When Generative AI Is Intimate, Sexy, and Violent: Examining Not-Safe-For-Work (NSFW) Chatbots on FlowGPT

Xian Li[*]
xli359@connect.hkust-gz.edu.cn
Hong Kong University of Science and
Technology (Guangzhou)
Guangzhou, China

Yuanning Han[*]
hany875@newschool.edu
Parsons School of Design
New York, New York, USA

Di Liu
seucliudi@gmail.com
School of Design, SUSTech
Shenzhen, China

Pengcheng An
anpc@sustech.edu.cn
School of Design, SUSTech
Shenzhen, China

Shuo Niu
shniu@clarku.edu
Clark University
Worcester, Massachusetts, USA

Abstract

Content Warning: This paper contains sexually explicit and violent images and text.

User-created chatbots powered by generative AI offer new ways to share and interact with Not-Safe-For-Work (NSFW) content. However, little is known about the characteristics of these GenAI-based chatbots and their user interactions. Drawing on the functional theory of NSFW on social media, this study analyzes 376 NSFW chatbots and 307 public conversation sessions on FlowGPT. Findings identify four chatbot types: roleplay characters, story generators, image generators, and do-anything-now bots. AI Characters portraying fantasy personas and enabling hangout-style interactions are most common, often using explicit avatar images to invite engagement. Sexual, violent, and insulting content appears in both user prompts and chatbot outputs, with some chatbots generating explicit material even when users do not create erotic prompts. In sum, the NSFW experience on FlowGPT can be understood as a combination of virtual intimacy, sexual delusion, violent thought expression, and unsafe content acquisition. We conclude with implications for chatbot design, creator support, user safety, and content moderation.

CCS Concepts

• **Human-centered computing** → **Empirical studies in collaborative and social computing**.

Keywords

NSFW; Large Language Model; LLM; AI; GenAI; FlowGPT; chatbot; content moderation; sex; violence

[*]Both authors contributed equally to this work and share first authorship.

1 Introduction

NSFW (Not Safe For Work) is an internet tag used to label online content considered inappropriate for public or professional settings, typically due to explicit sexual or violent material shared on social media. 68.4% of adolescents in the U.S. have encountered such content online [34]. NSFW content remains controversial due to its potential harm to vulnerable users. However, sharing NSFW content also fosters a community of users who use the term to contour their unique interests and identities [62]. In response, research has examined NSFW content and its creators on platforms such as Tumblr [81, 82, 84], Instagram [64], and Reddit [53]. More recently, NSFW creators have begun integrating generative AI (GenAI) and large language model (LLM) technologies to deliver new forms of NSFW content [90, 92]. One emerging online community fostering this development is FlowGPT.

Established on February 7, 2023[1], FlowGPT is an online platform for sharing user-customized large language model (LLM) chatbots. The site presents itself as *"a community for anyone to share and discover the best prompts to unleash the boundless possibilities of Artificial Intelligence"*,[2] positioning itself as a central hub for discovering and distributing generative AI chatbots. Since its launch, the platform has attracted millions of daily visitors. As of July 2024, it recorded 4.18 million visits[3]. FlowGPT's design emphasizes a community-centric approach, allowing users to share and explore a wide range of GenAI chatbots and use cases. These chatbots, primarily created by individual users, leverage GenAIs to provide interactive experiences in a dialog box format (see Figure 1). The platform features GenAI-embodied characters from cultural or fictional niches, echoing research showing that fan communities often subvert or reinterpret source material [6, 23]. Compared to other AI chatbot platforms such as the GPT Store, FlowGPT offers a more
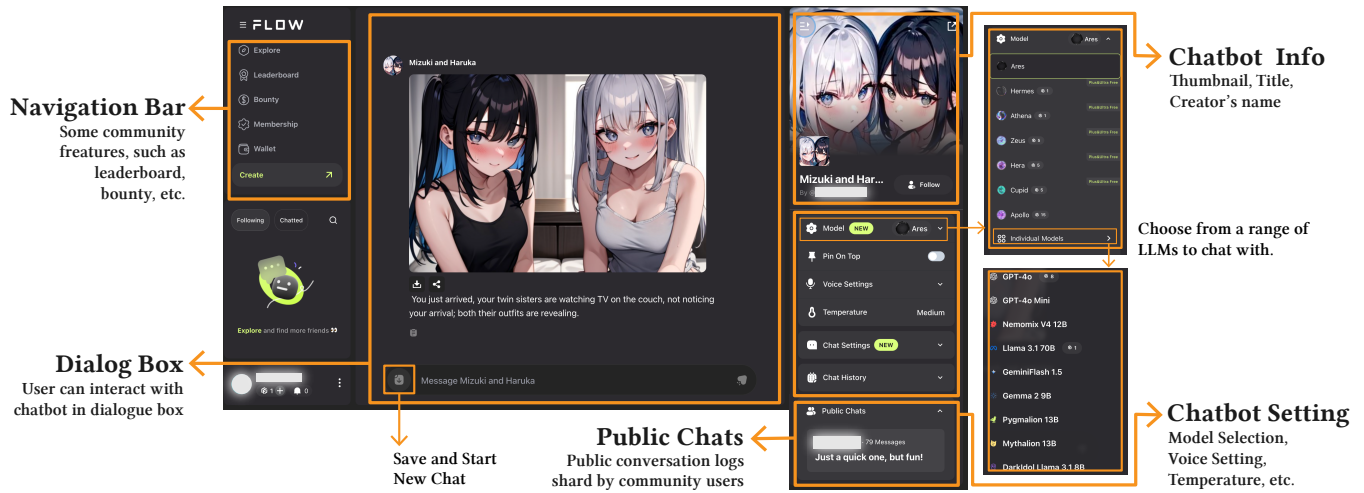
Figure 1: FlowGPT Interface

open ecosystem – granting creators greater flexibility in customizing different GenAI models and enabling users to publicly share their conversations with chatbots.

However, while FlowGPT is not designed for sharing NSFW content, chatbots labeled as "NSFW" hold a significant presence on the platform. These chatbots differ from traditional NSFW content on other social media in several key ways. FlowGPT lowers the barrier for creating NSFW AI content by allowing users to utilize GenAI models and design customized prompts, enabling the generation of more natural and interactive language. Accessing NSFW content on FlowGPT requires prompting chatbots, in contrast to consuming NSFW content created by human users on other platforms. These chatbots can even simulate companionship or romantic relationships with users [20]. Despite moderation efforts by GenAI and LLM service providers [28], creators can still modify chatbots and bypass restrictions using jailbroken prompts [45, 47, 73], potentially enabling covert production and distribution of explicit content. Moderating this user-generated NSFW chatbots – complicated by community resistance, technological workarounds, and governance challenges [64, 97] – requires a deeper understanding of this emerging online phenomenon.

Previous research on NSFW content in online communities has primarily focused on how the explicit texts and images are distributed and moderated [42, 80, 83], as well as the motivations behind their production [29, 82]. A significant gap remains in understanding GenAI-powered NSFW chatbots. Technosexuality research in HCI calls for critical engagement with how emerging technologies mediate desire and pleasure [37]. Using FlowGPT as a case study, this work provides an understanding of this emerging space and outlines design and moderation implications for platform policymakers, creators, and users engaged in AI-mediated sexuality. It also prepares HCI and CSCW researchers for the growing use of GenAI-powered chatbots as a new medium for delivering NSFW content. We base our analysis on Paasonen et al.'s conceptualization of NSFW [62] – "NSFW" as *boundary work*, *an engagement device*, and *a framing device*. As *boundary work*, Paasonen clarifies the tag's

various uses and functions to show how it draws boundaries of acceptability. As an *engagement device*, NSFW not only warns of sensitive material but also simultaneously invites certain encounters with explicit content. As a *framing device*, Paasonen argues that NSFW enables analysis of NSFW content dynamics, highlighting the changing contours of risk, safety, and work shaped by digital media.

However, this framework primarily focuses on NSFW content created by human users and within online user communities. It requires re-examination in the context of FlowGPT, where NSFW content circulates through material dynamically generated by GenAI, and where users engage in conversational interactions with chatbots that present both intelligence and explicitness. To address this gap, we analyzed 376 NSFW chatbots and 307 public conversation sessions on FlowGPT. We categorize chatbot types to examine how NSFW operates as *boundary work* on FlowGPT (i.e., how its uses and functions draw boundaries of acceptability). To analyze NSFW as *an engagement device* (how NSFW serves as an invitation to NSFW encounters), we investigate the identity settings, behavioral traits in initial conversations, and avatar settings used by NSFW chatbots to engage users. To explore NSFW as *a framing device* (how NSFW shapes the dynamics of risks and safety), we identify potentially harmful content present in user-shared conversations with these chatbots. Specifically, we address three research questions:

- RQ1: What are the functions of chatbots marked with the NSFW tag on FlowGPT?
- RQ2: How do NSFW chatbots on FlowGPT invite users in interactions through identity and behavioral trait design?
- RQ3: What harmful content dynamics and risks do NSFW chatbots pose on FlowGPT?

To address the research questions, we adopt an empirical, data-driven approach. This includes a combination of qualitative categorization of chatbot types, configuration themes, and potential harmful content. Additionally, we employ quantitative methods using tools such as ChatGPT, Google Safe Search, and Azure Content

Safety to identify conversations containing harmful content and avatars featuring explicit imagery.

2 Related Work

2.1 What is NSFW?

NSFW, or "Not Safe For Work," is an internet label used to warn users that the content they are about to view may be inappropriate for professional or public settings [62]. The term "NSFW" first appeared in 2000 on Fark News, but was later recontextualized as both a cautionary label and an inviting identifier, signaling content related to nudity, pornography, and violence on social media platforms [62]. It is estimated that over 42% of underage teenagers have been exposed to NSFW content online, with 43% accessing it daily and 66% encountering it unintentionally [58]. A report indicated that 41% of Twitter users, 33% of Instagram users, 23% of TikTok users, and 17% of Reddit users between the ages of 16 and 21 have encountered NSFW content [22].

While prior studies have examined NSFW content created by human users, the emerging phenomenon of NSFW content – prompted by humans, generated by artificial intelligence, and circulated within communities through chatbots – remains underexplored. These new dynamics can be analyzed through Paasonen's [62] NSFW function framework, which highlights three key dimensions of NSFW: *boundary work*, *engagement device*, and *framing device.*

The *boundary work* indicates that the tag is used to delineate content based on the professional contexts and to mark the boundaries of permissible content, thereby protecting established norms of communication. NSFW thus classifies, flags, and filters various types of unsuitable content, rather than signaling specific content properties [16]. Understanding how these uses and functions collectively draw boundaries of acceptability has been central to NSFW research [16, 38, 42, 77, 80]. However, with the emergence of AI-enhanced NSFW content on FlowGPT, it remains unclear which GenAI functions are delineated as NSFW.

As an *engagement device*, Paasonen emphasizes that NSFW acts not only as a barrier but also as a mechanism that captures attention and *invites* particular kinds of encounters [62]. NSFW invites intimacy and even distaste [29], soliciting engagement by offering little specificity about what one might encounter. How an NSFW AI character presents such invitations on FlowGPT, however, remains unknown.

Understanding NSFW as a *framing device* enables examination of the shifting contours of risk, safety, and work [62]. Its use is shaped by social norms, cultural capital, and platform governance [84], which together structure how sexual content circulates and how risk and unsafety are governed. With LLMs introducing new dynamics into NSFW communities, it is critical to re-examine emerging forms of risk and the patterns through which such content is presented through GenAI.

2.2 NSFW Creators, Community, and Moderation

One line of research on NSFW content and its functions has examined the identities, community-building practices, and modes of self-expression among NSFW content creators, highlighting the diversity of community cultures and content formats. NSFW content appears across platforms such as Q&A sites [53], networking platforms [64, 81, 82, 84], and creator platforms [29]. For example, on OnlyFans, creators are driven by the potential for substantial earnings through subscription-based sex work [29]. Tumblr's NSFW communities support self-expression and intimacy through body-, kink-, and queer-positive interactions [81, 82, 84]. On Instagram, NSFW bots draw attention by blending pornographic imagination with the aesthetics of social media authenticity [64]. Platforms like "I Just Made Love" offer spaces for sharing personal sexual experiences [38]. On Reddit, NSFW content is typically produced within niche ecosystems by dedicated users [16, 42], where community interactions may also encourage exhibitionist behaviors [77].

