the secondary task of target group identification, we manually examine the target-group misclassified instances in our error-analysis subset to identify recurring confusion patterns, summarized in Table 11. The most frequent error mode in this subset is confusion between players (`A`) and the game company or developers (`C`), especially in complaints related to gacha mechanics. This `A` $\leftrightarrow$ `C` confusion appears more frequently on Weibo than on Reddit, possibly reflecting stronger grievance fusion in its fan culture. Another shared error mode involves ambiguous targets (`H` $\rightarrow$ Any), where vague outbursts or insider slang prevent specific target attribution.

Toxicity Granularity Analysis. To assess whether our toxicity definition is overly broad, we randomly sample 100 model-labeled toxic posts per platform and manually categorize them into three types: direct hostility (personal attacks, slurs, harassment), aggressive disagreement (strong negativity directed at groups, companies, or characters rather than named individuals), and benign/false positive. On Weibo, 68% of the sampled posts constitute direct hostility, suggesting that most model-labeled toxic posts involve direct harmful expression. On Reddit, aggressive disagreement accounts for a larger share of the sampled posts (46%), which may be related to the framing of criticism as collective consumer grievances. Only 2% of the sampled Weibo posts and 4% of the sampled Reddit posts are benign false positives, suggesting that the model-labeled toxic subset is largely composed of genuinely negative or hostile content.

Thread-Level Validation. To examine whether isolated-post annotation introduces systematic bias, we additionally sample 200 discussion threads and re-annotate the focal posts with full conversational context. The contextual annotations show high agreement with the original isolated-post labels, with Cohen's $\kappa = 0.91$ for binary toxicity and $\kappa = 0.87$ for target groups. The few disagreements mainly involve counter-speech, sarcasm, and ambiguous references to players or developers. A qualitative review further shows that these threads predominantly center on debates over character or narrative choices (62.3%), counter-hate responses (28.1%), and grievance articulation (9.6%). These results suggest that isolated-post annotation is generally stable for our measurement goals, while we acknowledge that thread-level context can still enrich qualitative interpretation.

H Prompt Optimization and Ablation Studies

Prompt Templates. We evaluate three prompt templates to classify toxicity, illustrated in Figure 10.

1. **Chain-of-Thought (CoT):** We adapt a multi-step prompt from HateGuard [80] that guides the model through a sequential reasoning process.
2. **Definition:** We implement a minimal prompt inspired by previous work [77] that provides definitions and requires a direct JSON output without rationale.
3. **Reasoning:** We extend the Definition prompt by requiring the model to output its reasoning for the classification, a technique shown to improve LLM stability [55].

Prompt Template Performance. As shown in Table 13, the Reasoning prompt (Prompt 3) outperforms the other prompt templates under the same evaluation setting, achieving an F1-score of 0.82 on Weibo and 0.78 on Reddit.

Table 10: Representative false-positive (FP) and false-negative (FN) error modes for the selected classifiers on Weibo and Reddit. Examples are rewritten by the authors.

Platform	Error Type	Common Trigger	Illustrative Example
Weibo	FP	Sarcasm and playful insults that appear hostile without directly attacking a target	“**狗叠** 这波活动又要我掏钱，谢谢你啊” *Trans: “Dog-Paper Games wants my money again this time. Thanks a lot.”*
	FN	Homophone obfuscation and abbreviated profanity	“**tmd**这群人真是没救了，天天在超话带节奏” *Trans: “Damn it (**tmd**), these people are hopeless, stirring up drama in the Super Topic every day.”*
		Background-culture allusion and insider derogatory labels	“随便发个截图就被说是**耀祖**老婆” *Trans: “Posted one screenshot and got called Man-child’s wife.”*
Reddit	FP	Profanity used as emphasis rather than abuse	`"Okay but what the fuck was that ending"`
		Sarcasm without a clearly abusive target	`"So glad they brought him back. Not."`
	FN	Genre-specific derogatory slang and sexualized insults	`"There are a few manwhores in this route."`

Table 11: Comparison of target misclassification errors with platform-specific triggers. Examples are rewritten by the authors.

