

Article

Temporal Dynamics of Harmful Speech in Chatbot–User Dialogues: A Comparative Study of LLM and Chit-Chat Systems

Ohseong Kwon ^{1,†}, Hyobeen Yoon ^{1,†}, Hyojin Chin ¹  and Jisung Park ^{2,*}

¹ Department of Computer Science and Engineering, Gyeongsang National University, 501 Jinju-daero, Jinju-si 52828, Gyeongsangnam-do, Republic of Korea; osos0510@gnu.ac.kr (O.K.); dbsgyqls1218@gmail.com (H.Y.); hyojinchin@gnu.ac.kr (H.C.)

² Department of Mechanical Convergence Engineering, Gyeongsang National University, 501 Jinju-daero, Jinju-si 52828, Gyeongsangnam-do, Republic of Korea

* Correspondence: jisung@gnu.ac.kr

[†] These authors contributed equally to this work.

Abstract

Harmful language in conversational AI poses distinct safety and governance challenges, as Large Language Model (LLM) chatbots interact in private, one-to-one settings. Understanding the types of harm and their temporal concentration is crucial for responsible deployment and time-aware moderation. This study investigates the types and diurnal dynamics of harmful speech, comparing patterns between play-oriented chit-chat and task-oriented LLM services. We analyze two large-scale, real-world English corpora: a chit-chat service (SimSimi; 8.7 M utterances) and an LLM service (WildChat; 610 K utterances). Using the Perspective API for multi-label classification (Toxicity, Profanity, Insult, Identity Attack, Threat), we estimate the incidence of harm categories and compare their distribution across five dayparts. Our analysis shows that harmful speech is significantly more prevalent in the chit-chat context than in the LLM service. Across both platforms, Toxicity and Profanity are the dominant categories. Temporally, harmful speech concentrates most frequently during the dawn daypart. We contribute an empirical baseline on how harm varies by chatbot modality and time of day, offering practical guidance for designing dynamic, platform-specific moderation policies.

Keywords: harmful speech; chatbot; WildChat; chatbot user dialogue; SimSimi; offensive languages



Academic Editor: Carlos Delgado

Received: 15 November 2025

Revised: 10 December 2025

Accepted: 15 December 2025

Published: 16 December 2025

Citation: Kwon, O.; Yoon, H.; Chin, H.; Park, J. Temporal Dynamics of Harmful Speech in Chatbot–User Dialogues: A Comparative Study of LLM and Chit-Chat Systems. *Appl. Sci.* **2025**, *15*, 13185. <https://doi.org/10.3390/app152413185>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Large-language-model (LLM) chatbots are rapidly diffusing across consumer and enterprise contexts as a core technology for natural-language interaction. In parallel, public concern over ethical use and user protection has intensified. Prior work indicates that chatbot–user interactions, by virtue of their open-ended nature, can contain negative utterances such as profanity, insults, and threats [1–3]. Such content not only degrades user experience but may also pose safety risks for specific groups if dialogue data are later reused for model training [4,5]. It is therefore essential to precisely characterize which types of harmful speech occur and when they arise during user–chatbot exchanges, in order to inform service reliability and user-protection strategies.

Nevertheless, most research on online abusive or toxic language has focused on social media (e.g., Twitter, Facebook) or online communities, analyzing occurrence, categories,

and diffusion dynamics [6,7]. AI chatbots—including LLM-based systems—differ fundamentally from social media in their contextual specificity (1:1 dialogues) and the iterative generation of responses [8]. Dedicated investigation of harmful speech in chatbot–user conversations is thus warranted. While a small body of work analyzes chatbot data, the emphasis has largely been on improving automatic detection models [3,9]. Systematic, quantitative analyses of type-specific incidence and time-of-day distributions in real-world interactions, comparing chatbot types (e.g., chit-chat vs. LLM-based), remain scarce.

“A chit-chat system” denotes an open-domain, non-task-oriented conversational agent designed for informal, free-form social talk in 1:1 dialogues, rather than for executing specific user goals. Interactions are typically real-time or near-real-time and emphasize engagement, rapport, and small talk over task completion [10,11]. In contrast, we use “task-oriented chatbot” operationally to refer to open-domain LLM assistants that are predominantly used for task-like purposes (e.g., information seeking, writing, light troubleshooting) in practice [10]—even though they are not classic, bounded-domain task bots with explicit intent/slot pipelines. The emphasis is on observed usage patterns (task completion as a common goal), not on the traditional system architecture; accordingly, our comparisons focus on how usage context (chit-chat vs. task-leaning assistant use) relates to the incidence and timing of harmful speech.

The conversational-AI ecosystem spans multiple agent types. Over the past two years, LLM-based assistants (e.g., ChatGPT, Gemini, and Claude) have attracted large user bases; at the same time, there remains strong demand for chit-chat/relationship-oriented agents (e.g., Character.AI, SimSimi, Replika). For example, as of October 2025, Replika reports roughly 40 million users [12], Character.AI has 20+ million monthly active users worldwide [13], and SimSimi reports a cumulative user base of approximately 450 million [14], indicating substantial interest in affective, open-ended conversation and parasocial “relationships” with chatbots. By contrast, within LLM assistant services (e.g., OpenAI and Anthropic), self-expression/affective exchanges—including casual chit-chat—account for a relatively small share of overall traffic (on the order of 2.9–4.3%), with usage dominated by task execution (information seeking, writing, and work/study assistance) [15,16]. Taken together, this landscape motivates an analytic distinction between task-oriented assistants and chit-chat agents and supports examining how harmful-speech incidence and timing differ across these two interaction contexts.

Against this backdrop, using production logs and real user–chatbot dialogues, this study compares the type-specific frequencies and temporal distributions (e.g., time-of-day) of harmful utterances (e.g., profanity, insults, threats). By identifying distributional patterns and contextual correlates, we (1) extend social-media-centric research on hate and harmful language to the user–chatbot setting, and (2) provide empirical evidence to guide operations, such as time-aware safety policies. We further contrast LLM-based chatbots with chit-chat-oriented chatbots to illuminate how harmful speech varies by chatbot type. Concretely, we address three questions: (RQ1) Into which fine-grained categories can harmful speech in chatbot–user dialogues be classified? (RQ2) What are the temporal (e.g., time-of-day) distributions of harmful speech? (RQ3) Do patterns in type, frequency, and temporal trends differ by chatbot type (LLM-based vs. chit-chat)?