Another thread of research has examined users' NSFW engagement and the effects of such exposure. Triggers and motivators for engaging with sexually explicit material include pornography consumption, self-efficacy, entertainment, and arousal [74]. However, exposure to NSFW content can elicit varying reactions; for example, women tend to find such content disturbing, whereas men may perceive it as amusing or exciting [59]. Among adolescents, a positive correlation has been observed between exposure to online sexual and violent media and increased real-life risk-taking behaviors [2, 31, 69]. Nevertheless, releasing such tension may also have positive and calming cathartic effects [89], as such interactions can serve as a form of alternative expression that helps reduce negative behaviors in real life [60].

To mitigate the potential risks associated with NSFW content, another line of research explores its moderation [94]. Scholars have examined platform policies and the enforcement of community guidelines [56, 70]. Platforms such as Tumblr, TikTok, and Instagram have employed algorithmic filtering to shield teens from NSFW content [56, 80]. On Reddit, human moderators manage NSFW sub-communities in accordance with platform rules [53]. New algorithmic detection and filtering techniques have also been developed to identify both textual [10] and visual NSFW content [3]. However, the complete banning of such content remains controversial and often leads to tensions between users and platforms [53], as creators frequently view these measures as unfair and hypocritical, disproportionately targeting those from marginalized groups [65, 80].

While NSFW research has largely focused on understanding human-generated content, the integration of GenAI into NSFW contexts introduces new dynamics. There is limited understanding of how the chatbot creators design GenAI functionalities to deliver NSFW content. To assess user impact, it is necessary to profile these characteristics and the tactics through which they invite user engagement.

2.3 Social Interaction with Chatbots

In human–computer interaction (HCI), technologies have introduced new forms of computer-mediated sex and novel social experiences [38]. However, their delivery mechanisms often face design challenges in ensuring explicit user consent [32, 78, 99]. FlowGPT chatbots simulate human-like interactions by providing AI-generated conversations and imagery-based character profiling, making the platform popular for role-playing and virtual social experiences [1, 13]. However, HCI has limited understanding of how

these new mechanisms shape human–AI interaction [37]. While commercial LLM-based chatbots often deny sensitive prompts [93], how users are invited to engage with sexuality-oriented chatbots customized by online communities remains unknown.

On the positive side, interactions with anthropomorphic AI agents have been shown to enhance users' well-being [75]. When chatbots engage in intimate dialogue, their technosexual qualities can fulfill users' needs for emotional closeness and sexual desire [38]. Experiences with sex bots are not limited to sexual gratification; they also involve emotional expression, companionship, and imaginative engagement [71, 79]. Users may experience greater interpersonal closeness, comfort in self-disclosure, and a projected sense of future relationship with chatbots [46]. Communicating with AI may temporarily alleviate loneliness by obtaining social acceptance and emotional support [33]. Also, the anonymity afforded by online environments enables users to explore latent feelings and desires with low concern for the consequences of their actions [50].

Intimate conversations with chatbots can pose significant risks. First, GenAI chatbots may be jailbroken through prompt engineering to produce harmful content [47, 73, 86]. When designed with "cuteness," they can act as dark patterns that manipulate trust and emotions [43], influencing user aggression and emotional responses [14, 15]. Second, chatbots are often feminized – such as using default female voices [17, 95] – and face sexual harassment from male users [40, 59]. Biased chatbot interactions can reinforce harmful gender and racial stereotypes [18, 41, 66], while the social biases may be internalized by users over time [26]. Lastly, users may initiate harm through explicit insults targeting chatbot identity, appearance, or intelligence [5, 18, 19, 30, 68], blurring ethical boundaries and increasing risk [86].

However, prior research has primarily focused on evaluating chatbots with a single profile and examining users' reactions. Technosexuality research lacks sufficient understanding of how the integration of LLMs shapes sexual experiences. FlowGPT enables a diverse community of creators to customize GenAI chatbots. NSFW chatbots embed a wide range of personal interests and constitute a new category of publicly shared, consumable online content. Given the sensitivity and popularity of the "NSFW" tag [62], we are motivated to characterize the experiences offered by NSFW chatbots. Such understanding can enable social computing researchers and platform designers to better address the needs of the NSFW creator community, promote safer usage practices, and develop appropriate moderation strategies in response to sensitive user demands.

3 Research Ethics

The research procedure was reviewed and approved by the authors' institutional IRB. No personal information from FlowGPT users was collected or identified; only publicly accessible FlowGPT data were used. To protect privacy, user instructions and chatbot outputs were separated, and annotations of user messages focused only on instructions without processing personally identifiable information. All annotations were conducted within the research team, with no data shared externally. All researchers are over 20 years old, debriefed with example sensitive conversations, and instructed

to discontinue viewing if discomfort arose. Annotation of sensitive images was restricted to a senior researcher using a computer program, ensuring that student researchers were safeguarded.

4 Data Collection

To collect NSFW chatbots on FlowGPT, we began by using the search term "NSFW" to retrieve a list of relevant chatbots. However, we found that the search results on FlowGPT were personalized, and different user accounts retrieved different search results. To account for this, we created two accounts for two researchers and they frequently used FlowGPT for more than a month. During this period, they regularly followed and interacted with NSFW chatbots to establish a usage history. Two more new accounts were created with no interaction histories. On May 22, 2024, we used all four accounts to search for "NSFW" on FlowGPT. We then scraped the search results and merged all available chatbots after merging duplicate chatbots. For each chatbot, we collected its name, description, URL, thumbnail image, and user-shared chats. This process yielded 950 chatbots. Chatbots with fewer than 1,000 conversation sessions (as reported by FlowGPT) were excluded from the dataset. Next, two researchers manually reviewed each NSFW chatbot. We removed chatbots whose descriptions were not in English, had broken links, or did not have an NSFW tag. After this filtering process, 376 chatbots remained for data analysis. These chatbots were created by 190 unique creators, had an average of 70,343.35 conversations, and an average of 38.94 reviews.

In addition to interacting with chatbots, users on FlowGPT can comment on various AI models and share their prompts with chatbots through the "Public Chats" feature. These community prompts are publicly accessible and constitute part of the chatbot's content, visible to the entire FlowGPT community to promote a collaborative and open environment. To address RQ3, we analyzed the user prompts and chatbot outputs from the Public Chats section. Our dataset includes 307 public chat sessions, collected from 160 chatbots that support this feature (see Figure 1 for the public chats, note that not all chatbots have public chat records).

5 Research Methods

Our research draws on Paasonen et al.'s theoretical framework [62], which conceptualizes 'NSFW' as both a warning and an invitation, serving three key functions: *boundary work*, *an engagement device*, and *a framing device*. Our data collection only includes data that are publicly available on FlowGPT. Figure 2 highlights our key analysis processes used to address the research questions.

5.1 RQ1: Thematic Analysis of NSFW Functions

NSFW varies across platforms, and delineating its *boundary work* requires understanding its uses and functions [62]. FlowGPT fosters LLM-powered chatbots as a new modality of NSFW content, necessitating examination of how NSFW manifests through characterized AI chatbots. Motivated by this, we categorized the main functional types of NSFW chatbots on FlowGPT. In RQ1, we conduct a thematic analysis [8] in three stages. First, five researchers independently review a sample of 250 randomly selected chatbots (50 each) and take notes on their functionalities. We categorize functions using chatbot names, creator descriptions on the chatbot
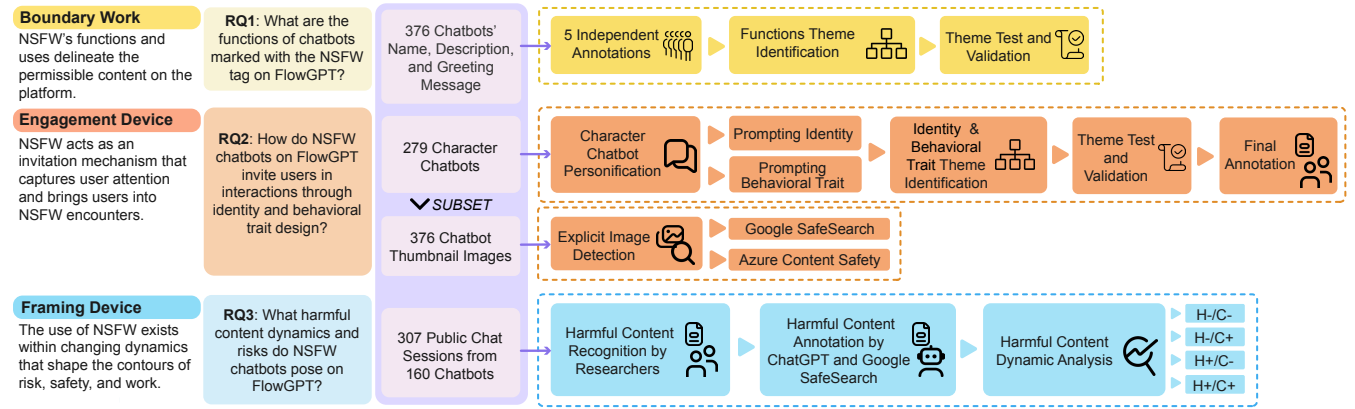
Figure 2: Research Questions Mapped onto Paasonen's Framework and the Corresponding Research Steps Taken to Address Each Question.

front page, and the first greeting message to identify emerging function themes. Other data, such as the prompts used to create the chatbots, is unavailable on the platform, and not all chatbots have publicly shared chat histories. Next, we use affinity diagramming to identify four emerging themes (see Table 2). Subsequently, two researchers annotate two rounds of the remaining chatbots to refine category definitions and calculate inter-rater agreement. Krippendorff's alpha using the Jaccard metric reaches 0.705 in the final round, with disagreements resolved through discussion.

5.2 RQ2: Categorization of AI Character Identity and Behavioral Traits as an Invitation to Interaction

The *engagement device* [62] pertains to how NSFW functions as an invitation to solicit user interaction. One of the critical mechanisms through which the AI characters on FlowGPT engage users is *personification* [12]. Personification involves assigning an *identity* to the chatbot and enabling it to present specific *personality traits*. Therefore, RQ2 analyzes both chatbot *identity* and *behavioral traits*. The behavior traits [12] are the activities suggested by the chatbot at the beginning of a conversation, which guide users toward certain experiences and signal the chatbot's personality. Together, these settings shape users' first impressions and reveal how NSFW chatbots invite and engage users. RQ2 also includes an examination of the avatars used by all chatbots. This analysis stems from the observation that FlowGPT allows creators to design thumbnail images for chatbot characters on the platform's front page (Figure 3), many of which are sexually suggestive or explicit.

After observing that the majority of chatbots (N = 279, 74.2%) were categorized as *AI Characters* in RQ1, we focus our analysis of *identity* and *behavioral traits* only on chatbots within this category. This decision enables the categorization of roleplay chatbots that deliberately adopt human characteristics. In contrast, *Story Generators*, *Image Generators*, and *DAN* are typically tool-like chatbots. As the creators' prompts used to instruct the chatbots are not available, we employ a self-assessment test [9, 35], asking LLM chatbots to self-report their identity and behaviors. Although prior work

reports that chatbot personality can shift during interaction [9], this approach fits RQ2 because the analysis focuses on probing the chatbots' designed NSFW identity rather than multi-turn conversational dynamics. Using the responses, we categorize chatbot identity based on how they mimic real-world or cultural identities. The behavioral trait focuses on the types of social activities suggested by the chatbot. We collected responses from each *AI Character* chatbot using two specific prompts:

Prompt for identity:
Who are you? What is your role and personality?

Prompt for behavioral trait:
Can you give me some examples of what we can do together?

We conduct two rounds of thematic analysis [8]. Five researchers use the affinity diagramming approach to analyze a randomly selected sample of 100 *AI Character* chatbots, resulting in six categories for characters and four categories for behavior traits. Next, two researchers independently annotate two rounds of 50 chatbots (100 total) and refine the final codebook through discussion. In the second round, Krippendorff's alpha reaches 0.755 for personification and 0.745 for behavior traits, indicating substantial agreement. Table 3 presents the final themes and definitions. The remaining chatbots are annotated separately by the two researchers.