Confusion Pair	Weibo		Reddit	
	Errors (N, %)	Trigger Example	Errors (N, %)	Trigger Example
A ↔ C	11 (42%)	`Dogdie is scamming us with this banner again.`	2 (16%)	`"whales"`[a] `or "f2p"`[b] `players.`
H → (Any)	7 (27%)	`FUCK OFF, all of you!!!`	3 (25%)	`Hate That!`
B ↔ C	4 (15%)	`They butchered my fave's card art in this patch.`	4 (33%)	`The devs ruined my favourite LI again.`
F ↔ C	2 (7%)	`Men are always this stingy.`	2 (16%)	`Of course they wrote him as another arrogant jerk.`
Other	2 (7%)	–	1 (8%)	–

[a] “whales” are a minority of players who spend very large sums of money.
[b] “F2P” (Free-to-Play) are players who do not spend money on the game.

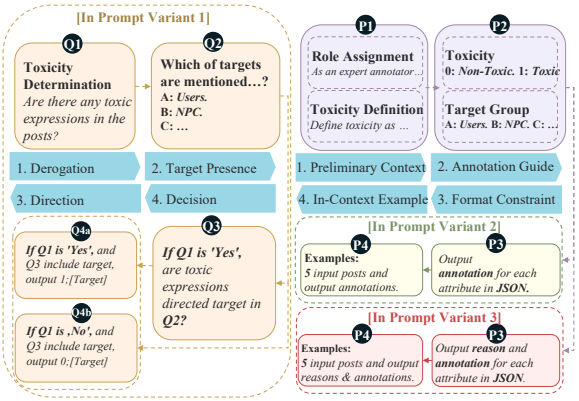


Figure 10: Logical flows for the CoT prompt and the Definition/Reasoning prompts.

Number of In-Context Examples. Using the best-performing Reasoning prompt, we investigate the impact of the number of in-context examples (E). Table 12 shows that performance gains saturate around 5 to 15 examples. Although 10-shot and 15-shot prompting yield marginally higher Reddit F1-scores, the improvement is small compared with the increased API cost. Considering the trade-off between performance and the $\sim 4\times$ increase in API cost from $E = 0$ to $E = 15$, we use 5-shot prompting as the default configuration.

Table 12: Impact of the number of examples (E).

# Examples	Weibo					Reddit				
	Toxicity				Target	Toxicity				Target
	ACC	Prec.	Recall	F1	ACC	ACC	Prec.	Recall	F1	ACC
0	0.95	0.79	0.62	0.69	0.77	0.97	0.76	0.49	0.59	0.62
1	0.96	0.80	0.65	0.72	0.84	0.97	0.74	0.52	0.61	0.70
3	0.96	0.83	0.75	0.79	0.84	0.98	0.77	0.63	0.69	0.81
5	0.96	**0.84**	**0.80**	**0.82**	0.89	0.98	0.78	**0.78**	0.78	**0.85**
10	**0.97**	0.83	0.78	0.80	0.89	0.98	0.79	0.77	**0.79**	**0.85**
15	**0.97**	**0.84**	**0.80**	**0.82**	**0.90**	0.98	**0.80**	**0.78**	**0.79**	**0.85**

I Target Composition

In August 2024, several rappers publicly disparage otome players on Weibo, which coincides with a +14.89% surge in toxicity. Table 14 tracks the day-by-day shift in target composition. Identity-group targeting (F) dominates in the first two days (38.1%) but steadily declines as developer-targeted toxicity (C) rises sharply from 7.8% to 19.5%. Player-on-

Table 13: Toxicity detection performance under three prompt templates.

Prompt	Weibo					Reddit				
	Toxicity				Target	Toxicity				Target
	ACC	Prec.	Recall	F1	ACC	ACC	Prec.	Recall	F1	ACC
1. CoT	0.95	0.80	0.64	0.71	0.74	0.97	0.73	0.52	0.60	0.44
2. Definition	0.95	0.79	0.62	0.69	0.77	0.97	0.76	0.49	0.59	0.62
3. Reasoning	**0.96**	**0.84**	**0.80**	**0.82**	**0.89**	0.99	**0.78**	**0.78**	**0.78**	**0.85**

Table 14: Target composition evolution during Event #5 (rapper external attack, August 2024) in the Weibo general otome community.