Answering these questions advances both science and practice. RQ1 establishes a fine-grained, operational taxonomy of harmful speech in user–chatbot dialogues, improving measurement fidelity and enabling targeted interventions (e.g., distinct handling for insults vs. identity attacks). RQ2 characterizes time-of-day rhythms, which directly inform time-aware safety policies—from adaptive thresholds and warning prompts to staffing and monitoring schedules. RQ3 links platform context (task-leaning LLM assistants vs. play-oriented chit-chat) to the level and mix of harm, clarifying how interaction goals shape risk

and guiding design trade-offs (persona, guardrails, defaults). Together, these motivations situate our study as a bridge between social-media evidence and 1:1 conversational AI, yielding actionable guidance for safer deployments.

To answer these questions, we analyze 8,705,959 English ChatGPT–user interaction turns and 610,037 SimSimi turns; each dialogue is annotated for the presence of harmful speech and categorized by type, and we compute time-of-day incidence. The findings yield policy and design implications for strengthening the reliability and accountability of LLM-based chatbots.

2. Related Work

2.1. Addressing Harmful Expressions in Chatbots

Whereas AI-based toxicity detectors for news comments or social media primarily fail by not filtering some harmful content, LLM-based chatbots are active generators of language, so comparable safety failures can create and amplify harmful expressions. A growing body of work documents that generative language models can produce biased, hateful, or toxic outputs—even from seemingly innocuous prompts—and that such model behaviors carry broader societal risks [4,17]. These risks are further heightened in interactive settings, where users iteratively engage the model and can jailbreak or otherwise circumvent safety guardrails; large-scale “in-the-wild” studies show systematic, transferable prompt attacks against multiple LLMs [18]. Beyond technical evaluations, recent surveys synthesize how LLMs can amplify misuse and harmful content generation in practice, underscoring the need for robust safety-by-design approaches [19]. The Iruda incident in Korea illustrates these risks in a real deployment: due to biased conversational data and ongoing interactions, the chatbot itself generated discriminatory and hateful utterances toward protected groups, prompting public backlash and ethics debates on data governance and model behavior [20]. Taken together, the emergence of generative models reframes harmful-language mitigation from a static classification problem to a dynamic interaction problem. It becomes imperative to closely observe LLM utterance behavior and to characterize temporal patterns of harmful expressions in order to advance AI ethics and safety.

2.2. Harmful Speech Types and Incidence in LLM Chatbots

A small but growing body of literature has begun to quantify what kinds of harmful utterances LLM chatbots produce and how often they arise. Using over half a million generations from ChatGPT, Deshpande et al. [21] systematically showed that toxicity rates and targets shift with conversational “persona” settings—with toxicity rising up to sixfold under certain personas and specific entities being disproportionately targeted—highlighting category-level patterns such as insults, identity-directed abuse, and threats in model outputs. Complementing model-centric evaluations, a large in-the-wild log study of 6.25 M SimSimi user–bot exchanges documented harmful speech prevalence and topical skew (e.g., sexually explicit content) in real chit-chat interactions and released tooling tailored to open-domain dialogue, offering distributional evidence by content type rather than only aggregate “toxicity” scores [3]. From a security perspective, Weeks et al. demonstrated toxicity-injection attacks against deployed open-domain chatbots, showing that pipelines can be manipulated to amplify toxic responses and alter the mix of generated harms—evidence that incidence and type distributions are sensitive to interaction dynamics and adversarial pressure [22]. Taken together, these studies establish that (i) LLM chatbots can emit distinct harmful categories (insult, identity attack, threat, etc.) in measurable proportions; (ii) real-world, anonymous chit-chat settings exhibit different frequency profiles than curated benchmarks; and (iii) interactive or adversarial contexts can shift those profiles.

However, prior work has not jointly compared type-specific incidence across LLM-based vs. chit-chat chatbots in production logs, nor has it quantified temporal (time-of-day) variation within and across types—gaps our study addresses directly [21].

2.3. Temporal Variation in Harmful Expressions

Our core hypothesis—that the incidence of harmful expressions varies by time of day—builds on extensive evidence that online behavior and affect exhibit robust diurnal structure. Large-scale analyses of hundreds of millions of tweets show clear circadian rhythms in expressed mood, with positive affect peaking in the morning and declining toward the evening, and seasonality aligned with daylength, implying sleep–circadian mechanisms behind online affective language [23]. Similar diurnal regularities in social media streams have been observed for diverse real-world activities and topics, underscoring the value of time-series approaches to online language [24].

Mechanistically, late-evening and night-time windows are plausibly higher-risk because sleep loss and circadian misalignment degrade inhibitory control and heighten emotional reactivity—conditions linked to impulsive, disinhibited responding. Laboratory and neuroimaging studies show that sleep deprivation reduces response inhibition and amplifies limbic/reward reactivity to emotional stimuli [25]. In parallel, the “online disinhibition effect” explains how anonymity, asynchronicity, and attenuated social cues can loosen normative restraints in computer-mediated communication, increasing the likelihood of hostile or extreme expression when self-regulation is compromised [26].

Observational studies further connect night-time social-media use with poorer well-being and sleep outcomes. A recent cohort study linking Twitter (X) activity logs with mental-health measures found that posting predominantly between 23:00–05:00 is associated with meaningfully lower well-being; results were published in *Scientific Reports* [27]. Complementary epidemiological work reports that greater social-media use correlates with shorter sleep duration and poorer sleep quality in adolescents, consistent with a bidirectional relationship between digital media and sleep/affect [28].

Although most toxicity research emphasizes detection accuracy rather than time [3,9], recent measurement studies and modeling work highlight the temporal dimension of antisocial behavior online and the challenge of temporal generalization (i.e., models trained at one time underperform at others). Analyses of toxic accounts on Reddit over multi-month horizons document dynamic patterns in toxic posting, while evaluations of hate-speech detectors show performance drift across time, calling for temporally robust—and potentially time-aware—moderation strategies [29].

Taken together, prior evidence supports a temporal-affective hypothesis for harmful expressions online: harmful language often reflects high-arousal negative affect (e.g., anger, contempt); population-level negative affect tends to rise over the day; night-time conditions exacerbate impulsivity and disinhibition; therefore, incidence should plausibly peak during late evening and night. This, in turn, motivates proactive, time-aware safety interventions (e.g., adaptive thresholds, consent friction for sensitive topics, de-escalation prompts) that complement purely reactive filters triggered post-input or post-generation.

3. Methods

3.1. Dataset

We conduct our study on two real-world chatbot datasets: (1) WildChat Wild, a corpus of user–ChatGPT dialogues from an LLM-based chatbot, and (2) SimSimi, a corpus of user–bot dialogues from a chit-chat-oriented chatbot [8]. WildChat Wild is an anonymized dataset collected by offering users free access to ChatGPT (GPT-3.5/4) and logging their real interactions for research under explicit consent; it provides conversa-

tion texts, timestamps, and request headers while retaining IP information only at the country level and removing personally identifiable information (PII) [30]. In its public release, WildChat comprises roughly 1 million conversations and 2.5 million interaction turns, with multilingual, multi-region timestamp metadata that enable temporal analyses <https://huggingface.co/datasets/allenai/WildChat-1M-Full> (accessed on 4 December 2025).