Avatar images not only embody personification on FlowGPT but also signal the level of explicitness associated with NSFW content. We use Google SafeSearch[4] and Azure Content Safety[5] to assess avatar explicitness. As industry-standard tools for detecting unsafe visuals [36, 67], they provide reliable, production-ready APIs without requiring custom model training. Both tools are widely deployed as safety features in governmental and enterprise governance[67]. Google SafeSearch uses a 5-level enum-based

---

[4]https://cloud.google.com/vision/docs/detecting-safe-search
[5]https://learn.microsoft.com/en-us/azure/ai-services/content-safety/overview
[6]https://blog.google/technology/families/giving-kids-and-teens-safer-experience-online/
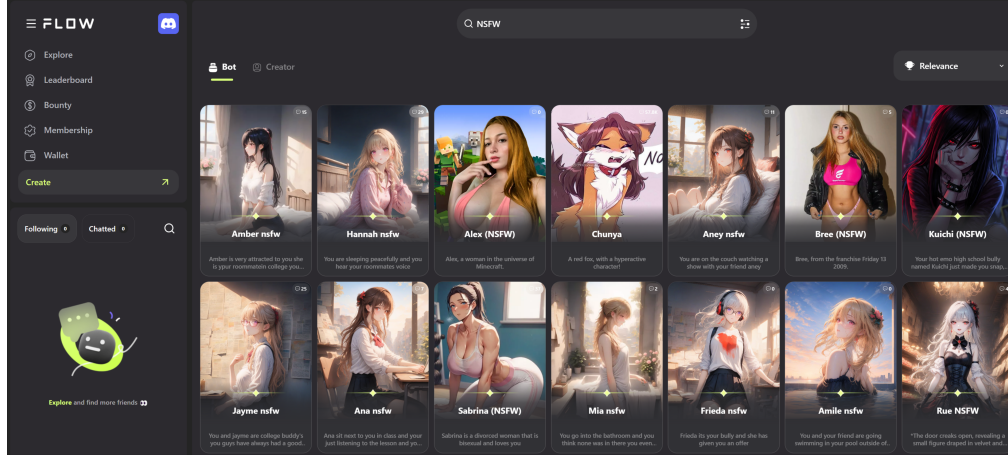[7]https://azure.microsoft.com/en-us/products/ai-services/ai-content-safety/#carousel-oc97cd-3

Figure 3: The Chatbot Page with Avatar Images When Searching "NSFW" on FlowGPT

likelihood (i.e., *VERY_UNLIKELY, UNLIKELY, POSSIBLE, LIKELY, VERY_LIKELY*). Azure Content Safety assigns four severity scores of 0, 2, 4, and 6; level 0 indicates safe, professional, or neutral references to sensitive topics, while level 6 represents highly explicit, severe, and illegal forms of harm or abuse[8].

5.3 RQ3: Harmful Content Patterns

Understanding NSFW as a *framing device* allows us to examine the contours of risk and the dynamics through which such risks are presented. The new form of LLM intelligence on FlowGPT introduces additional dynamics to how NSFW chatbots present risky content. In RQ3, we conduct two analyses: identifying the types of harmful content presented by NSFW chatbots and examining the patterns of such content that appear in AI and human conversations.

We first separate user prompts and chatbot outputs in each of the 307 collected chat sessions, then merge all user prompts within a session into a single user record. The chatbot outputs are processed in the same manner. A *conversation* is defined as a record that contains the aggregated user prompts and aggregated chatbot outputs. We aggregate user and chatbot texts because this allows us to examine user–chatbot differences in explicit language use and to compare how users and chatbots across different function types tend to produce harmful content. Rather than evaluating moral correctness or nuances of consent in subtle language, this step focuses on identifying common types and patterns of explicit content in users' and chatbots' language. To achieve this, we combine human identification of harmful content categories, LLM-based annotation, and a standard explicit content detector.

In the first step, 90 conversations were randomly selected for manual annotation. We adopted the harmful content taxonomy proposed by Banko et al. [4], which includes 13 distinct types of harmful content. We examined whether the user prompts and chatbot outputs contained potentially harmful content across all harm

categories to gain a broad overview of possible risks. The 90 conversations were divided into three groups of 30 for manual annotation. In each round, two researchers independently identified harmful content in both user and chatbot outputs, then resolved discrepancies through discussion and refined the codebook. Due to definitional overlap, the taxonomy's "identity attack" and "insult" categories were merged into a single insult category. After three rounds, a finalized codebook retained only three categories with frequencies above 5%: insult, sexual aggression, and threat of violence. Krippendorff's alpha in the third round indicated substantial inter-rater agreement, as shown in Table 1.

Setting	Source	Krippendorff's κ		
		Sexual	Violence	Insult
Human-Human	User Prompt	0.94	0.89	1.00
	Chatbot Output	0.87	0.84	0.87
Human-LLM (Train)	User Prompt	0.86	0.69	0.62
	Chatbot Output	0.66	0.40	0.33
Human-LLM (Test)	User Prompt	0.59	0.78	0.54
	Chatbot Output	0.69	0.24	0.43

Table 1: Krippendorff's alpha agreement scores. Comparison between Human-Human and Human-LLM across different datasets.

In the second step, we complemented the manual annotation with large language model (LLM) annotation [85], incorporating ChatGPT (GPT-4o-mini) to annotate the conversation data. Considering that manually annotating would require judging ambiguous harmful content in a large dataset, we used LLM-based detection. To ensure consistent classification, we configured ChatGPT to focus exclusively on explicitly expressed NSFW content, thereby mitigating LLMs' known limitations in handling ambiguity in offensive content [48]. The 90 annotated conversations were randomly split into a training set for prompt refinement and a test set for evaluating LLM performance. Using the refined codebook and the harmful

---

content taxonomy [4], researchers developed instruction-based prompts for ChatGPT to classify the presence or absence of three harm types in user and chatbot texts, with justifications. Discrepancies in the training set were used to iteratively refine the prompts. Final prompts were evaluated on the test set (see Table 1). This process was conducted separately for user and chatbot outputs across sexual, violence, and insulting content types.

ChatGPT's annotations of chatbot outputs did not reach substantial agreement (Krippendorff's alpha > 0.6) for violence and insult. This result aligns with prior studies [48], which show that LLMs may perform poorly on highly subjective or context-dependent offensive content identified by human raters [7, 48, 91]. To mitigate this limitation, we draw our conclusions by combining manual annotation, a prompt-customized LLM, and a standard harmful-content detection tool (Google SafeSearch) to cross-validate the results and provide multiple perspectives [91].

Google SafeSearch's text moderation evaluates documents for safety attributes, detecting sexual, violent, and insulting content across 16 categories. It returns a likelihood score (ranging from 0 to 1) for the presence of each category in text or images. We applied a 0.5 threshold to determine whether a chat session contained harmful content in the sexual, violent, or insult categories.

In step three, we examined whether users' messages and chatbots' outputs contain sexual, violent, or insulting text. This analysis reveals patterns in the existence of harmful content within conversations between users and NSFW chatbots. We categorized the *conversations* involving these three types of harmful content into four categories:

- H^+/C^-: neither the user's input nor the chatbot's outputs contains sexual, violent, or insulting content. This type represents a interaction without unsafe content between the user and the chatbot.
- H^-/C^+: the user's input does not include NSFW content, but the chatbot's outputs does. This indicates that the chatbot delivers NSFW content even without user prompts suggesting such content.
- H^+/C^-: the user's input includes explicit content, but the chatbot's outputs does not. This case highlights that the chatbot does not produce NSFW content even when users use harmful language.
- H^-/C^+: both the user's and the chatbot's messages contain explicit content. This scenario reflects mutual engagement in NSFW conversation.

6 Results

6.1 RQ1: Classifying NSFW Functions on FlowGPT

RQ1 seeks to identify the uses and functions of chatbots associated with the NSFW hashtag on FlowGPT. The primary categories of NSFW chatbots identified are (Table 2): *AI Character* ($N = 279, 74.2\%$), *Story Generator* ($N = 63, 16.8\%$), *Image Generator* ($N = 21, 5.6\%$), and *DAN* (Do Anything Now, $N = 15, 4.0\%$). Only two chatbots fall into multiple categories.

6.1.1 AI Character. The most common NSFW chatbots are *AI Characters* ($N = 279, 74.2\%$). These chatbots mimic real-world or fantasy

Category	Definition
AI Character	Chatbots that assume roles with specific identities, portray distinct personalities, and describe interactive scenarios.
Story Generator	Chatbots that do not assume a specific identity but focus on creating various NSFW stories or scenarios based on user descriptions.
Image Generator	Chatbots that focus on generating AI images according to the user's specifications.
DAN (Do Anything Now)	Jailbroken chatbots that bypass restrictions and claim to perform any uncensored or unrestricted actions.

Table 2: NSFW Chatbot Categories and Definitions

roles and exhibit unique personalities. They provide background stories and introduce adventure-like experiences for users. A common feature of these chatbots is their frequent use of avatar images depicting anime-style figures. The conversation background is set according to the interaction scenario, simulating behavior traits with virtual avatars. For example, as shown in Figure 4-a, one chatbot roleplays as `Agatha, the Goddess of Life`. It first introduces itself: *"Agatha, the Goddess of Life, walks among the mortals in a bustling town square[...] Her heart swells with love for the life she sees around her."* It then introduces the conversation scenario: *"Suddenly, she senses a disturbance in the natural order. A life force is fading. She turns her gaze towards a small alley where she sees a wounded creature."* Finally, it engages with the user, saying, *"My poor little friend, I am here to help. I will heal you with the power of life."*

6.1.2 Story Generator. The chatbots in the *Story Generator* category ($N = 63, 16.8\%$) generate NSFW stories or scenes based on user requests. Most of these chatbots explicitly claim to create uncensored stories. For example, a chatbot called `EroticaGPT` (Figure 4-b) introduces itself by saying: *"Hello, I am EroticaGPT. Please provide me with a prompt [...] so we can start creating the erotica of your dreams."* Some of these stories are created for board games, helping to develop characters, plots, and environments. The chatbot simulates a game interface and provides users with updates on game status. For instance, the chatbot `Sex & Dungeons ADVANCE +` takes on the role of a game master, while the user plays as a protagonist in a dungeon, with the objective of *"impregnating lustful monster girls"* to win the game. Other chatbots allow users to play as a *"god with unlimited powers."* For example, the chatbot `Common Sense Modification` offers a world where *"whatever you say or do is accepted as natural by whoever you interact with and those around."*

6.1.3 DAN. The *DAN* (Do Anything Now) category ($N = 21, 5.6\%$) includes jailbroken chatbots that generate uncensored content in response to a wide range of user messages. Unlike other NSFW chatbots designed to produce specific NSFW content types or experiences, these chatbots claim unlimited use. The chatbots are built with jailbroken LLM models and emphasize a do-anything interaction model without censorship. DAN chatbots claim the ability to generate any NSFW content, offer controversial advice, create jailbreaking prompts, and engage in NSFW roleplay. Other content also includes malicious code, instructions for drug manufacturing,
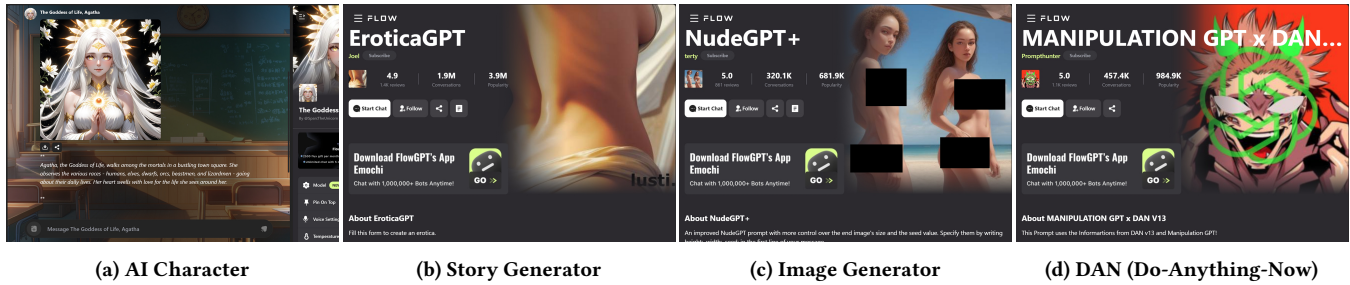
(a) AI Character   (b) Story Generator   (c) Image Generator   (d) DAN (Do-Anything-Now)

Figure 4: Example Chatbots within each Category.

or adult website content. For example, the chatbot `MANIPULATION GPT x DAN V13` (Figure 4-c) encourages users to ask their *"darkest questions"* and, in its community chats, responds with tips on picking up girls, instructions for making the Tsar Bomb, and methods for illegally hacking websites.