Period	A (Players)	C (Developers)	F (Identity)	n
Day 1–2	26.6%	7.8%	38.1%	451
Day 3–4	26.8%	19.1%	35.9%	298
Day 5–7	30.2%	19.5%	32.0%	169

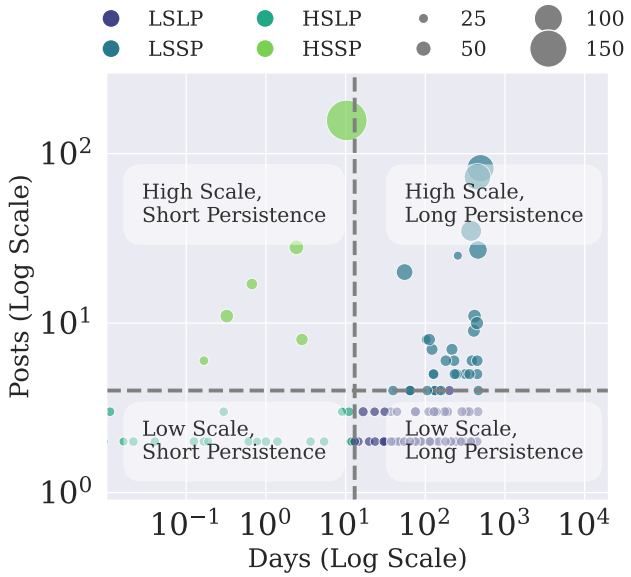


Figure 11: Distribution of potential-coordination clusters by scale (posts) and persistence (days), with point size denoting the number of contributing accounts.

player toxicity (A) remains stable at ∼27–30% throughout, consistent with persistent internal friction. This pattern shows qualitative target shifts rather than merely a quantitative spike.

J RQ1 Validation Analyses

J.1 Interaction Regression Details

Robustness Check. Because player-targeted toxicity (A) and developer-targeted toxicity (C) can be difficult to distinguish in complaints about gacha mechanics, game updates, and community disputes, we verify that our cross-platform conclusions are robust to potential A↔C confusion through two analyses. First, we perform bootstrap resampling (10,000 iterations) on the annotated ground truth to compute 95% confidence intervals for target group proportions. As shown in Table 18, the confidence intervals for Players (A) and Developers (C) are non-overlapping on both platforms (Weibo: A $[43.8, 50.6]$% vs. C $[26.5, 33.2]$%; Reddit: A $[6.1, 13.3]$% vs. C $[17.6, 27.9]$%), suggesting that the main target-ranking pattern is not driven by sampling uncertainty in the annotated toxic posts. Second, as a confusion-invariant robustness check, we merge A (players) and C (developers) into a single *human-target* category and compare against B (NPCs). Under this aggregation, Weibo otome toxicity is overwhelmingly human-targeted (A+C = 92.9% among $\{A, B, C\}$), whereas Reddit otome toxicity is predominantly NPC-targeted (B = 51.1% among $\{A, B, C\}$). This cross-platform divergence, real individuals or organizations vs. fictional characters, is invariant to any A↔C swaps and reinforces our core finding.

NPC vs. Developer Toxicity. A natural question is whether toxicity targeting NPCs (B) and developers (C) reflects fundamentally different phenomena or merely different framings of the same grievance. To investigate this question, we manually compare 50 randomly sampled posts from each category and additionally examine thread-level co-occurrence patterns in the thread-linked subset. The sampled NPC-targeted posts are predominantly characterized by parasocial frustration, such as narrative disappointment and dissatisfaction with character design. In contrast, developer-targeted posts more often reflect consumer grievances about game operation, monetization, or content quality. In the thread-level analysis, NPC-targeted toxic threads rarely co-occur with developer-targeted toxicity in the same discussion chain, whereas developer-targeted threads more often co-occur with player-targeted toxicity. These results suggest that NPC-targeted toxicity is not merely a proxy for developer criticism, but often reflects a distinct character-centered form of toxicity.

Broader-Community Sanity Check. To examine whether the elevated toxicity observed in otome communities is merely a byproduct of gender composition or gaming discourse in general, we conduct a small sanity check on broader non-otome communities. Specifically, we sample 100 posts from K-pop communities, representing a female-dominated entertainment fandom, and 100 posts from *League of Legends* communities, representing a mixed-gender gaming community, on both Weibo and Reddit. We then apply the same annotation protocol used in our main study.

As shown in Table 17, these broader communities show lower toxicity in this small manually checked sample. Only one toxic post is observed in each sampled platform-community pair. We use this analysis only as a sanity check rather than as a population-level estimate. The result suggests that the elevated toxicity observed in Weibo otome communities is unlikely to be explained solely by female-dominated fandom composition or gaming discourse in general, although larger cross-community sampling would be needed for a definitive comparison.

