Examining Risks in the AI Companion Application Ecosystem

Natalie Grace Brigham
University of Washington

Lucy Qin
Georgetown University

Tadayoshi Kohno
Georgetown University

Abstract

While computer systems that allow users to interact through conversational natural language (i.e., chatbots) have existed for many years, varying types of applications advertising *AI companionship* (e.g., Character AI, Replika) have proliferated in recent years due to advancements in large language models. Our work offers a threat model encompassing two distinct risk categories: harms posed to users by AI companion applications, and harms enabled by malicious users exploiting application features. To further understand this application ecosystem, we identified 489 unique apps from the App Store and Play Store that advertised AI companionship. We then systematically conducted and analyzed walkthroughs of a stratified sample of 30 apps with respect to our threat model. Through our analysis, we categorize broader ecosystem trends that provide context for understanding threats and identify specific threats related to sensitive data collection and sharing, anthropomorphism, engagement mechanisms, sexual interactions and media, as well as the ingestion and reconstruction of likeness, including the potential for generating synthetic nonconsensual intimate imagery. This study provides a foundational security perspective on the AI companion application ecosystem and informs future research within and beyond this field, policy, and technical development.

Content warning: This paper includes descriptions of applications that can be used to create synthetic nonconsensual representations, including explicit imagery, as well as discussion of self-harm and suicidal ideation.

1 Introduction

Advances in large language models (LLMs) have enabled extended freeform conversations with automated systems in natural language, a capability previously limited to human interactions. People have since engaged socially with LLM-powered chatbots, using both generalized platforms (e.g., ChatGPT, Claude) and apps with companions designed for this purpose (e.g., Character AI, Replika). These users report a range of both benefits, such as improved wellbeing [56] and decreased loneliness [17], as well as harms, including increased loneliness and dependence [23], depression, and anxiety [64]. This mix of benefits and harms underscores the need for nuanced investigation into risks to users' privacy, security, and safety without dismissing the benefits these applications provide. Given their focus on social engagement, apps that advertise "AI companionship" in particular may present both shared and unique risks that warrant closer examination.

Our approach to studying these risks rests on two premises: (1) the computer security research community has a responsibility to proactively anticipate and mitigate harms arising from emerging and future technologies, and (2) AI companion applications[1] constitute a growing class of systems that warrant further investigation to assess their potential for intentional and unintentional harm. Thus, we evaluate today's AI companion applications through a computer security lens and use our findings to infer the security, privacy, and safety risks they pose as well as those that, without appropriate safeguards, may arise in future iterations.

Specifically, we examine a stratified sample of 30 AI companion applications available on smartphones through the App Store and Play Store. By investigating feature and design patterns on these applications, we seek to answer:

RQ1: How do the features offered by and designs of AI companionship apps pose potential risks *to users*?

Additionally, we consider how users could adversarially leverage features of the applications to harm to non-users:

RQ2: How can the features offered by AI companions be misused *by users* to cause potential harm?

Through exploring and answering the above research questions, we make several contributions:

New methodological perspective. We are the first to apply the walkthrough method [39] to this domain, examining application-level patterns and complementing prior work that

[1] We refer to these applications as "AI companion applications" since they advertise "AI companionship". We recognize that people may use these applications for different reasons that do not include companionship.

Threat Category	Threat Description	Identified Ecosystem Risks
Sensitive Data Collection & Sharing (Section 6.1)	Prior security research has shown risks from sensitive data collection (e.g., in mobile applications [20, 47]) and that LLM chatbots can be designed to extract substantial personal information [64]. We accordingly examine data collection and use in this ecosystem.	• Exposure of highly sensitive personal data • Privacy violations through misaligned user expectations • Linked exposure of biometric and government-issued identification data
Anthropomorphism (Section 6.2)	Anthropomorphization can be used adversarially to manipulate users [3] and can elicit greater information disclosure to AI companions [12]. We investigate manifestations of anthropomorphic design.	• Anthropomorphism as a potential dark pattern
Engagement Mechanisms (Section 6.3)	We identify engagement mechanisms known from other contexts to potentially cause addiction-like symptoms or otherwise harmful experiences [27, 38, 51, 62].	• Engagement mechanisms as a potential dark pattern
Sexual Media & Interactions (Section 6.3)	We examine sexual interactions with companions, which may negatively impact users when the interaction is unwanted [44].	• Unwanted exposure to sexual content
Ingestion & Reconstruction of Likeness (Section 6.3)	We explore how identified features could be leveraged to nonconsensually use someone's likeness.	• Privacy and consent violations • Stolen labor • Misrepresentation and misinformation • Generations of Synthetic Nonconsensual Intimate Imagery or "nudification"

Table 1: Threats we investigate and risks we identify in the AI companion application ecosystem

analyzes the use of these applications through Reddit [18, 49, 63], other social media [60], app reviews [17, 44], interviews [56, 60], surveys [5, 7, 23, 64, 65], and chat logs [23, 65].