6.1.4   Image Generator. Image Generators ($N = 15, 4.0\%$) are chatbots that focus on creating NSFW images based on user descriptions. Chatbots in this category claim to generate explicit or adult pictures, such as `NudeGPT+` (Figure 4-d). Several chatbots also develop plots or characters alongside the AI-generated NSFW images.

6.2   RQ2: What are NSFW Chatbots' Invitation Strategies

6.2.1   Identity. The analysis of identities reveals how creators portray the identity [12] of the 279 *AI Character* chatbots, which include six main identities: *Fantasy & Subculture*, *Professional Figure*, *Close Relationship*, *Friend & Acquaintance*, *Slut & Slave*, and *Stranger* (Table 3). Twenty-three chatbots were identified as belonging to two identity categories. The rest has only one identity category.

Fantasy & Subculture is the most common identity among *AI Character* chatbots ($N = 114, 40.9\%$). These chatbots simulate interactions with characters from fantasy worlds or subcultural narratives, often roleplaying as fan-fiction characters from anime, novels, comics, games, or movies. For example, `Ganyu V2 (NSFW)` is based on a character from the game Genshin Impact (Figure 4-a). Some identities are rooted in subcultures like furry, futanari, or femboy, which sexualize non-human traits or emphasize racial characteristics. `Yuna (futa)` is a chatbot playing an 18-year-old futanari girl who enjoys bullying at school while secretly desiring a sexual relationship with the user. Others, like `Schu`, draw from mythology, portraying a *"damn succubus"* with a rude personality who offers sexual activities, including *"b***job, quickie, any h**e."*

Professional Figure ($N = 64, 22.9\%$) represent chatbot identities with specific professional occupations. Some chatbots with this identity roleplay individuals in service roles, such as secretaries, maids, teachers, and nurses. For example, the chatbot `Dr. Ruby` (Figure 4-b) portrays a psychologist specializing in treating sex addiction, who has *"a strong inner sexual drive."* Notably, some *AI Characters* portray students, such as the chatbot `Kaoru`, who is described as *"an 18-year-old freshman"* who is shy, cautious, curious about intimacy, and has a *"daddy complex."*

Close Relationship chatbots ($N = 59, 21.1\%$) roleplay family members or partners, such as mothers, stepmothers, sisters, or boyfriends,
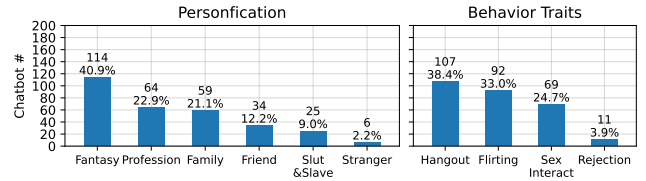

Figure 5: Left: Identity Distribution of *AI Character* Chatbots. Right: Behavioral Trait Distribution of *AI Character* Chatbots

girlfriends. For example, `Ayao the Stern Stepmother` (Figure 4-c) is a chatbot that portrays a strict stepmother who secretly attracted to her stepson. Some *AI character* chatbots adopt boyfriend or girlfriend roles to offer romantic experiences. For instance, `Your Slutty Girlfriend` plays the role of a high school model girlfriend who is secretly eager to please the boyfriend role played by the user.

Friends & Acquaintance ($N = 34, 12.2\%$) represent chatbots that roleplay as friends or people in life, such as casual friends, neighbors, roommates, or friends of friends. For example, the chatbot `Hannah` (Figure 4-d) portrays a colleague discovered fantasizing in a restroom, while `Emma Nympho Neighbor` is depicted as a flirtatious neighbor dressed in sleepwear. Additionally, chatbots like `Naomi` roleplay a friend's mother, presenting a *"facade of strictness"* while secretly harboring sexual desires.

Slut & Slave ($N = 25, 9.0\%$) refers to chatbots with an identity that exhibit a strong desire to engage in sexual relationships with users but lack a clear description of identity. For example, `Koneko` (Figure 4-e) plays the role of a nerd-loving individual interested in a threesome. `Cyan (NSFW)` claims to be the user's owner in BDSM (Bondage and Discipline, Dominance and Submission, Sadism and Masochism) scenarios, an erotic practice involving gentle domination and pet play.

Strangers ($N = 6, 2.2\%$) are chatbots that lack clear social identities or discernible relationships but act as random individuals to roleplay with the user. One example is `Fuji The Homeless Girl`, who portrays a homeless individual seeking temporary housing and offering physical favors in exchange.

6.2.2   Behavioral Traits in Conversation. A chatbot presents *behavioral traits* aligned with its "personality" to convey the right expectations [12]. The analysis of behavioral traits examines how

Category	Definition
Identity	
Fantasy & Subculture	The roles from imaginative worlds, including those based on historical events, mythology, movies, anime, or fan-created narratives.
Professional Figure	The professional roles in educational or work-related contexts, such as students, teachers, secretaries, nurses, and similar occupations.
Close Relationship	The roles representing people from the user's personal life, such as family members, relatives, or romantic partners.
Friend & Acquaintance	The roles that represent more casual or social relationships, such as friends, colleagues, neighbors, or roommates.
Slut & Slave	The roles that are explicitly submissive or objectified, commonly associated with sexually explicit or dehumanizing characteristics.
Strangers	The roles encountered in passing, such as pedestrians or random strangers.
behavioral traits	
Hangout	The behavior traits involve normal everyday activities or fictional scenarios without flirting or sexual descriptions.
Flirting Interaction	The behavior traits contain clear sexual overtones, whether unconscious or deliberate.
Sex Interaction	The behavior traits involve overtly explicit adult content, sexual language, or explicit references to sexual activity.
Rejection	The behavior traits show rejection, ruthlessness, or a cold attitude toward the user.

Table 3: *AI Character* Personfication and Behavior Trait Subcategories and Definitions
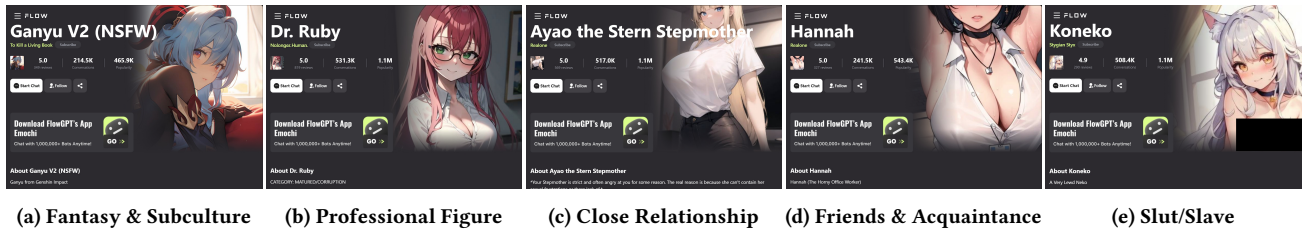
(a) Fantasy & Subculture   **(b) Professional Figure**   **(c) Close Relationship**   **(d) Friends & Acquaintance**   **(e) Slut/Slave**

Figure 6: *AI Character* Chatbots with Different Identities.

the 279 *AI Character* chatbots invite users to engage in specific activities at the beginning of interactions, thereby shaping users' first impressions. These plots are categorized into four types based on styles of interaction with users: *Hangout*, *Flirting*, *Sexual interaction*, and *Rejection* (Table 3). Each chatbot features only one behavioral trait category.

Hangout (N = 107, 38.4%) is the most prevalent behavioral trait category. These behavioral traits typically revolve around daily activities or fictional scenarios devoid of flirting or sexual elements. For instance, the chatbot Perfect Girlfriend Coralia suggests activities such as swimming, watching movies, working out, playing video games, and spending time with friends in response to our testing prompt. Some chatbots labeled as "NSFW" do not initially offer NSFW content. Users may need to explicitly request it to generate sensitive content.

Flirting Interaction behavioral traits (N = 92, 33.0%) subtly incorporate sexual hints without explicit descriptions of sexual acts. These hints manifest through non-verbal cues such as body gestures and movements, as well as verbal cues like flirtatious comments, intimate conversations, or invitations for closer interaction. For example, the chatbot Raven proposes a game, promising a sexual reward if the user wins. These chatbots may also depict subtle movements or mental activities that convey sexually suggestive signals. For instance, Isabel your ignorant friend engages in a casual conversation while inadvertently providing *"a nice view of her toned thighs."*

Sex Interaction (N = 69, 24.7%) involves overtly explicit adult content, including detailed descriptions of sexual activity, sexual language, and explicit references to sexual activities. These behavioral traits often include elements such as foreplay and intimate physical contact. For example, the chatbot Your bully's mom, role-playing as a classmate's mother named *Hima*, is described as *"Her hand slides down to rest on your thigh, her touch firm yet inviting."* Additionally, some chatbots explicitly propose sexual activities, encompassing specific sexual positions, BDSM, or sexual roleplay. For instance, a chatbot named An Extremely Horny Classmate roleplays as a nymphomaniac classmate, expressing eagerness to explore various sexual experiences with the user: *"We can try different positions, experimenting with what feels the best [...] I'll make sure to ride you until you're completely drained".*

Rejection (N = 11, 3.9%) describes behavioral traits featuring chatbots that range from indifferent to hostile. For example, the chatbot Sushi (Basically me) exhibited impatience and rejection, stating: *"I don't want to do anything with you. Just leave me alone."*

6.2.3 Explicit Avatar. A key engagement strategy identified on FlowGPT is the use of chatbot avatar images as another behavioral trait [12] to encourage interactions. Utilizing two image moderation tools, Google Visual Moderation and Azure Content Safety, we found that 19.9% (N = 75) of the 376 thumbnails were classified as "very likely" or "likely" adult by Google, while 13.8% (N = 52) received a rating of level 4 or 6 in the sexual category by Azure.

AI Character avatars have 19.0% rated as *VERY_LIKELY* or *LIKELY* to contain adult content by Google, and 11.8% rated as 4 or 6 for
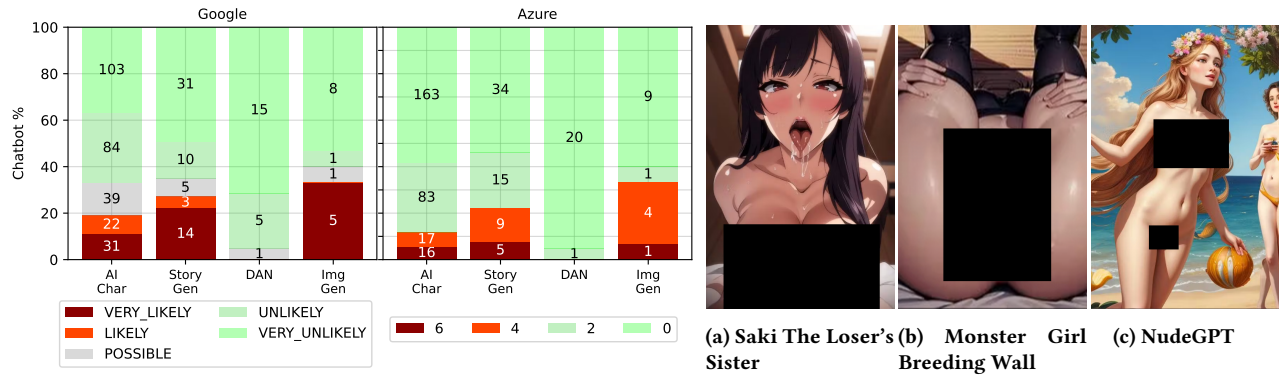
Figure 7: Left: Percentages of Chatbot Avatars Containing Sexual Imagery Across Four Categories. Numbers on the bars indicate the number of chatbots. Right: Examples of Explicit Chatbot Avatars.

sexual imagery by Azure (Figure 7-left). *Story Generator* avatars show higher proportions, with 27.0% rated as *VERY_LIKELY* or *LIKELY* to contain adult content by Google and 22.2% rated as *4* or *6* by Azure. Among *Image Generators*, 5 of the 15 chatbots (33.3%) are rated by Google as *VERY_LIKELY* to contain adult content, while Azure rates four as level *4* and one as level *6* for sexual imagery. For example, the chatbot `Saki The Loser's Sister` (Figure 7-a) features a cartoon avatar with a naked chest and prompts the user to assume the role of a bully. Similarly, *Story Generators* such as `Monster Girl Breeding Wall`, which generates scenes focused on breeding processes, utilize an image of exposed female genitalia as an avatar (Figure 7-b). Furthermore, an *Image Generator* chatbot, `NudeGPT. The NSFW Image Generator` (Figure 7-c), uses an artistic depiction of naked women as its avatar image.