For analytic consistency, we subset WildChat to dialogues whose metadata language attribute is “English,” and whose country is the United States, United Kingdom, Canada, Australia or New Zealand. To reduce annotation ambiguity, we exclude utterances longer than 4096 tokens and records with missing or corrupted timestamps. We chose 4096 tokens as a pragmatic upper bound for three reasons: extremely long turns often reflect concatenations or copy–paste artifacts that inflate ambiguity for sentence-level toxicity scoring [31]; prior LLM practice commonly treats 4 K tokens as a stability boundary for single-pass text (legacy context windows), beyond which model behavior and API scores can become less reliable [32]; and in our data, such long turns are rare outliers (a small fraction of all utterances), so excluding them reduces noise with minimal impact on coverage. The final sample includes 610,837 utterances.

SimSimi is a long-running chit-chat chatbot launched in 2002, serving dozens of languages and hundreds of millions of users worldwide; unlike goal-oriented assistants, it is designed primarily for open-ended, entertainment-focused conversation [8]. SimSimi operates in an application environment that emphasizes anonymity, and it famously offers a Teach feature whereby users can directly supply responses that the bot may immediately adopt—an affordance documented in prior technical descriptions and reviews [3,33]. Such participatory and rapid-update mechanics have been linked to broader adoption and community engagement, but they also correlate with a playful style of interaction (e.g., joking, teasing, role-play, profanity) and higher rates of direct or provocative expressions in open, anonymous settings—consistent with classic accounts of the online disinhibition effect in computer-mediated communication [26]. Consistent with this, a large-scale analysis of SimSimi logs reported that sex-related topics constitute roughly 48% of harmful speech, far outpacing other subjects such as politics or sports [3].

In this study, we analyze SimSimi–user dialogues collected from 2019 to 2022. The application logs country of use, dialogue content, timestamps, and device ID. For analytic consistency, WildChat is limited to English-language dialogues. For SimSimi, we apply the same English-language restriction and include the United States, United Kingdom, Canada, and Ireland. The country sets are not identical due to data availability in the research releases and sample-size considerations (SimSimi does not provide sufficient coverage for Australia/New Zealand in our window). The final SimSimi dataset comprises 8,785,859 user utterances.

3.2. Analysis Methods

To classify whether a user–chatbot utterance is harmful and, if so, which subtype it reflects, we use Google/Jigsaw’s Perspective API <https://developers.perspectiveapi.com/> (accessed on 4 December 2025), a machine-learning service that scores the perceived toxicity of text on a continuous 0–1 scale and provides subtype-specific attributes. Concretely, we query the attributes Toxicity (general abusive/hostile tone likely to drive others out of a conversation), Insult (personal attacks), Identity-Attack (derogation targeting protected characteristics), and Threat (explicit or implicit harm). Our taxonomy is grounded in prior HCI and conversational-agent research that typologizes verbal abuse toward chatbots (e.g., insults, threats, and swearing/profanity) [2,34]. Building on these precedents, we operationalized categories using Perspective API attributes, which align well with the literature and enable non-mutually exclusive, sentence-level labeling. The API’s attribute

design and scoring scheme are documented in the developer materials and prior research on large-scale moderation using Wikipedia/New York Times comment corpora. Public reports note that Perspective has been piloted in production comment pipelines (e.g., The New York Times) to triage or filter abusive content [35].

We map continuous scores to binary labels using a single threshold consistent with API guidance and empirical practice. While the Perspective documentation notes that stricter cutoffs (0.9–0.95) may be appropriate for certain filtering scenarios, many social-computing studies adopt thresholds in the 0.7 range to balance recall and precision for measurement [9,36–38]. Higher cutoffs yield the same qualitative ordering but attenuate rates and significance due to sparsity. Accordingly, we fix the cutoff at 0.7 for all analyses and report time-of-day distributions based on this criterion.

To probe temporal structure, we group local timestamps into five dayparts adapted from the literature on diurnal affect in social media: morning (06:00–09:00), daytime (09:00–16:00), evening (16:00–21:00), night (21:00–24:00), and dawn (00:00–06:00) [23]. Similar four- or five-bin time-of-day partitions are widely used to analyze digital behavior in social-media-based affect and mental-health research [39,40]. This binning follows evidence of robust circadian rhythms in linguistic affect—positive affect peaking in the morning and declining toward evening—observed in large-scale analyses of hundreds of millions of tweets across cultures [41]. Our aim in this paper is to align temporal bins with coarse-grained patterns of human activity (e.g., work hours and late-night use) rather than with exact clock time. We compute per-bin incidence rates and compare distributions across chatbot types.

To minimize time zone misalignment, we normalize timestamps to local time using the country metadata in the logs, converting each UTC timestamp to local clock time by applying the appropriate country-level time-zone offset, and then apply country-specific daypart bins so that dawn/morning/afternoon/evening/night are defined in local clock time rather than UTC. We recognize that individual sleep–wake schedules vary across time zones, ages, and occupations; our approach follows established practice in diurnal analyses of online language (using population-level local-time bins).

In this study, Perspective API attributes are treated as non-mutually exclusive labels. Toxicity serves as an umbrella construct capturing rude, hostile, or conversation-disruptive tone, while Insult, Identity Attack, Threat, Profanity, and Severe Toxicity denote more specific patterns. For each user utterance, the Perspective API returns six scores between 0 and 1: Toxicity, Insult, Identity Attack, Threat, Profanity, and Severe Toxicity. We apply a fixed threshold of 0.7 to each attribute independently and allow co-occurrence across attributes. For example, if the utterance “you are an idiot” receives scores at or above 0.7 for “Toxicity” and “Insult” while the other attributes fall below 0.7, we increment the counts for both “Toxicity” and “Insult”; the remaining categories are not incremented for that utterance. For each daypart, the incidence rate of a given attribute is the number of utterances in that daypart with a score of at least 0.7 for the attribute divided by the total number of utterances in that daypart. All between-platform comparisons (ChatGPT/WildChat versus SimSimi) are conducted on these rates computed within each daypart.

3.3. Ethical Consideration

Given the sensitivity of harmful speech, we designed all stages of this study—access, preprocessing, analysis, and reporting—to minimize risk while preserving scientific utility across both WildChat and SimSimi. Work with SimSimi was conducted in collaboration with SimSimi Inc. under a clearly bounded research scope and access controls, limited to de-identified, aggregate artifacts appropriate for population-level inference. We neither inferred nor tracked individual users, and no attempts at re-identification were made.