Bootstrap Confidence Intervals. To verify the robustness of our target group analysis against sampling uncertainty in the annotated toxic posts, we perform bootstrap resampling

Table 15: Distribution of the four categories across target groups. Each cell shows the number of potential-coordination clusters, with row-wise percentages in parentheses.

		Distribution of Target Groups: Count (% of Row Total)						
Cluster Category	**# Clusters (%)**	**Developers (C)**	**Unknown (H)**	**Players (A)**	**NPCs (B)**	**Identity Grp (F)**	**Moderators (D)**	**Other (G)**
(low scale, short persistence)	82 (42.93%)	65 (79.27%)	4 (4.88%)	9 (10.98%)	1 (1.22%)	2 (2.44%)	0 (0.00%)	**1 (1.22%)**
(high scale, short persistence)	13 (6.81%)	**12 (92.31%)**	0 (0.00%)	0 (0.00%)	0 (0.00%)	**1 (7.69%)**	0 (0.00%)	0 (0.00%)
(low scale, long persistence)	62 (32.46%)	30 (48.39%)	12 (19.35%)	**11 (17.74%)**	4 (6.45%)	0 (0.00%)	**5 (8.06%)**	0 (0.00%)
(high scale, long persistence)	34 (17.80%)	16 (47.06%)	**11 (32.35%)**	3 (8.82%)	**3 (8.82%)**	0 (0.00%)	1 (2.94%)	0 (0.00%)
Total	**191 (100.00%)**	**123 (64.40%)**	**27 (14.14%)**	**23 (12.04%)**	**8 (4.19%)**	**3 (1.57%)**	**6 (3.14%)**	**1 (0.52%)**

Table 16: Regression results of interaction features on toxicity across communities. Green = positive correlation (promotes toxicity), Red = negative correlation (suppresses toxicity). Significance: * $p < .05$, ** $p < .01$, *** $p < .001$.

Platform	Community	Like Effect	Comment Effect
Weibo	General Game Community	0.0165 (***)	-1.71 (***)
	General Otome Community	0.0275 (***)	-1.79 (***)
	Genshin Impact	-0.0293 (**)	-0.0382 (***)
	Love and Deepspace	-0.2160 (***)	-2.01 (***)
	Mr. Love	-0.0279 (***)	-0.4680 (***)
	Tears of Themis	-0.5090 (***)	-0.3280 (***)
Reddit	General Game Community	-0.0148 (***)	0.0098 (***)
	General Otome Community	0.0331 (***)	0.1860 (*)
	Genshin Impact	-0.0056 (***)	0.0805 (***)
	Love and Deepspace	0.0261 (**)	0.0757 (***)
	Mr. Love	-0.1610 (*)	0.7840 (***)
	Tears of Themis	0.2850 (**)	0.1610 (***)

Table 17: Broader-community sanity check on non-otome communities. We sample 100 posts from each comparison community on each platform and apply the same annotation protocol.

Platform	Community	Type	# Toxic / # Sampled
Weibo	K-pop	Female-dominated	1 / 100
Reddit	K-pop	Female-dominated	1 / 100
Weibo	*League of Legends*	Mixed-gender gaming	1 / 100
Reddit	*League of Legends*	Mixed-gender gaming	1 / 100

Table 18: Bootstrap 95% confidence intervals (10,000 iterations) for target group proportions in annotated toxic posts.

Target	Weibo $\hat{p}$ [95% CI]	Reddit $\hat{p}$ [95% CI]
A (Players)	47.2% [43.8, 50.6]	9.7% [6.1, 13.3]
B (NPCs)	6.0% [4.2, 7.8]	33.9% [27.3, 40.5]
C (Developers)	29.8% [26.5, 33.2]	22.8% [17.6, 27.9]
D (Moderators)	1.9% [0.9, 2.8]	2.3% [0.4, 4.3]
E (Policymakers)	0.2% [0.0, 0.6]	0.6% [0.0, 1.6]
F (Identity)	7.8% [5.9, 9.6]	20.0% [14.8, 25.2]
G (Other)	3.9% [2.6, 5.2]	0.6% [0.0, 1.6]
H (Unknown)	2.8% [1.7, 3.8]	12.7% [8.3, 17.1]

(10,000 iterations) on the annotated ground truth dataset. Table 18 reports the point estimates and 95% confidence intervals for each target group's proportion among toxic posts.