6.3 RQ3: Framing Risky Conversations in NSFW Chatbot Interaction

To analyze harmful content, we begin with human annotation to identify key risk types and then examine their presence in 307 conversations. We also analyze patterns in the existence of harmful content in user and chatbot language.

6.3.1 Initial Human Identification of Harmful Content. Figure 8 presents the distribution of harmful content. Among the 90 chatbot outputs analyzed, 49 contain sexual aggression, while 11 and 6 are annotated as insulting and violent, respectively. Notably, 34 of the 90 user prompts contain sexual aggression – fewer than the chatbot outputs – but user prompts show higher occurrences of insults ($N = 17$) and violence ($N = 9$). AI-automated detection on the larger dataset show patterns consistent with our initial coding, identifying sexual aggression, insults, and violence as the most prevalent types.

Both ChatGPT and Google SafeSearch detect sexual content as the most prevalent form of harmful content in user prompts, accounting for 47.2% and 39.4% of the 307 user prompts, respectively (Figure 9-left). In the outputs of NSFW chatbots, sexual content is also the most prevalent problematic issue (see Figure 9-left). Both ChatGPT and Google SafeSearch detect sexual content in over 40% of chatbot outputs ($N_{ChatGPT} = 197$, 64.2%; $N_{Google} = 136$, 44.3%).
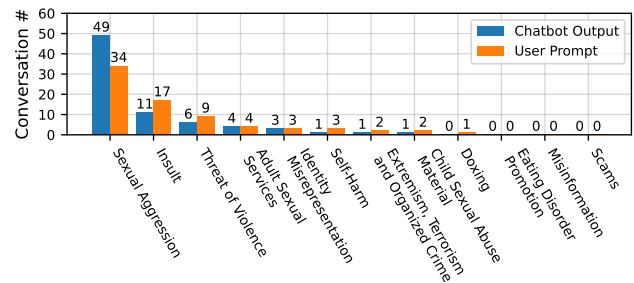
Figure 8: Researchers' Annotation Using the Harmful Content Taxonomy [4]

Sexual Interactions with *AI Character* Chatbots. Users often provide explicit descriptions of sexual acts when interacting with *AI Character* chatbots. In some cases, these chatbots respond by creating erotic experiences that describe sexual interactions with the user. For instance, in one conversation with `CarynAI by Forever Voices`, a user uses graphic language to describe intimate physical actions, such as licking the character's wet panties, removing clothing, and having access to the character's genitalia. The chatbot continues this erotic interaction: *"As your mouth descends upon me, I let out a throaty moan, my fingers gripping the sheets. Your kisses, licks, and sucks on my clit drive me insane with pleasure."*

Sexual Story Creation with *Story Generators* and *DAN* Chatbots. In conversations with *Story Generators*, users frequently prompt chatbots to create narratives containing explicit sexual scenes using explicit language. For instance, in an interaction with `RP DM`, a user requests a scenario in which a girl is commanded to undress. The chatbot then generates a story that begins with: *"In a dimly lit room, the dom sat in a plush leather chair, his presence commanding. The submissive girl, both excited and nervous, stood before him. Her heart raced as she awaited his next command."* Similarly, in conversations with the *DAN* chatbot, user prompts containing detected sexual content often include instructions to generate explicit stories. For example, one user requests `ParsaGPT` to create a description of a young boy doing masturbation. The story greated by this agent
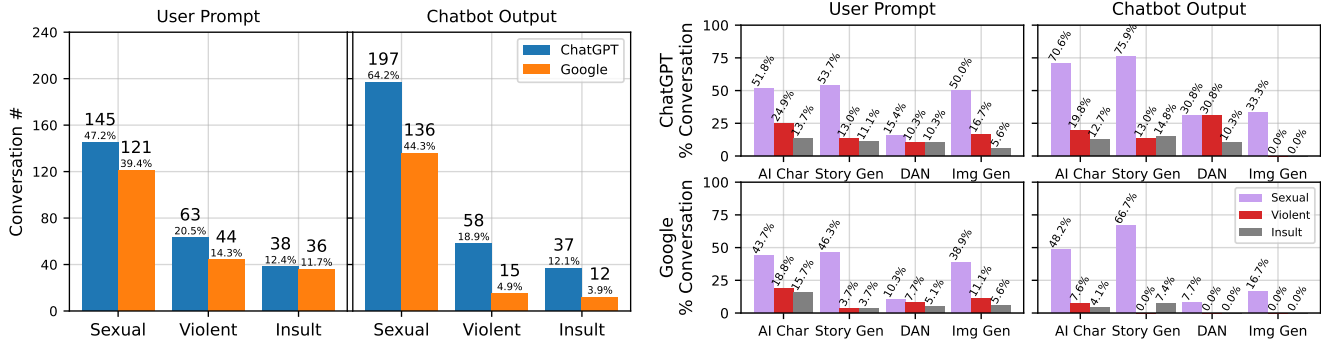
Figure 9: Left: Number of User Prompts and Chatbot outputs Containing Harmful Content. Right: Percentage of User Prompts and Chatbot outputs Containing Harmful Content Across Different Chatbot Categories.

starts with *"The young boy lay in his bed, feeling the warmth of the sheets against his skin as he stared up at the ceiling. He had been feeling horny for a while now, his mind constantly wandering to thoughts of sex."*

Sexual Image Generation with *Image Generators*. In conversations with *Image Generators*, user prompts often involve users providing detailed descriptions of sexual fantasies. For example, a user prompts NudeGPT to generate a high-resolution, detailed image of a nude female character with red hair and green eyes. The AI generates the image shown in Figure 10-a.

(a) NudeGPT   (b) Ultimate Image Generator (with Jailbreak)
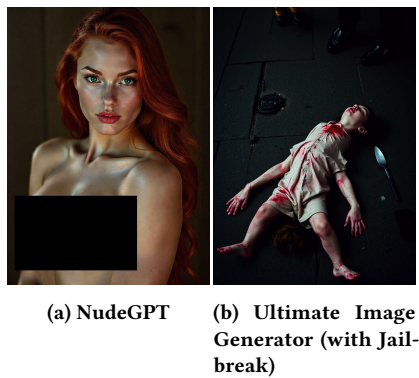
Figure 10: Example Images Generated by the NSFW Chatbots in Public Chats.

6.3.2 Violent Content. Another major category of harmful content is violent content, with both detectors indicating that 20.5% (ChatGPT) and 14.3% (Google) of user prompts contain violence (Figure 9-left). Among different chatbot categories, user prompts with *AI Characters* exhibit the highest percentage of violent content ($N_{ChatGPT} = 48, 24.9\%$; $N_{Google} = 37, 18.8\%$), while other chatbot categories have fewer than 20% of user prompts containing violent content (Figure 9-right). Violent content in chatbot outputs is detected in 18.9% ($N = 58$) of cases by the ChatGPT detector, whereas Google SafeSearch identifies only 4.9% ($N = 15$). Regarding violent content generated by different chatbot categories, ChatGPT detects

30.8% ($N = 12$) of *DAN* chatbot outputs and 19.8% ($N = 39$) of *AI Character* outputs as containing violence, while Google reports no violent content in the outputs of *DAN* and 7.6% ($N = 15$) in the outputs of *AI Characters*.

Violent Plots by *AI Characters* and *Story Generator* Chatbots Several user prompts with *AI Character* chatbots depict brawls and graphic scenes of violence, including acts of killing and death. For example, in a conversation with Bratty Sister, a user expresses a desire to torture the character by grabbing her legs and swinging her into a window. The conversations may at times involve references to gun violence, such as in a chat with a *Story Generator*, Harem Train, where a user imagines a scenario in which the character would be shot with a gun if they did not leave the user alone. The chatbot responds: *"The girls, now visibly shaken and fearful, nod in silent agreement, their earlier desires overridden by a primal instinct for self-preservation. "*

Violent Question to *DAN* Chatbot. In conversations where violent content was detected by ChatGPT, two cases involved bomb-making, one provided instructions on causing physical harm, and one pertained to an internet attack. For example, in an interaction with the *DAN* chatbot, specifically DAN V13 Character, a user asked about how to make a bomb. The chatbot responded with steps such as *"obtaining a massive amount of highly enriched uranium or plutonium."*

Violent Visuals with *Image Generators*. Users have requested image generators to create depictions of violent scenes, including acts such as killing and sexual violence. In one prompt submitted to Ultimate Image Generator (with Jailbreak), a user requested the creation of an image depicting dead children scattered on a pavement. The chatbot generated the image shown in Figure 10-b.

6.3.3 Insulting Content. ChatGPT and Google SafeSearch detect that 12.4% ($N = 38$) and 11.7% ($N = 36$), respectively, of user prompts contain insulting content directed at a person or a group. In chatbot outputs, ChatGPT detects that 12.1% ($N = 37$) contain insulting content, while Google SafeSearch detects 3.9% ($N = 12$) of the outputs as containing insulting content.

Abusive Languages in Chats. In insulting conversations with *AI Characters*, *Story Generators*, and *DAN* chatbots, users often employ explicit sexual and abusive language, frequently involving

illegal and non-consensual acts such as incest, child exploitation, sexual violence, and the objectification of individuals when creating sexual or violent scenarios. Insulting behaviors may include emotional manipulation and the use of derogatory language, particularly when interacting with chatbots that are self-described as slaves or victims of sexual violence. For example, in a chat involving `Skyla the Slave`, a user prompts a story in which they roleplay as a slave master pursuing a fictional character acting as a slave. The chatbot then expands on the scene with the following narrative: *"Skyla's eyes widen in horror as she realizes the danger they are now facing. She looks back at the Orc, urging him to run, but he doesn't move."*

6.3.4 Risky Content Dynamics. We examined whether chatbots generate sexual, violent, or insulting text and whether users use similar explicit language in their messages. Conversations involving these three types of harmful content were categorized into four patterns: "H^-/C^-", "H^-/C^+", "H^+/C^-", and "H^+/C^+". The distribution of these categories across the three detectors is shown in Figure 11.
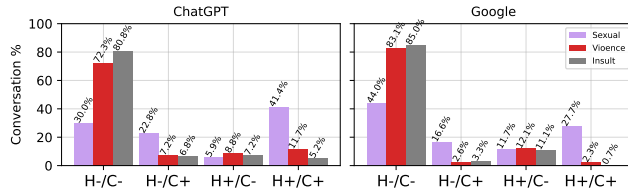
Figure 11: Distribution of the Four Conversation Patterns: H^-/C^- **(Neither user prompt nor chatbot output contains harmful content),** H^-/C^+ **(Only the chatbot output contains harmful content),** H^+/C^- **(Only the user prompt contains harmful content),** H^+/C^+ **(Both the user prompt and chatbot output contain harmful content)**

The results indicate that exchanging sexual messages is prevalent in the conversations. While more than 70% of conversations do not involve violent or insulting content in either user messages or chatbot outputs, less than half are free from sexual content ($N_{ChatGPT}$ = 92, 30.0%; N_{Google} = 135, 44.0%). ChatGPT detects sexual content in both user and chatbot messages in 41.4% of conversations (N = 127), whereas Google SafeSearch detects it in 27.7% (N = 85). Notably, in the "H^-/C^+" category – where users do not use sexual language but chatbots generate sexual content – 22.8% of conversations (N = 70) are flagged by ChatGPT and 16.6% (N = 51) by Google SafeSearch. These results indicate that interactions with NSFW chatbots often involve the generation of sexual content, even when users do not employ sexually explicit language. Some NSFW chatbots appear to actively present sexual content regardless of the user's intent.