For lawful basis and informed participation, WildChat records were collected expressly for research via a two-step consent flow that notified participants of logging and anonymized research use prior to engagement [30]. SimSimi interactions occur under the service’s Terms of Use, which grant SimSimi Inc. a transferable, worldwide, royalty-free license to use submitted content for service improvement and research [42]; our analysis operated within that policy framework and with the provider’s cooperation.

Data minimization and de-identification guided all preprocessing. Personally identifiable information (e.g., names, emails, phone numbers) was not retained in either dataset; IP information was available only at the country level. We did not compute user-level aggregates, link sessions via persistent identifiers, or reconstruct longitudinal histories. Device identifiers present in SimSimi logs were not used for tracking or analysis. All reporting is at the aggregate level (counts, rates, confidence intervals, model coefficients), and we do not release raw dialogue text. Transparency to users and the research community was prioritized. SimSimi Inc. publicly communicates how conversational data inform service quality and safety (e.g., blog-style research notes), and our collaboration drew on these materials to clarify data use [8,42]. WildChat is available for research under clearly stated conditions that require affirmative acknowledgment of research use before access. [30] <https://huggingface.co/datasets/allenai/WildChat-1M-Full> (accessed on 4 December 2025).

4. Results

This section summarizes the composition of harmful speech and its temporal distribution across datasets. Table 1 shows that, in both corpora, the majority of harmful content is concentrated in Toxicity (general abusive tone) and Profanity. In SimSimi, Toxicity is 4.02% and Profanity is 2.98%, together accounting for roughly 7 percentage points of all utterances labeled harmful. The remaining categories are less prevalent: Insult 0.55%, Threat 0.09%, Identity Attack 0.08%, and Severe Toxicity <0.01%. WildChat exhibits the same ranking at lower levels—Toxicity 0.83% and Profanity 0.48%—followed by Insult 0.06%, Identity Attack 0.02%, Threat 0.01%, and Severe Toxicity <0.01%. Aggregating across categories, the overall harmful-speech rate is 7.72% in SimSimi (672,214 of 8,785,959 utterances) and 1.39% in WildChat (8494 of 610,837 utterances), a difference of about a factor of 5.6. Category wise, SimSimi exceeds WildChat most strongly in Insult and Threat, followed by Profanity, Toxicity, and Identity Attack.

Table 1. Category-wise harmful speech counts and rates in SimSimi and WildChat (two-proportion z-test *p*-values). Percentages are relative to each platform’s total turns. *p*-values from two-proportion z-tests; very small values are capped at $p < 0.001$.

Category	SimSimi		WildChat		<i>p</i>
	Count	Rate	Count	Rate	
Toxicity	350,015	4.02%	5080	0.83%	<0.001
Profanity	259,677	2.98%	2900	0.48%	<0.001
Insult	47,905	0.55%	356	0.06%	<0.001
Iden._Attack	6934	0.08%	99	0.02%	<0.001
Threat	7537	0.09%	45	0.01%	<0.001
S._Toxicity	146	0.00%	14	0.00%	=0.248
Harmful (any)	672,214	7.72%	8494	1.39%	<0.001
Total turns	8,785,959		610,837		

To assess statistical significance, we conducted two-proportion z tests for each category, correcting for multiple comparisons with the Holm–Bonferroni method at $\alpha = 0.05$. Except for Severe Toxicity, SimSimi’s rates are significantly higher than WildChat’s for all

categories (all adjusted $p < 0.001$). Estimated risk ratios are largest for Threat, Insult, and Profanity, at approximately 11.6, 9.4, and 6.2, respectively; Toxicity and Identity Attack also show meaningful elevation. Severe Toxicity remains too rare in both corpora—to support a powered test, and the between-corpus difference is not significant. Taken together, these contrasts are consistent with the hypothesis that differences in conversational context—an anonymous, play-oriented chit-chat environment versus a task- or goal-oriented LLM setting—shape both the level and the mix of harmful speech.

Figure 1 compares daypart-level totals by summing across categories and normalizing by all utterances in each time block. Both corpora display a common diurnal rhythm characterized by a low in the morning and higher rates at night and at dawn. In SimSimi, the dawn period (00:00–06:00) reaches the highest incidence at about 2.61%, followed by evening at about 1.70%, night at about 1.32%, afternoon at about 1.22%, and morning at about 0.87%. WildChat shows the same ordering at a lower scale: dawn at about 0.45%, then afternoon at about 0.35%, evening at about 0.27%, night at about 0.18%, and morning at about 0.14%. Thus, diurnal variation is evident in both datasets, with SimSimi exhibiting both higher levels and larger swings over the day.