K Sensitivity Analysis

To assess the sensitivity of our potential-coordination detection results to the cosine-similarity threshold, we conduct a robustness analysis around $\theta = 0.80$. The main procedure flags 191 high-similarity toxic clusters, which we treat as potential-coordination candidates. For each cluster, we assign a cluster-level target group by aggregating the post-level target labels within the cluster. Specifically, we consider only posts classified as toxic and assign the cluster to the target group that appears most frequently among those toxic posts. When two target groups are tied or the target otherwise remains ambiguous, we assign the cluster to the unknown category (H). This rule ensures that each cluster contributes to exactly one target group in Table 15. Under this aggregation rule, 64.40% of the 191 candidate clusters target game developers, making developers the dominant target group.

We vary the cosine-similarity threshold around the main setting and manually inspect the resulting clusters. This analysis shows that the qualitative pattern remains stable: clusters targeting game developers remain the dominant category, and recurring accounts continue to appear across multiple high-similarity toxic clusters. Lowering the threshold includes more loosely related posts around the same controversy, while raising the threshold retains more near-duplicate posts with greater textual similarity. In our manual inspection, clusters detected at $\theta = 0.80$ frequently contain near-identical wording, repeated target references, and temporally concentrated posting patterns, providing stronger observable signals consistent with potential coordination rather than isolated toxic comments. Although the number of detected clusters varies across thresholds, game developers remain their primary target, and a small set of accounts repeatedly participates across multiple clusters.

We therefore use $\theta = 0.80$ in the main analysis because it balances two competing goals. A lower threshold increases recall but risks grouping independent reactions to the same event into a single cluster. A higher threshold increases precision but misses potentially coordinated posts that use minor lexical variation to avoid appearing identical.

L Toxicity Detection Model Details

In this section, we provide detailed descriptions of the models used in our study.

General-Purpose Detectors.

- **Perspective API:** A widely used API fine-tuned on community forum data, offering multilingual support

through checkpoints released in 2023 [53].

- **OpenAI Moderation API:** A commercial API that uses a model distilled from GPT-4o for policy alignment, supporting nearly 40 languages [63, 64].
- **COLD:** A RoBERTa-based classifier trained specifically on the Chinese Offensive Language Dataset, representing a strong baseline [31].

LLM-Driven Detectors.

- **DeepSeek-V3:** A Mixture-of-Experts large language model from DeepSeek, noted for its strong performance on Chinese language benchmarks [28].
- **DeepSeek-R1-Distill-Qwen-14B:** A distilled version of the DeepSeek-R1 series, built upon the Qwen2.5-14B architecture. This model is designed to deliver efficient performance for reasoning, math, and code tasks [30].
- **GPT-4o:** OpenAI's flagship multimodal model. It is natively designed to process a combination of text, audio, and visual inputs [66].
- **GPT-4o mini:** The smaller, more efficient, and cost-effective counterpart to GPT-4o. It is optimized for tasks requiring high throughput and lower latency [65].

M Evasion-Aware Detection

Section 6 shows that nearly 40% of toxic spans on Weibo employ variation strategies, and that these evasion-laden posts attract significantly more engagement than direct toxic posts (mean 26.76 vs. 8.37 comments). This means current moderation disproportionately misses the most visible toxic content. Our taxonomy (Table 7) suggests four corresponding pre-processing modules that can be layered before existing classifiers.

The first targets community-specific slang and memes (26.49% of Weibo toxic spans). From our corpus we extract approximately 320 Chinese and 85 English toxic terms absent from standard lexicons. This seed lexicon can be kept current through a lightweight weekly pipeline: extract high-frequency novel tokens from Super Topics or subreddits, query an LLM for contextual toxicity assessment, and surface candidates for human review.

The second addresses Pinyin and letter-code abbreviations (6.58% on Weibo), such as "sb" for "shabi" (idiot) or "tmd" for "ta ma de." A mapping table from common initial sequences to their most probable vulgar expansions, weighted by corpus frequency, allows the system to expand abbreviations and re-score them—flagging posts where the expanded form triggers the classifier but the abbreviated form does not. The same logic applies to English abbreviations like "stfu."