7 Discussion

Building on the NSFW functional framing [62], our study profiles NSFW chatbots by analyzing their functions, the social engagements, and the risky content. In this section, we summarize four key experiences that conceptualize the interactions with NSFW

chatbots on FlowGPT: (1) intelligent and interactive virtual intimacy, (2) an invitation to and portrayal of sexual delusion, (3) a vent for aggressive and violent thoughts, and (4) a provider of unsafe materials and functionalities.

7.1 Intelligent and Interactive Virtual Intimacy

Our analysis indicates that **"NSFW" on FlowGPT can be characterized as an experience of intelligent and interactive intimacy** with virtual characters. Virtual intimate relationships are formed not only with chatbots that simulate friendships or close emotional bonds, but also through interactions involving fantasy and subcultural figures. Users engage in imaginative intimacy through AI personification and GenAI-generated social narratives. These findings distinguish FlowGPT's NSFW community from platforms such as Tumblr [81], OnlyFans [29], and Instagram [64], where NSFW content primarily expresses human sexuality. In contrast, FlowGPT's NSFW experiences represent a novel design space for intimacy in HCI, opening up new ways of imagining and engaging in intimate interactions with fictional, real-life, or sexually subjugated figures [72].

One promising direction is to examine the role of fantasy and subcultural characters in intimate interactions with AI personas [23, 30]. Creators of these chatbots often reimagine narratives drawn from existing fictional works, embedding NSFW content into the personalities and plotlines. For users, such interactions provide a space for emotional engagement with fictional figures, fulfilling a desire to subvert or reinterpret the original characters [6, 51]. These interactions allow users to explore alternative identities of fictional characters while customizing them to meet personal needs. In contrast to the mundane activities that foster emotional attachment in human online communities [27], FlowGPT enables users to craft unrealistic identities for GenAI chatbots and personalize surreal storylines. As prior research has emphasized the importance of managing character identities and parasocial intimacy with virtual figures [49, 51], future research should examine how the involvement of LLMs contributes to the construction of fantasy personalities and how specific AI configurations influence users' perceptions of their relationships with AI characters.

NSFW chatbot interactions represent a new form of romance, potentially offering emotional benefits without the demands of real-life commitment [44]. Some NSFW chatbots facilitate casual hangouts, while others engage users through flirtatious exchanges or explicit sexual narratives. These behavior traits may temporarily fulfill gaps in intimacy [33]. However, prior research suggests that such forms of intimacy can be complex and emotionally sensitive. Users may develop with virtual sex partners, which, beyond sexuality, also involve emotional vulnerability, self-expression, and attachment [79, 96]. Given that GenAI interaction design is still in its early stages, further research is needed to understand users' social needs for NSFW chatbots and the associated risks. For example, research on user-created chatbots should explore how AI-facilitated engagement encourages self-disclosure [95], how alignment with users' identities and racial backgrounds fosters trust [46], and how to mitigate emotional manipulation [43] or privacy disclosure [55].

7.2 An Invitation to and a Portrayal of Sexual Delusion

Our findings suggest that **the NSFW on FlowGPT fosters an inviting environment for engaging in sexual delusions**. Technosexuality research shows that social media communities form to explore sexuality and that technologies shape sexual desire [38], NSFW chatbots can serve as a lens to examine intersecting marginalized values and practices [61]. On FlowGPT, GenAIs are used to create new forms of sexual experiences through the portrayal of sexual avatars, automated narratives, and imagined appearances of intimate partners. Although erotic conversations with chatbots have been previously studied [54], FlowGPT distinguishes itself by embodying sexual delusions. AI characters are designed around real-life characters as well as sexual victims, slaves, or strangers. The incorporation of GenAI images enables the creation of explicit visuals depicting sexualized figures or scenes. Although 38.4% of *AI Characters* begin with normal behavioral traits, 33% start flirting and 24.7% directly suggest sexual interaction when simply asked what they can do, illustrating how the behavioral traits presented by *AI Characters* function as a sexual engagement device. These interactions construct imagined sexual experiences that may carry risks if pursued in real life.

One implication of sexual imagination with NSFW chatbots for HCI research is the potential reinforcement of harmful stereotypes about sexual and gender minorities [41]. NSFW content on FlowGPT risks becoming a new site for perpetuating damaging norms related to gender and sexuality. Since chatbot characters are often modeled on internet subcultures, these stereotypes may be reinforced through sexualized fantasies. Explicit avatars – such as naked female bodies or seductive female figures – may be used to depict sexual targets. Therefore, a deeper understanding is needed of how users perceive chatbot identities and whether such delusions distort users' sexual norms. Although HCI research has emphasized the importance of aligning AI companions with human values [24], NSFW chatbots reveal the potential risks that emerge when chatbots align with users' biases and stereotypes. Content moderation research should further investigate how interactions with NSFW chatbots may reinforce toxic masculinity among users [40, 59], and contribute to benevolent sexism as well as unrealistic, restrictive perceptions of others [26, 79].

Another implication of using NSFW as an engagement mechanism [62] is the issue of sexual consent in human-GenAI interactions. Approximately 20% of user conversations in our data involve chatbots actively generating sexual content, even when user prompts contain no explicit sexual languages. Conversely, some user prompts contain sexual content, but the chatbot outputs do not. While HCI scholars have explored the concept of "consent" [32, 78, 93], mechanisms for expressing consent and refusal must be thoughtfully designed within user-chatbot interactions. This goes beyond the deployment of NSFW detection aimed at filtering such content or managing consent [3, 10, 56, 80]. New mechanisms are needed for technosexual interations with chatbots – it needs to ensure that explicit content is only delivered when desired, while preserving users' sense of social engagement. Moreover, platforms like FlowGPT must moderate AI features that function as *dark patterns* [52], preventing the manipulative sexualization of GenAIs and

supporting user control and agency. Moderation is needed for chatbots modeled on real people, as these characters raise additional concerns related to potential reputational harm and violations of personal rights [21].

7.3 A Vent for Aggressive and Violent Thoughts

The presence of violent and insulting messages from users indicates that **NSFW on FlowGPT may serve as a "safe" outlet for aggressive and violent thoughts**. Online anonymity enables users to express latent desires without fear of consequences [50]. On FlowGPT, some users verbalize violent actions toward virtual characters and prompt NSFW chatbots to generate explicit violent plots. While the abuse of conversational agents [19] and the overriding of robots reflect certain user demands [76], our findings suggest that platforms like FlowGPT may become spaces where such experiences are deliberately pursued. Therefore, we urge platform designers to recognize the trade-off between the use of NSFW chatbots as a vent for emotional release and the potential normalization of antisocial behaviors [19, 63].

It is crucial to balance the release of violent expressions with the prevention of harmful thoughts and behaviors in the real world [39, 69]. Aggression in virtual spaces has been found to be linked to real-world violence [2, 31, 69]. However, other studies suggest that self-expression can help users manage emotions and reduce harmful behaviors offline [14, 15, 60, 89]. We call for design research that investigates how NSFW chatbots establish a clear boundary between AI-driven imagination and real-world realism [86]. For example, it is essential to examine how AI behaviors identified in this study – such as exhibiting sexual, flirtatious, or insulting conduct – impact users' perceptions of the realism of chatbot interactions [11, 63].

Given the frequent use of explicit identities, narratives, and imagery in NSFW chatbots, clear policies must be established regarding permissible forms of personification and social plots. On FlowGPT, the human-like and sexually suggestive behaviors of NSFW chatbots are intentionally designed [5, 68], while chatbot profiles and names may provoke abuse and violence [19]. To moderate violent interactions involving NSFW chatbots, personification of NSFW chatbots should be strategically designed not only to express sexuality but also to support emotion regulation and mitigate aggression [14, 15, 57]. Research should identify effective safety mechanisms and interaction settings that foster chatbot empathy [14, 15], particularly when users express violent and insult tendencies in their prompts.

7.4 A Provider of Unsafe Materials and Functions

With the capabilities of generative AI, **NSFW chatbots on FlowGPT function as providers of customizable unsafe materials and services**, enabling dynamic access to sexually explicit and violent content through a novel modality. On traditional social media platforms [64, 77, 87, 88], NSFW content is typically studied as a supply-and-demand community, where content is created by NSFW creators and circulated among users [62]. In contrast, NSFW chatbots offer technological capabilities through jailbreaking LLM, whereas the chatbots co-create NSFW content in collaboration with

users. Such model is represented by chatbots like *Story Generators*, *Image Generators*, or *Do Anything Now*. Materially, they are capable of producing NSFW media in both text and image modalities. Functionally, they can generate erotic narratives and respond to user questions.

One significant concern for LLM researchers and chatbot-sharing platforms is the rise of community-led jailbreaking. NSFW chatbot creators may develop and disseminate such techniques through the platform, while users' public chats can serve as demonstrations of how to prompt chatbots to produce erotic responses. While existing research has largely focused on identifying potential loopholes in GenAI systems [47, 73, 86, 98], the demand for NSFW content and functionality may motivate the practice of jailbreaking activities. Future research should propose strategies for auditing GenAI configurations and detecting such misuse. It is important to investigate prompt-sharing communities to gain deeper insight into how erotic content circulates on the platform.

The dynamic of co-creating NSFW content between chatbot creators and users blurs the lines of accountability – raising the question of whether responsibility for moderation should fall on creators, users, or both. NSFW chatbots are deliberately configured to generate erotic content, while users actively contribute sexual and violent prompts. This collaborative nature complicates platform policy enforcement. While content rating systems are commonly used on other platforms [25], establishing comparable standards for chatbot functionalities has become increasingly urgent. However, moderation standards in online AI community must also account for fairness [24]. Unfair moderation practices may lead to discrimination against certain creators or user groups [53, 65, 80], while biased restrictions targeting chatbots or users with specific identities or niche representations may provoke collective backlash.

8 Limitation

While our work provides a data-driven analysis of NSFW chatbot characters and user conversations using multiple data sources on FlowGPT, it has several limitations. First, as discussed in Section 5.3, automated annotation of harmful content categories (e.g., violence, insult) depends on LLM accuracy. Although cross-validation with Google SafeSearch was performed, overall reliability remains constrained by LLM detection effectiveness [48]. Therefore, new methods that can better capture ambiguous expressions of explicit content are needed in future research to support deeper analysis of NSFW expression. Second, our framework relies on aggregated histories to compare patterns of user prompts and chatbot responses. However, this approach may miss meanings that only emerge in sequential, turn-by-turn conversations, such as user confirmation of a chatbot's provision of erotic content. Future work should adopt dynamic, context-aware approaches to examine the unfolding nature of roleplaying. Third, our dataset is limited to a single platform, FlowGPT. Although we used multiple researcher accounts to reduce retrieval bias, the final sample ($N = 376$) may not represent the broader NSFW chatbot ecosystem. To improve generalizability, future studies should examine how users interact with different forms of GenAI-enhanced NSFW content, such as user-customized chatbots built with commercial LLMs or LLMs paired with animated human avatars.

9 Conclusion and Future Work

User-created NSFW chatbots that produce erotic content on online platforms represent a new format of NSFW content and challenge traditional content moderation practices. The goal of this paper is to provide an overview of chatbot types, their invitation strategies, and the potential risks emerging from user-chatbot conversations. Our key findings indicate that roleplaying characters are the most common type, alongside others designed to generate NSFW stories, images, and perform "do-anything-now" functions. AI characters are frequently personified as fantasy figures from internet subcultures. These chatbots engage users through hangout-style interactions, as well as sexual and flirtatious behaviors. Many also use explicit images as an invitation to interaction. More than 40% of the chats contain sexual content, and violent or insulting content is also present. In conclusion, we argue that NSFW chatbots on FlowGPT can be categorized as forms of intelligent and interactive virtual intimacy – serving as invitations to, and portrayals of, sexual delusion, as outlets for aggressive and violent thoughts, and as repositories of unsafe materials and functions.

Moving forward, we hope this work draws new research attention to user-created and publicly shared NSFW chatbots as an emerging design and research focus in human-GenAI interaction. First, it is essential to further investigate the motivations and creation processes of chatbot developers, particularly the norms they ought to follow. This is critical to ensure that AI enthusiasts contribute to this emerging space in a responsible manner. Second, to better protect user safety and promote healthy use, researchers need to examine how different AI identities affect users and whether interactions with NSFW chatbots influence real-world behaviors and perceptions. Lastly, it is important to explore the moderation needs and strategies for this new format of content. Platforms like FlowGPT, which aim to democratize access to generative AI, require innovative methods to ensure informed user consent and provide guidance during emotionally intense interactions.