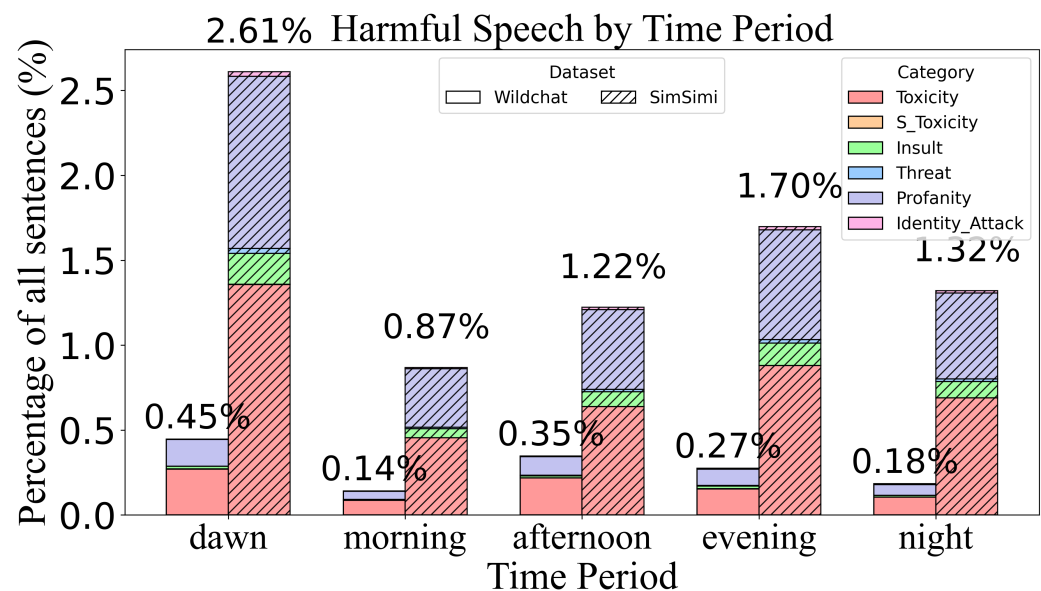
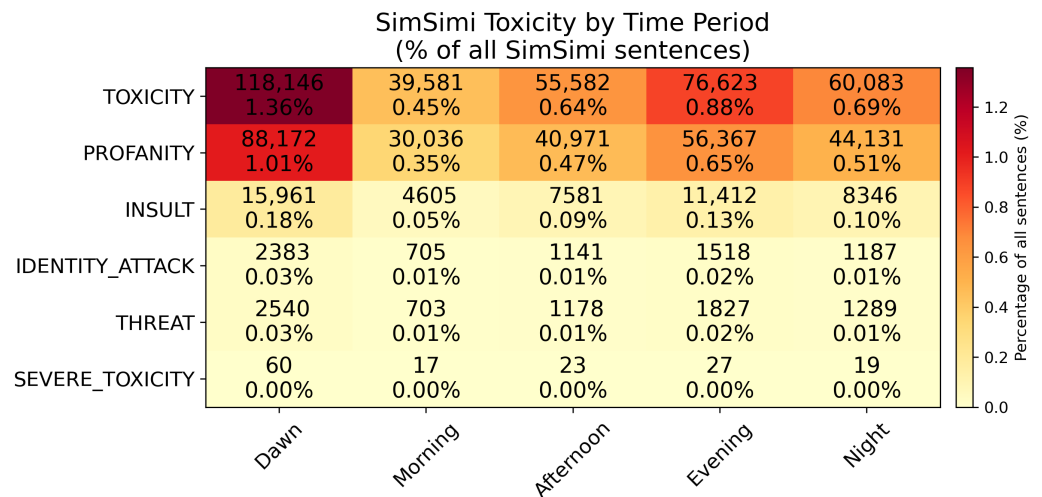
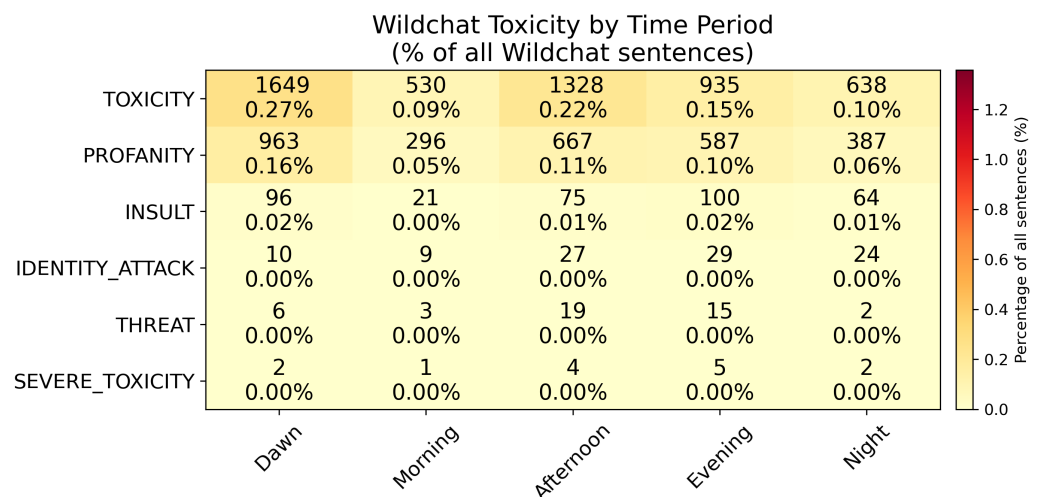


Figure 1. Temporal Patterns of Harmful Speech (Stacked by Category).

Figure 2 details the joint distribution by category and daypart. In SimSimi, Toxicity and Profanity dominate across all time blocks, with dawn showing a pronounced increase (Toxicity roughly 1.36%; Profanity roughly 1.01%). Evening remains elevated (around 0.8–0.9% for the combined core categories), and night is moderately high, indicating a concentration of general abuse and coarse language during hours typically associated with heightened arousal and reduced self-regulation. Insult remains low-frequency across dayparts—around 0.1%—with a relative uptick at night, while Identity Attack and Threat stay sparse at roughly 0.01–0.03%. WildChat exhibits the same qualitative hierarchy—Toxicity, then Profanity, then the remaining categories—though at substantially lower rates; dawn shows a peak again (Toxicity about 0.27%; Profanity about 0.16%), and afternoon shows a secondary peak for Toxicity at about 0.22%. Insult, Identity Attack, and Threat remain near 0.00–0.02% across time blocks.



(a) SimSimi harmful speech by time period (% of all SimSimi sentences).



(b) WildChat harmful speech by time period (% of all WildChat sentences).

Figure 2. Temporal distribution of harmful speech across two datasets.

Figure 3 summarizes these patterns in terms of relative risk ratios between WildChat and SimSimi for each category–daypart cell, where values below 1 indicate that WildChat is safer than SimSimi. Across most categories and times of day, the ratios lie well under 1 (e.g., 0.06–0.34× for Identity Attack and 0.02–0.23× for Threat), confirming that WildChat generally exhibits substantially lower rates of harmful content. The main exception is Severe Toxicity, for which WildChat shows elevated risk in the afternoon, evening, and night (2.48×, 2.64×, and 1.50×, respectively), suggesting that while overall levels are lower, the most extreme toxic behaviors remain a relatively concentrated concern in specific time windows.

In summary, both datasets are structurally dominated by general abuse and profanity, rare in severe forms, and display consistent temporal organization with morning lows and dawn or evening highs. SimSimi’s absolute rates and diurnal amplitude are markedly larger than WildChat’s, reinforcing the interpretation that platform context and user intent are linked to the incidence and composition of harmful speech.

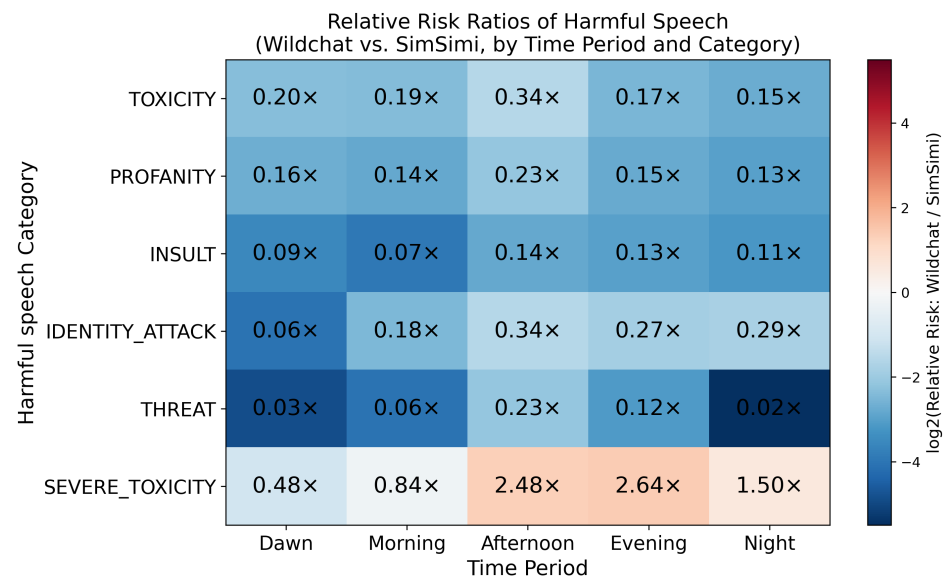


Figure 3. Heatmap of relative risk ratios of harmful speech categories between Wildchat and SimSimi across time periods. Each cell shows the relative risk $RR = p_{wildchat}/p_{simsimi}$ for a given harm type and time period, where p is the proportion of all sentences in the dataset. Colors represent $\log_2(RR)$: red cells indicate categories that are more prevalent in Wildchat, whereas blue cells indicate categories more prevalent in SimSimi.