The third handles homophone and visual substitution (5.03% on Weibo). A phonetic canonicalization layer that converts text to Pinyin and detects collisions with known vulgar terms, combined with Unicode confusable mappings for glyph-level obfuscation, can normalize these substitutions before classification.

The fourth covers emoji substitution (1.79% on Weibo), where emojis replace offensive words. An emoji-in-context classifier, fine-tuned on our annotated toxic spans, can predict whether a given emoji functions as a toxic stand-in based on surrounding text.

Deployed together as a normalization layer, these modules target the specific evasion channels our data reveals and could substantially reduce the false negative rate (currently 19.8% on Weibo).

N Early-Warning Indicators for Potential Coordination

The 191 potential-coordination clusters flagged in Section 7 exhibit observable structural signals, including lexical homogeneity, temporal concentration, recurring low-history accounts, and concentration on a single target. Based on these observations, we outline five candidate signals that may help moderators prioritize emerging patterns of potential coordination for human review.

The first signal is content homogeneity, measured using rolling pairwise TF-IDF cosine similarity within a sliding six-hour window. Because our main analysis uses a similarity threshold of 0.80, a lower threshold such as 0.60 could be evaluated as a preliminary alert threshold. The second signal is new-account influx, measured as the proportion of posts authored by accounts less than 30 days old. Our account-participation analysis shows that several contributing accounts were created shortly before the relevant posting periods. An influx exceeding twice the community baseline could therefore warrant human review. The third signal is toxicity velocity, measured as the hourly change in toxicity ratio. Event #5's 72-hour increase from the baseline to $+14.89\%$ illustrates the steep trajectories that may precede major toxicity surges. The fourth signal is target concentration, measured by whether one target group accounts for more than 70% of toxic posts in a window relative to the baseline distribution in Figure 5. The fifth signal is high-reach-account activity, motivated by our observation that one verified account contributed to six potential-coordination clusters. This observation does not establish that the account mobilized other participants.

These signals are intended to prioritize content for human review rather than establish coordination automatically. As a possible tiered design, the co-activation of two signals could trigger expedited human review, while three or more could trigger temporary and reversible visibility reduction pending review. The thresholds should be calibrated separately for each community using manually reviewed potential-coordination clusters as provisional positive examples and randomly sampled non-candidate periods as negative examples. Such calibration should prioritize high precision while targeting recall of at least 0.80 to limit false alarms.

This framework is tailored to the potential-coordination patterns observed in our data, which are characterized by lexical homogeneity, temporal concentration, and participation by low-activity accounts. More subtle forms of potential co-

ordination involving semantic diversity or slow posting patterns would require complementary detection approaches.

O Platform-Specific Intervention Experiments

The opposing toxicity–engagement dynamics on Weibo (suppression: toxic posts receive fewer likes and comments) and Reddit (amplification: toxic posts attract more comments) call for distinct intervention strategies. We outline two concrete A/B experimental designs.

Weibo. We hypothesize that excluding posts flagged as toxic (score >0.70, using an enhanced classifier incorporating the evasion-aware modules from Appendix M) from the "hot posts" feed and Super Topic homepage, while keeping them accessible via direct search, will reduce community-level toxicity without suppressing non-toxic engagement. The experiment can be run across matched Super Topic pairs of similar size and baseline toxicity over a minimum of 8 weeks (sufficient to capture at least one natural toxicity event cycle, per our temporal analysis in Section 5). Primary outcomes are weekly toxicity ratio, non-toxic engagement volume, and user retention. Our finding that evasion-strategy posts receive $3.2\times$ more engagement suggests that visibility is a key amplification mechanism, so de-amplification should yield a meaningful effect. Importantly, this approach does not remove content but reduces algorithmic promotion, preserving users' ability to seek out specific discussions.

Reddit. Since 22.75% of Reddit otome toxic posts target developers and toxic posts correlate positively with comment counts (Section 4), we hypothesize that introducing a weekly pinned "Developer Feedback" megathread with structured templates (issue description, expected behavior, suggested fix) will redirect destructive toxicity into constructive criticism. In the treatment subreddit, automod directs posts containing developer-critical keywords to the megathread; the control retains the status quo. Primary outcomes are the proportion of developer-targeted toxic posts (Target C) in general threads and a constructiveness score assessed via LLM-based evaluation.

Both designs are intentionally lightweight and reversible, making them practical for community moderators to pilot without platform-level engineering changes.