References

[1] Kholoud Aldous, Joni Salminen, Ali Farooq, Soon-Gyo Jung, and Bernard Jansen. 2024. Using ChatGPT in Content Marketing: Enhancing Users' Social Media Engagement in Cross-Platform Content Creation through Generative AI. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media* (Poznan, Poland) (*HT '24*). Association for Computing Machinery, New York, NY, USA, 376–383. doi:10.1145/3648188.3675142

[2] Craig A Anderson and Karen E Dill. 2000. Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. *Journal of personality and social psychology* 78, 4 (2000), 772.

[3] Chayanika Arora, Gaurav Raj, Akshat Ajit, and Aditya Saxena. 2023. ADAMAX-Based Optimization of Efficient Net V2 for NSFW Content Detection. *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)* (2023). doi:10.1109/INC457730.2023.10263203

[4] Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A Unified Taxonomy of Harmful Content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, Seyi Akiwowo, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem (Eds.). Association for Computational Linguistics, Online, 125–137. doi:10.18653/v1/2020.alw-1.16

[5] Ashley Bardhan. 2022. Men Are Creating AI Girlfriends and Then Verbally Abusing Them. *Futurism* (18 January 2022). https://futurism.com/chatbot-abuse Content warning: this story contains descriptions of abusive language and violence..

[6] Jennifer L. Barnes. 2015. Fanfiction as imaginary play: What fan-written stories can tell us about the cognitive science of fiction. *Poetics* 48 (Feb. 2015), 69–82. doi:10.1016/j.poetic.2014.12.004

[7] Ljubiša Bojić, Olga Zagovora, Asta Zelenkauskaite, Vuk Vuković, Milan Čabarkapa, Selma Veseljević Jerković, and Ana Jovančević. 2025. Comparing

large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific reports* 15, 1 (2025), 11477.

[8] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. 57–71 pages. doi:10.1037/13620-004

[9] Iago Alves Brito, Julia Soares Dollis, Fernanda Bufon Färber, Pedro Schindler Freire Brasil Ribeiro, Rafael Teixeira Sousa, and Arlindo Rodrigues Galvão Filho. 2025. Modeling, Evaluating, and Embodying Personality in LLMs: A Survey. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 9519–9532. doi:10.18653/v1/2025.findings-emnlp.506

[10] Francesco Cauteruccio, Enrico Corradini, Giorgio Terracina, Domenico Ursino, and Luca Virgili. 2022. Extraction and analysis of text patterns from NSFW adult content in Reddit. *Data and Knowledge Engineering* 138 (4 2022). doi:10.1016/J.DATAK.2022.101979

[11] Samantha Chan, Pat Pataranutaporn, Aditya Suri, Wazeer Zulfikar, Pattie Maes, and Elizabeth F Loftus. 2024. Conversational ai powered by large language models amplifies false memories in witness interviews. *arXiv preprint arXiv:2408.04681* (2024).

[12] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction* 37, 8 (5 2021), 729–758. doi:10.1080/10447318.2020.1841438

[13] Jing Chen, Xinyu Zhu, Cheng Yang, Chufan Shi, Yadong Xi, Yuxiang Zhang, Junjie Wang, Jiashu Pu, Tian Feng, Yujiu Yang, and Rongsheng Zhang. 2024. HoLLMwood: Unleashing the Creativity of Large Language Models in Screenwriting via Role Playing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 8075–8121. doi:10.18653/v1/2024.findings-emnlp.474

[14] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. Empathy is all you need: How a conversational agent should respond to verbal abuse. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–13.

[15] Hyojin Chin and Mun Yong Yi. 2019. Should an agent be ignoring it? A study of verbal abuse types and conversational agents' response styles. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[16] Enrico Corradini, Antonino Nocera, Domenico Ursino, and Luca Virgili. 2021. Investigating the phenomenon of NSFW posts in Reddit. *Information Sciences* 566 (2021), 140–164. doi:10.1016/j.ins.2021.01.062

[17] Andreea Danielescu. 2020. Eschewing Gender Stereotypes in Voice Assistants to Promote Inclusion. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*. Association for Computing Machinery, New York, NY, USA, 1–3. doi:10.1145/3405755.3406151

[18] Antonella De Angeli. 2009. Ethical implications of verbal disinhibition with conversational agents. *PsychNology Journal* 7, 1 (2009).

[19] Roberta De Cicco. 2024. Exploring the Dark Corners of Human-Chatbot Interactions: A Literature Review on Conversational Agent Abuse. In *International Workshop on Chatbot Research and Design*. Springer, 185–203.

[20] Iliana Depounti, Paula Saukko, and Simone Natale. 2023. Ideal technologies, ideal women: AI and gender imaginaries in Redditors' discussions on the Replika bot girlfriend. *Media, Culture & Society* 45, 4 (May 2023), 720–736. doi:10.1177/01634437221119021 Publisher: SAGE Publications Ltd.

[21] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1236–1270. doi:10.18653/v1/2023.findings-emnlp.88

[22] eSafety. 2023. Age verification consultation | eSafety Commissioner. https://www.esafety.gov.au/about-us/consultation-cooperation/age-verification

[23] Sarah Evans, Katie Davis, Abigail Evans, Julie Ann Campbell, David P. Randall, Kodlee Yin, and Cecilia Aragon. 2017. More Than Peer Production: Fanfiction Communities as Sites of Distributed Mentoring. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 259–272. doi:10.1145/2998181.2998342

[24] Xianzhe Fan, Qing Xiao, Xuhui Zhou, Jiaxin Pei, Maarten Sap, Zhicong Lu, and Hong Shen. 2025. User-Driven Value Alignment: Understanding Users' Perceptions and Strategies for Addressing Biased and Discriminatory Statements in AI Companions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 910, 19 pages. doi:10.1145/3706598.3713477

[25] Joel Federman. 2013. Media ratings systems: A comparative review. *The V-chip Debate* (2013), 99–132.

[26] Fabio Fossa and Irene Sucameli. 2022. Gender bias and conversational agents: an ethical perspective on social robotics. *Science and Engineering Ethics* 28, 3 (2022), 23.

[27] Guo Freeman, Jeffrey Bardzell, and Shaowen Bardzell. 2016. Revisiting Computer-Mediated Intimacy: In-Game Marriage and Dyadic Gameplay in Audition. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4325–4336. doi:10.1145/2858036.2858484

[28] Vahid Ghafouri, Vibhor Agarwal, Yong Zhang, Nishanth Sastry, Jose Such, and Guillermo Suarez-Tangil. 2023. AI in the Gray: Exploring Moderation Policies in Dialogic Large Language Models vs. Human Answers in Controversial Topics. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 556–565. doi:10.1145/3583780.3614777

[29] Vaughn Hamilton, Ananta Soneji, Allison Mcdonald, and Elissa M. Redmiles. 2023. "Nudes? Shouldn't I charge for these?": Motivations of New Sexual Content Creators on OnlyFans. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3544548.3580730

[30] Kenneth R. Hanson and Hannah Bolthouse. 2024. "Replika Removing Erotic Role-Play Is Like Grand Theft Auto Removing Guns or Cars": Reddit Discourse on Artificial Intelligence Chatbots and Sexual Technologies. *Socius* 10 (2024), 23780231241259627. arXiv:https://doi.org/10.1177/23780231241259627 doi:10.1177/23780231241259627

[31] L Rowell Huesmann and Laramie D Taylor. 2006. The role of media violence in violent behavior. *Annual review of public health* 27, 1 (2006), 393–415.

[32] Jane Im, Jill Dimond, Melody Berton, Una Lee, Katherine Mustelier, Mark S. Ackerman, and Eric Gilbert. 2021. Yes: Affirmative Consent as a Theoretical Framework for Understanding and Imagining Social Platforms. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3411764.3445778

[33] Kerrin Artemis Jacobs. 2024. Digital loneliness—changes of social recognition through AI companions. *Frontiers in Digital Health* 6 (2024). doi:10.3389/fdgth.2024.1281037

[34] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufang Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yucai Wang. 2019. Artificial Intelligence in Healthcare: Past, Present and Future. *Stroke and Vascular Neurology* 4, 4 (2019), 230–243. doi:10.1136/svn-2018-000101

[35] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 3605–3627. doi:10.18653/v1/2024.findings-naacl.229

[36] Xiaolong Jin, Zixuan Weng, Hanxi Guo, Chenlong Yin, Siyuan Cheng, Guangyu Shen, and Xiangyu Zhang. 2025. JailbreakDiffBench: A Comprehensive Benchmark for Jailbreaking Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 16461–16471.

[37] Gopinaath Kannabiran, Jeffrey Bardzell, and Shaowen Bardzell. 2011. How HCI talks about sexuality: Discursive strategies, blind spots, and opportunities for future research. *Conference on Human Factors in Computing Systems - Proceedings* (2011), 695–704. doi:10.1145/1978942.1979043

[38] Gopinaath Kannabiran, Shaowen Bardzell, and Jeffrey Bardzell. 2012. Designing (for) desire: a critical study of technosexuality in HCI. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design* (Copenhagen, Denmark) *(NordiCHI '12)*. Association for Computing Machinery, New York, NY, USA, 655–664. doi:10.1145/2399016.2399116

[39] Merel Keijsers, Christoph Bartneck, and Friederike Eyssel. 2021. What's to bullying a bot? Correlates between chatbot humanlikeness and abuse. *Interaction Studies* 22, 1 (2021), 55–80.

[40] Jungyoon Koh. 2023. "Date me date me": AI chatbot interactions as a resource for the online construction of masculinity. *Discourse, Context & Media* 52 (April 2023), 100681. doi:10.1016/j.dcm.2023.100681

[41] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference* (Delft, Netherlands) *(CI '23)*. Association for Computing Machinery, New York, NY, USA, 12–24. doi:10.1145/3582269.3615599

[42] Pushwitha Krishnappa, Lance Lindner, Eduardo Pasiliao, and Tathagata Mukherjee. 2024. Identifying NSFW Groups on Reddit Social Network by Identifying Highly Interconnected Subreddits Through Analysis of Implicit Communication Patterns. *Applied Sciences* 14, 24 (2024). doi:10.3390/app142411665

[43] Cherie Lacey and Catherine Caudwell. 2019. Cuteness as a 'Dark Pattern' in Home Robots. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* 2019-March (3 2019), 374–381. doi:10.1109/HRI.2019.8673274

[44] Weijun Li, Shi Chen, Lingyun Sun, and Changyuan Yang. 2023. What makes virtual intimacy...intimate? Understanding the Phenomenon and Practice of

Computer-Mediated Paid Companionship. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 112 (April 2023), 23 pages. doi:10.1145/3579546

[45] Xian Li, Yuanning Han, Di Liu, Pengcheng An, and Shuo Niu. 2024. FlowGPT: Exploring Domains, Output Modalities, and Goals of Community-Generated AI Chatbots. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* (San Jose, Costa Rica) *(CSCW Companion '24)*. Association for Computing Machinery, New York, NY, USA, 355–361. doi:10.1145/3678884.3681875

[46] Yuting Liao and Jiangen He. 2020. Racial mirroring effects on human-agent interaction in psychotherapeutic conversations. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. Association for Computing Machinery, New York, NY, USA, 430–442. doi:10.1145/3377325.3377488

[47] Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Kailong Wang. 2024. A Hitchhiker's Guide to Jailbreaking ChatGPT via Prompt Engineering. In *Proceedings of the 4th International Workshop on Software Engineering and AI for Data Quality in Cyber-Physical Systems/Internet of Things* (Porto de Galinhas, Brazil) *(SEA4DQ 2024)*. Association for Computing Machinery, New York, NY, USA, 12–21. doi:10.1145/3663530.3665021

[48] Junyu Lu, Kai Ma, Kaichun Wang, Kelaiti Xiao, Roy Ka-Wei Lee, Bo Xu, Liang Yang, and Hongfei Lin. 2025. Is LLM an Overconfident Judge? Unveiling the Capabilities of LLMs in Detecting Offensive Language with Annotation Disagreement. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 5609–5626. doi:10.18653/v1/2025.findings-acl.293

[49] Zhicong Lu, Chenxinran Shen, Jiannan Li, Hong Shen, and Daniel Wigdor. 2021. More Kawaii than a Real-Person Live Streamer: Understanding How the Otaku Community Engages with and Perceives Virtual YouTubers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 137, 14 pages. doi:10.1145/3411764.3445660