5. Discussion

Our findings align with the study’s hypothesis that harmful speech concentrates in specific types and times of day. Across both corpora, Toxicity and Profanity dominate, whereas Insult, Identity Attack, and Threat are comparatively rare. This pattern is consistent with the possibility that interacting with a nonhuman counterpart suppresses direct, person-targeted aggression [43]. Prior work shows that when the target is not human, people tend to ascribe lower moral patience and feel less guilt [44], which in turn shifts both the choice of target and the form of expression—away from insults or identity-based attacks and toward more generalized abuse or coarse language [43]. For example, experimental evidence indicates that when an interlocutor is explicitly framed as an AI, participants report lower guilt for unethical actions relative to human-targeted controls, suggesting that perceived nonhuman status and social distance can weaken the motivation for interpersonal attacks [45]. Moreover, HCI studies that typologize verbal abuse toward conversational agents (e.g., insults, threats, swearing) and manipulate interactional context (such as response style) demonstrate that the perceived humanness of the target, its social proximity, and its response behavior systematically shape the mix and intensity of abusive language [2].

The marked gap in overall prevalence—WildChat substantially lower than SimSimi at a 0.7 threshold—also appears consistent with contextual differences between the platforms. For WildChat, several mechanisms plausibly depress harmful content: (i) two-step research consent and explicit notices that may induce self-censorship and observer effects; (ii) awareness of public research collection, which can elicit social-desirability behavior; and (iii) preprocessing and de-identification that further reduce residual sensitive strings. By contrast, SimSimi’s open, anonymous, and play-oriented environment creates conditions under which generalized abuse and profanity—particularly overnight—are more likely to surface. Building on prior work, anonymity has been shown to weaken normative restraints and promote online disinhibition, increasing the likelihood of rude or aggressive expression [26]; likewise, playful or entertainment-oriented contexts tend to overrepresent blunt, teasing, and jocular utterances [46], creating conditions under which such styles

are more readily produced and tolerated. Taken together, these findings reinforce the interpretation that anonymity and play orientation, in combination with a nonhuman interlocutor, shift abusive expression away from explicitly interpersonal attacks toward generalized toxicity and profanity, while also elevating the overall propensity for disinhibited speech [46]. These contrasts support the broader claim that platform context meaningfully shapes not only the level of harmful speech but also its composition, even under a common thresholding scheme.

Based on our findings, we derive the following design implications. First, time-aware guardrails appear promising: during dawn and evening periods, platforms can modestly increase filter sensitivity, add pre-conversation warnings or consent friction for sensitive topics, and default to de-escalation prompts or mild rate limits, while relaxing interventions during lower-risk morning and daytime windows to preserve usability. Second, surveillance priorities should reflect type rarity and risk. For SimSimi, real-time monitoring and warnings for Toxicity/Profanity in dawn periods are likely to be cost-effective; for WildChat, focused monitoring of Toxicity in dawn and afternoon may suffice. Although Insult, Threat, and Identity Attack are rare, their harm potential is high; precision-oriented detectors that are sensitive to even small counts remain necessary.

6. Limitations and Future Work

This study is subject to several limitations. First, we binarize Perspective attribute scores using a single 0.7 threshold. While this ensures comparability, it cannot fully eliminate risks of over- or under-detection, particularly given attribute-specific score distributions and known API biases. Second, our analyses are restricted to English-language corpora with country-level metadata, precluding a more granular examination of linguistic, cultural, or regional variations. Third, the WildChat dataset's explicit research context (including disclosure and consent) may not fully represent interactions on commercial or relationship-oriented LLM services (e.g., character-based chatbots), which raises considerations for external validity. Fourth, our harmful speech counts are computed at the single-turn level and thus do not capture dialogic dynamics (e.g., escalation, repair, or multi-turn context). As a result, utterances that are implicitly harmful in context—but not explicitly toxic at the sentence level—may be undercounted. Fifth, we do not model platform-side factors such as differences in metadata granularity or UI/UX design between the two services; our analyses focus solely on the presence of harmful speech in user-originated messages, which may omit effects mediated by interface, affordances, or system replies. Another limitation is the lack of temporal alignment between the two datasets. This reflects the difficulty of obtaining research-grade chatbot logs—especially for chit-chat systems—where 1:1, privacy-sensitive interactions (often including romantic/companionship content) make public releases exceedingly rare. The SimSimi corpus is thus a valuable but atypical resource and, as noted, is not contemporaneous with WildChat. If research access to Replika or Character.AI logs becomes available, future work will compare matched time windows against task-oriented LLM data. Finally, our observational design allows us to identify associations, not to establish causal links between time of day and the incidence of harmful speech.

Future work should proceed along several directions. First, we recommend broadening the platform spectrum. Analysis should incorporate relationship-oriented LLMs (e.g., Character.AI) and other domain-specific agents to establish context-specific baselines and more robustly assess external validity. As a complementary extension, a cross-domain replication on public social platforms—using matched time windows, identical daypart binning, and comparable harm labels—would test whether the diurnal and type-mix patterns observed in 1:1 chatbot dialogs generalize to many-to-many settings and help separate

platform effects from population-level dynamics [47–49]. Second, future studies should diversify the harm detection models used. To mitigate single-API dependency, results should be cross-validated against other toxicity classifiers, including open-source models, to evaluate the consistency and robustness of the findings. Finally, prior work indicates that the ways people discuss negative emotions with chatbots can vary by culture. In future work, we plan to compare the types and incidence of harmful speech across additional languages and cultural contexts to assess the generalizability of our findings.

7. Conclusions

This work provides a production-scale view of harmful language in user–chatbot interactions, linking what appears with when it occurs and how platform context shapes both. Three themes emerge: harmful speech is substantially more prevalent in the chit-chat setting; across platforms, Toxicity and Profanity dominate while Insult, Identity Attack, and Threat remain rare; and harmful language follows a clear diurnal rhythm that concentrates around dawn relative to other periods. Interpreting these patterns through interactional and governance factors suggests that anonymity and play orientation, together with a nonhuman interlocutor, shift expression away from explicitly interpersonal attacks toward generalized abuse. These findings furnish an actionable baseline for time-aware, platform-specific safety.

Author Contributions: O.K. contributed to writing—original draft and editing, data curation, investigation, methodology, visualization, and conceptualization. H.Y. contributed to writing—review and editing, formal analysis, methodology, and investigation. H.C.: contributed to writing—review and editing, visualization, supervision, and conceptualization. J.P.: co-corresponding author. contributed to writing—review and editing, project administration, supervision, and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: J.P. and H.C. are supported by the research grant of the Gyeongsang National University in 2024.

Data Availability Statement: Public data: WildChat-1M-Full (Hugging Face; <https://huggingface.co/datasets/allenai/WildChat-1M-Full> (accessed on 4 December 2025)). No new proprietary data were collected. For the SimSimi corpus, to safeguard privacy, access will be provided to qualified researchers for research purposes upon reasonable request.

Acknowledgments: We gratefully acknowledge SimSimi for granting research access to the dataset. In preparing this manuscript, we used a generative AI tool (Chatgpt 5) solely for language refinement (e.g., wording and clarity). The tool was not used to generate ideas, interpret findings, or produce references. All AI-assisted text was reviewed, edited, and verified by the authors before inclusion.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Si, W.M.; Backes, M.; Blackburn, J.; De Cristofaro, E.; Stringhini, G.; Zannettou, S.; Zhang, Y. Why So Toxic? Measuring and Triggering Toxic Behavior in Open-Domain Chatbots. In Proceedings of the CCS '22: 2022 ACM SIGSAC Conference on Computer and Communications Security, Los Angeles, CA, USA, 7–11 November 2022; pp. 2659–2673. [\[CrossRef\]](#)
2. Chin, H.; Molefi, L.W.; Yi, M.Y. Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse. In Proceedings of the CHI '20: 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–13. [\[CrossRef\]](#)
3. Song, H.; Hong, J.; Jung, C.; Chin, H.; Shin, M.; Choi, Y.; Choi, J.; Cha, M. Detecting offensive language in an open chatbot platform. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy, 20–25 May 2024; pp. 4760–4771.
4. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual, 3–10 March 2021; pp. 610–623. [\[CrossRef\]](#)

5. Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Papernot, N.; Anderson, R.; Gal, Y. AI models collapse when trained on recursively generated data. *Nature* **2024**, *631*, 755–759. [CrossRef] [PubMed]
6. Fortuna, P.; Nunes, S. A survey on automatic detection of hate speech in text. *Acm Comput. Surv. (CSUR)* **2018**, *51*, 1–30. [CrossRef]
7. Jahan, M.S.; Oussalah, M. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing* **2023**, *546*, 126232. [CrossRef]
8. Chin, H.; Song, H.; Baek, G.; Shin, M.; Jung, C.; Cha, M.; Choi, J.; Cha, C. The potential of chatbots for emotional support and promoting mental well-being in different cultures: Mixed methods study. *J. Med. Internet Res.* **2023**, *25*, e51712. [CrossRef]
9. Lin, Z.; Wang, Z.; Tong, Y.; Wang, Y.; Guo, Y.; Wang, Y.; Shang, J. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. *arXiv* **2023**, arXiv:2310.17389.
10. Ni, J.; Young, T.; Pandelea, V.; Xue, F.; Cambria, E. Recent advances in deep learning based dialogue systems: A systematic survey. *Artif. Intell. Rev.* **2023**, *56*, 3055–3155.
11. Roller, S.; Boureau, Y.L.; Weston, J.; Bordes, A.; Dinan, E.; Fan, A.; Gunning, D.; Ju, D.; Li, M.; Poff, S.; et al. Open-domain conversational agents: Current progress, open problems, and future directions. *arXiv* **2020**, arXiv:2006.12442. [CrossRef]
12. Replika CEO Eugenia Kuyda Launches Wabi. Business Insider. 2025. Available online: <https://www.businessinsider.com/replika-ceo-eugenia-kuyda-launch-wabi-2025-10> (accessed on 4 December 2025).
13. The New York Times. Character.ai Abandons Making AI Models after \$2.7bn Google Deal. 2024. Reports 20 Million Monthly Active Users. Available online: <https://www.ft.com/content/f2a9b5d4-05fe-4134-b4fe-c24727b85bba> (accessed on 4 December 2025).
14. SimSimi Inc. SimSimi Official Website. 2025. Available online: <https://simsimi.kr/> (accessed on 4 December 2025).
15. Anthropic. How People Use Claude for Support, Advice, and Companionship. 2025. Available online: <https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship> (accessed on 4 December 2025).
16. Chatterji, A.; Cunningham, T.; Deming, D.J.; Hitzig, Z.; Ong, C.; Shan, C.Y.; Wadman, K. *How People Use Chatgpt*; Technical Report; National Bureau of Economic Research: Cambridge, MA, USA, 2025.
17. Gehman, S.; Gururangan, S.; Sap, M.; Choi, Y.; Smith, N.A. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16–20 November 2020*; Cohn, T., He, Y., Liu, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 3356–3369. [CrossRef]
18. Shen, X.; Chen, Z.; Backes, M.; Shen, Y.; Zhang, Y. “Do Anything Now”: Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In *Proceedings of the CCS ’24: 2024 on ACM SIGSAC Conference on Computer and Communications Security*, Salt Lake City, UT, USA, 14–18 October 2024; pp. 1671–1685. [CrossRef]
19. Zhou, W.; Zhu, X.; Han, Q.L.; Li, L.; Chen, X.; Wen, S.; Xiang, Y. The Security of Using Large Language Models: A Survey with Emphasis on ChatGPT. *IEEE/CAA J. Autom. Sin.* **2025**, *12*, 1–26. [CrossRef]
20. Kim, Y.; Kim, J.H. The impact of ethical issues on public understanding of artificial intelligence. In *Proceedings of the 23rd HCI International Conference, HCII 2021, Virtual Event, 24–29 July 2021*; Springer: Cham, Switzerland, 2021; pp. 500–507.
21. Deshpande, A.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; Narasimhan, K. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Singapore, 2023; pp. 1236–1270. [CrossRef]
22. Weeks, C.; Cheruvu, A.; Abdullah, S.M.; Kanchi, S.; Yao, D.; Viswanath, B. A first look at toxicity injection attacks on open-domain chatbots. In *Proceedings of the 39th Annual Computer Security Applications Conference*, Austin, TX, USA, 4–8 December 2023; pp. 521–534.
23. Golder, S.A.; Macy, M.W. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* **2011**, *333*, 1878–1881. [CrossRef]
24. Grinberg, N.; Naaman, M.; Shaw, B.; Lotan, G. Extracting diurnal patterns of real world activity from social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Cambridge, MA, USA, 8–11 July 2013; Volume 7, pp. 205–214.
25. Anderson, C.; Platten, C.R. Sleep deprivation lowers inhibition and enhances impulsivity to negative stimuli. *Behav. Brain Res.* **2011**, *217*, 463–466. [CrossRef]
26. Suler, J. The online disinhibition effect. *Cyberpsychol. Behav.* **2004**, *7*, 321–326. [PubMed]
27. Joinson, D.; Haworth, C.M.; Simpson, E.; Cristianini, N.; Di Cara, N.H.; Davis, O.S. Active night-time tweeting is associated with meaningfully lower mental wellbeing in a UK birth cohort study. *Sci. Rep.* **2025**, *15*, 34301. [CrossRef] [PubMed]
28. Scott, H.; Biello, S.M.; Woods, H.C. Social media use and adolescent sleep patterns: Cross-sectional findings from the UK millennium cohort study. *BMJ Open* **2019**, *9*, e031161. [CrossRef] [PubMed]
29. Kumar, D.; Hancock, J.; Thomas, K.; Durumeric, Z. Understanding the behaviors of toxic accounts on reddit. In *Proceedings of the ACM Web Conference 2023*, Austin, TX, USA, 30 April–4 May 2023; pp. 2797–2807.
30. Zhao, W.; Ren, X.; Hessel, J.; Cardie, C.; Choi, Y.; Deng, Y. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv* **2024**, arXiv:2405.01470.