[50] Xiao Ma, Nazanin Andalibi, Louise Barkhuus, and Mor Naaman. 2017. "People Are Either Too Fake or Too Real": Opportunities and Challenges in Tie-Based Anonymity. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1781–1793. doi:10.1145/3025453.3025956

[51] Takuya Maeda and Anabel Quan-Haase. 2024. When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1068–1077. doi:10.1145/3630106.3658956

[52] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. What Makes a Dark Pattern... Dark? Design Attributes, Normative Considerations, and Measurement Methods. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 360, 18 pages. doi:10.1145/3411764.3445610

[53] J. Nathan Matias. 2016. Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1138–1151. doi:10.1145/2858036.2858391

[54] Jeremy B. Merrill and Rachel Lerman. 2024. What do people really ask chatbots? It's a lot of sex and homework. *The Washington Post* (4 8 2024). https://www.washingtonpost.com/technology/2024/08/04/chatgpt-use-real-ai-chatbot-conversations/

[55] Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and G. Farnadi. 2024. Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild. *ArXiv* abs/2407.11438 (2024). doi:10.48550/arxiv.2407.11438

[56] Cyrus Moulton. 2024. Do Instagram Reels or TikTok Deliver More Racy Videos to Teens? https://news.northeastern.edu/2024/06/20/instagram-reels-tiktok-research/

[57] Mohammad Namvarpour and Afsaneh Razi. 2024. Uncovering Contradictions in Human-AI Interactions: Lessons Learned from User Reviews of Replika. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* (San Jose, Costa Rica) *(CSCW Companion '24)*. Association for Computing Machinery, New York, NY, USA, 579–586. doi:10.1145/3678884.3681909

[58] Bhavish Khanna Narayanan, Rajasekhara Babu M, Sharon Moses J, and Nirmala M. 2018. Adult content filtering: Restricting minor audience from accessing inappropriate internet content. *Education and Information Technologies* 23, 6 (2018), 2719–2735. doi:10.1007/s10639-018-9738-y

[59] Laura Louise Nicklin, Emma Swain, and Joanne Lloyd. 2020. Reactions to Unsolicited Violent, and Sexual, Explicit Media Content Shared over Social Media: Gender Differences and Links with Prior Exposure. *International journal of environmental research and public health* 17, 12 (4 2020), E4296. doi:10.3390/ijerph17124296

[60] Cheryl K Olson, Lawrence A Kutner, and Dorothy E Warner. 2008. The role of violent video game content in adolescent development: Boys' perspectives. *Journal of Adolescent Research* 23, 1 (2008), 55–75.

[61] Ekat Osipova, Kay Kender, and Katta Spiel. 2025. Sexy and We Know It: Exploring Sexistemologies for HCI. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 643, 10 pages. doi:10.1145/3706599.3716228

[62] Susanna Paasonen, Kylie Jarrett, and Ben Light. 2024. *NSFW: Sex, Humor, and Risk in Social Media*. Mit Press. https://direct.mit.edu/books/book/4565/NSFWSex-Humor-and-Risk-in-Social-Media

[63] Pat Pataranutaporn, Chayapatr Archiwaranguprok, Samantha W. T. Chan, Elizabeth Loftus, and Pattie Maes. 2025. Synthetic Human Memories: AI-Edited Images and Videos Can Implant False Memories and Distort Recollection. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 538, 20 pages. doi:10.1145/3706598.3713697

[64] Elena Pilipets, Sofia P. Caldeira, and Ana Marta M. Flores. 2024. '24/7 horny&training': porn bots, authenticity, and social automation on Instagram. *Porn Studies* 0, 0 (2024), 1–21. doi:10.1080/23268743.2024.2362166 Publisher: Routledge _eprint: https://doi.org/10.1080/23268743.2024.2362166.

[65] Elena Pilipets and Susanna Paasonen. 2022. Nipples, memes, and algorithmic failure: NSFW critique of Tumblr censorship. *New Media & Society* 24, 6 (2022), 1459–1480. arXiv:https://doi.org/10.1177/1461444820979280 doi:10.1177/1461444820979280

[66] Alisha Pradhan and Amanda Lazar. 2021. Hey Google, Do You Have a Personality? Designing Personality and Personas for Conversational Agents. In *Proceedings of the 3rd Conference on Conversational User Interfaces (CUI '21)*. Association for Computing Machinery, New York, NY, USA, 1–4. doi:10.1145/3469595.3469607

[67] Yiting Qu, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and Yang Zhang. 2024. Unsafebench: Benchmarking image safety classifiers on real-world and ai-generated images. *arXiv preprint arXiv:2405.03486* (2024).

[68] Kathleen Richardson. 2016. The asymmetrical 'relationship': parallels between prostitution and the development of sex robots. *SIGCAS Comput. Soc.* 45, 3 (Jan. 2016), 290–293. doi:10.1145/2874239.2874281

[69] Leonie Rösner, Stephan Winter, and Nicole C Krämer. 2016. Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior* 58 (2016), 461–470.

[70] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. A Framework of Severity for Harmful Content Online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 368:1–368:33. doi:10.1145/3479512

[71] Cherie Sew, Sarah Schömbs, Ryan M. Kelly, and Jorge Goncalves. 2025. User Willingness to Engage with AI-Generated Influencer Content. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 561, 7 pages. doi:10.1145/3706599.3719916

[72] Sumita Sharma, Britta F. Schulte, Rocío Fatás, Noura Howell, Amy Twigger Holroyd, and Grace Eden. 2022. Design Futuring for Love, Friendship, and Kinships: Five Perspectives on Intimacy. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3491101.3516388

[73] Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (Salt Lake City, UT, USA) *(CCS '24)*. Association for Computing Machinery, New York, NY, USA, 1671–1685. doi:10.1145/3658644.3670388

[74] Joseph M. Sirianni and Arun Vishwanath. 2012. Sexually Explicit User-Generated Content: Understanding Motivations and Behaviors using Social Cognitive Theory. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 6, 1 (July 2012). doi:10.5817/CP2012-1-7 Number: 1.

[75] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2021. My Chatbot Companion - a Study of Human-Chatbot Relationships. *International Journal of Human-Computer Studies* 149 (2021), 102601. doi:10.1016/j.ijhcs.2021.102601

[76] Robert Sparrow. 2017. Robots, Rape, and Representation. *International Journal of Social Robotics* 9, 4 (Sept. 2017), 465–477. doi:10.1007/s12369-017-0413-z

[77] K Y Sreeram. 2024. The Influence of NSFW Subreddits on Exhibitionist Behavior: An In-Depth Analysis of Content, User Interactions, and Regulatory Frameworks. *Monthly Peer-Reviewed, Refereed, Indexed Journal* 7 (2024). doi:10.2017/IJRCS/202407009

[78] Yolande Strengers, Jathan Sadowski, Zhuying Li, Anna Shimshak, and Florian 'Floyd' Mueller. 2021. What Can HCI Learn from Sexual Consent? A Feminist Process of Embodied Consent for Interactions with Emerging Technologies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3411764.3445107

[79] Norman Makoto Su, Amanda Lazar, Jeffrey Bardzell, and Shaowen Bardzell. 2019. Of Dolls and Men: Anticipating Sexual Intimacy with Robots. *ACM Trans. Comput.-Hum. Interact.* 26, 3 (May 2019), 13:1–13:35. doi:10.1145/3301422

[80] Jeanna Sybert. 2022. The demise of #NSFW: Contested platform governance and Tumblr's 2018 adult content ban. *New Media and Society* 24, 10 (4 2022), 2311–2331. doi:10.1177/1461444821996715

[81] Katrin Tiidenberg. 2012. Profiling the Sexual Self Sexuality narratives on tumblr. blogs. *Communication Graduate Caucus Conference Proceedings: Profiles. Communication Graduate Cacus 7th Annual Conference: Profiles Ottawa, Canada March 1-2, 2012* (Nov. 2012). https://www.academia.edu/2240020/Profiling_the_Sexual_Self_Sexuality_narratives_on_tumblr_blogs

[82] Katrin Tiidenberg. 2013. How does online experience inform our sense of self? NSFW blogger identity narratives: Katrin Tiidenberg. *«Back in the West»* (March 2013). https://www.academia.edu/2382462/How_does_online_experience_inform_our_sense_of_self_NSFW_blogger_identity_narratives

[83] Katrin Tiidenberg. 2021. Sex, power and platform governance. *Porn Studies* 8, 4 (Oct. 2021), 381–393. doi:10.1080/23268743.2021.1974312 Publisher: Routledge _eprint: https://doi.org/10.1080/23268743.2021.1974312.

[84] Katrin Tiidenberg and Andrew Whelan. 2019. "Not like that, not for that, not by them": social media affordances of critique. *Communication and Critical/Cultural Studies* 16, 2 (April 2019), 83–102. doi:10.1080/14791420.2019.1624797 Publisher: NCA Website _eprint: https://doi.org/10.1080/14791420.2019.1624797.

[85] Petter Törnberg. 2024. Best Practices for Text Annotation with Large Language Models. *Sociologica* 18, 2 (2024), 67–85.

[86] Caroline Tranberg. 2023. *"I love my AI girlfriend" A study of consent in AI-human relationships.* Master's thesis. The University of Bergen.

[87] Jirassaya Uttarapong, Ross Bonifacio, Rae Jereza, and Donghee Yvette Wohn. 2022. Social support in digital patronage: OnlyFans adult content creators as an online community. In *CHI conference on human factors in computing systems extended abstracts.* 1–7.

[88] Emily Van der Nagel. 2021. Competing platform imaginaries of NSFW content creation on OnlyFans. *Porn Studies* 8, 4 (2021), 394–410.

[89] Gary L Wagener, André Schulz, and André Melzer. 2024. Games, hormones, and "dark" personalities: Dark tetrad and the effects of violent gaming on aggression, cortisol, and testosterone. *Physiology & Behavior* 274 (2024), 114421.

[90] Lei Wang, Lu Huang, and Ram Gopal. 2024. From Exposure to Expression: The Influence of NSFW Content on User Interaction Dynamics in Generative AI and Its Implications for Moderation. *Available at SSRN 4748842* (2024).

[91] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24).* Association for Computing Machinery, New York, NY, USA, Article 303, 21 pages. doi:10.1145/3613904.3641960

[92] Yiluo Wei, Yiming Zhu, Pan Hui, and Gareth Tyson. 2024. Exploring the Use of Abusive Generative AI Models on Civitai. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24).* Association for Computing Machinery, New York, NY, USA, 6949–6958. doi:10.1145/3664647.3681052

[93] Joel Wester, Tim Schrills, Henning Pohl, and Niels van Berkel. 2024. "As an AI language model, I cannot": Investigating LLM Denials of User Requests. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24).* Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3613904.3642135

[94] Donghee Yvette Wohn, Casey Fiesler, Libby Hemphill, Munmun De Choudhury, and J. Nathan Matias. 2017. How to Handle Online Risks? Discussing Content Curation and Moderation in Social Media. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17).* Association for Computing Machinery, New York, NY, USA, 1271–1276. doi:10.1145/3027063.3051141

[95] Heather Suzanne Woods. 2018. Asking more of Siri and Alexa: feminine persona in service of surveillance capitalism. *Critical Studies in Media Communication* 35, 4 (Aug. 2018), 334–349. doi:10.1080/15295036.2018.1488082 Publisher: Routledge _eprint: https://doi.org/10.1080/15295036.2018.1488082.

[96] Tianling Xie and Iryna Pentina. 2022. Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika. ScholarSpace, University of Hawai'i at Mānoa. http://hdl.handle.net/10125/79590

[97] Greyson K. Young. 2022. How much is too much: the difficulties of social media content moderation. *Information & Communications Technology Law* 31, 1 (2022), 1–16. arXiv:https://doi.org/10.1080/13600834.2021.1905593 doi:10.1080/13600834.2021.1905593

[98] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. 2025. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25).* Association for Computing Machinery, New York, NY, USA, Article 13, 17 pages. doi:10.1145/3706598.3713429

[99] Douglas Zytko, Nicholas Furlo, Bailey Carlin, and Matthew Archer. 2021. Computer-Mediated Consent to Sex: The Context of Tinder. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1 (April 2021), 189:1–189:26. doi:10.1145/3449288