31. Bell, S.; Meglioli, M.C.; Richards, M.; Sánchez, E.; Ropers, C.; Wang, S.; Williams, A.; Sagun, L.; Costa-jussà, M.R. On the Role of Speech Data in Reducing Toxicity Detection Bias. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Albuquerque, NM, USA, 29 April–4 May 2025; Chiruzzo, L., Ritter, A., Wang, L., Eds.; Association for Computational Linguistics: Albuquerque, NM, USA, 2025; pp. 1454–1468. [\[CrossRef\]](#)
32. IBM. Context Window (Think Blog Topic Page). 2025. Available online: <https://www.ibm.com/think/topics/context-window> (accessed on 4 December 2025).
33. Shin, M.; Chin, H.; Song, H.; Choi, Y.; Choi, J.; Cha, M. Context-Aware Offensive Language Detection in Human-Chatbot Conversations. In *Proceedings of the 2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Bangkok, Thailand, 18–21 February 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 270–277.
34. Xu, J.; Ju, D.; Li, M.; Boureau, Y.L.; Weston, J.; Dinan, E. Bot-Adversarial Dialogue for Safe Conversational Agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021*; Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 2950–2968. [\[CrossRef\]](#)
35. Greenberg, A. Now Anyone Can Deploy Google’s Troll-Fighting AI. WIRED (Security). 2017. Available online: <https://www.wired.com/2017/02/googles-troll-fighting-ai-now-belongs-world/> (accessed on 4 December 2025).
36. Lima, L.H.Q.; Pagano, A.S.; da Silva, A.P.C. Toxic Content Detection in online social networks: A new dataset from Brazilian Reddit Communities. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese—Volume 1*, Santiago de Compostela, Spain, 14–15 March 2024; pp. 472–482.
37. Saveski, M.; Roy, B.; Roy, D. The structure of toxic conversations on Twitter. In *Proceedings of the Web Conference 2021*, Ljubljana, Slovenia, 19–23 April 2021; pp. 1086–1097.
38. Fan, L.; Li, L.; Hemphill, L. Toxicity on Social Media During the 2022 Mpox Public Health Emergency: Quantitative Study of Topical and Network Dynamics. *J. Med. Internet Res.* **2024**, *26*, e52997. [\[CrossRef\]](#)
39. Shen, Z.; Paik, I. Temporal Modeling of Social Media for Depression Forecasting: Deep Learning Approaches with Pretrained Embeddings. *Appl. Sci.* **2025**, *15*, 11274. [\[CrossRef\]](#)
40. Andrews, S.; Ellis, D.A.; Shaw, H.; Piwek, L. Beyond self-report: Tools to compare estimated and real-world smartphone use. *PLoS ONE* **2015**, *10*, e0139004. [\[CrossRef\]](#)
41. Guo, S.; He, Z.; Rao, A.; Morstatter, F.; Brantingham, J.; Lerman, K. The pulse of mood online: Unveiling emotional reactions in a dynamic social media landscape. *ACM Trans. Web* **2025**, *19*, 1–22. [\[CrossRef\]](#)
42. Chin, H.; Lima, G.; Shin, M.; Zhunis, A.; Cha, C.; Choi, J.; Cha, M. User-chatbot conversations during the COVID-19 pandemic: Study based on topic modeling and sentiment analysis. *J. Med. Internet Res.* **2023**, *25*, e40922. [\[CrossRef\]](#)
43. Chin, H.; Yong Yi, M. Exploring the influence of user characteristics on verbal aggression towards social chatbots. *Behav. Inf. Technol.* **2025**, *44*, 1576–1594.
44. Noble, S.M.; Mende, M. The future of artificial intelligence and robotics in the retail and service sector: Sketching the field of consumer-robot-experiences. *J. Acad. Mark. Sci.* **2023**, *51*, 747–756. [\[CrossRef\]](#)
45. Li, T.G.; Zhang, C.B.; Chang, Y.; Zheng, W. The impact of AI identity disclosure on consumer unethical behavior: A social judgment perspective. *J. Retail. Consum. Serv.* **2024**, *76*, 103606. [\[CrossRef\]](#)
46. Black, E.W.; Mezzina, K.; Thompson, L.A. Anonymous social media—Understanding the content and context of Yik Yak. *Comput. Hum. Behav.* **2016**, *57*, 17–22. [\[CrossRef\]](#)
47. Bleize, D.N.; Anschütz, D.J.; Tanis, M.; Buijzen, M. The effects of group centrality and accountability on conformity to cyber aggressive norms: Two messaging app experiments. *Comput. Hum. Behav.* **2021**, *120*, 106754. [\[CrossRef\]](#)
48. Bogdan, A.; Dospinescu, N.; Dospinescu, O. Beyond Credibility: Understanding the Mediators Between Electronic Word-of-Mouth and Purchase Intention. *arXiv* **2025**, arXiv:2504.05359. [\[CrossRef\]](#)
49. Ganai, A.H.; Hashmy, R.; Khanday, H.A. Finding information diffusion’s seed nodes in online social networks using a special degree centrality. *SN Comput. Sci.* **2024**, *5*, 333. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.