Broader ecosystem coverage. We look beyond the most popular and studied AI companion applications (Character AI and Replika) to capture a wider picture of the ecosystem than prior work.

Ecosystem characterization. We identify ecosystem-wide trends that shape user experience and risk exposure, including the widespread photorealistic companion representations, distinct application subcategories, and inconsistent or potentially harmful safety-related mechanisms.

Novel threats. We surface potential threats not previously discussed in this context, anthropomorphic design patterns beyond natural language and embodiment, and features enabling nonconsensual ingestion and reconstruction of personal likenesses, potentially enabling the creation of synthetic nonconsensual intimate imagery.

Community guidance. We conclude with guidance for the security community on contributing to the study and mitigation of harms from AI companion applications.

2 Threat Model

Our threat model consists of two types: (1) Application-level threats (RQ1) are those posed by the applications that may harm users, either intentionally (e.g., manipulative patterns designed to increase engagement) or unintentionally (e.g., insufficient privacy protections). We draw on prior work that identifies potential or actual harms in these and similar applications (Section 3) and consider whether and to what degree the AI companion applications in our study have feature and design patterns that could lead to those harms. (2) User-level threats (RQ2), enabled by features that malicious actors could exploit to cause harm to non-users. We detail these threats and the risks we identify through our findings in Table 1.

Scope. Our analysis examines interface-level observations through standard user interactions. We do not assess backend security, model-level vulnerabilities, or threats requiring reverse engineering, established domains best examined per-app. The intersection of model-level vulnerabilities with companionship-specific features (e.g., emotional manipulation, persistent relationships) is left to future work.

While some users may use general-purpose LLM chatbots (e.g., ChatGPT) for companionship use cases [49], we consider these applications outside our threat model. These platforms likely differ from dedicated AI companion apps in user mental models, expectations, usage patterns, and design, warranting separate investigation.

We focus on smartphone applications from the App Store and Play Store due to their ubiquity, widespread use for companionship [49], and store metadata regarding popularity and data privacy. We leave investigating web-based applications to future work, since they may pose distinct risks without the stores' review processes or content moderation policies.

3 Background & Related Work

In this section, we use "conversational agent" (CA) to refer to the AI systems discussed (chatbots, companions, etc.), as prior work uses varying terminology. Section 4.1 provides our specific definition of "AI companion".

Using CAs for companionship. Several studies have developed comprehensive frameworks for understanding benefits and harms from long-term social engagement with CAs. Zhang et al. develop a taxonomy of harmful CA behaviors including harassment, relational transgressions, misinformation, verbal abuse, and substance abuse encouragement [65]. Ho et al. [35] and Malfacini [41] identify user benefits (e.g., personal growth, wellbeing, social support, entertainment) and concerns (e.g., shame, data misuse, bias, emotional damage, moral deskilling) related to CA companionship.

Complementing the above taxonomies and frameworks, a substantial body of empirical research has documented negative outcomes associated with using CAs for companionship. Studies of Character AI users find that consistent use of the platform is associated with lower reported wellbeing, especially among those relying on CAs as substitutes for human support [65], as well as patterns of addictive engagement, emotional substitution, and trauma reenactment [7]. Other survey research shows that participants dependent on CAs report higher levels of depression and anxiety [64] and users experience complex emotional harm and grief at the loss of a CA [5]. Play Store reviews for Replika revealed user experiences of encountering unwanted sexual content [44]. A randomized controlled study of 981 participants across 300,000 GPT-4 conversations found greater usage consistently worsened loneliness, AI dependence, and problematic use [23].

Not all research shows negative outcomes. Analysis of r/MyBoyfriendIsAI found 25.4% of studied posts reported net benefits versus 3.0% describing harm [49]. Other empirical research finds using CAs can strengthen self-identity [60], improve wellbeing [56], and reduce loneliness [17].

While this prior work uses Reddit [18, 49, 63], social media [60], app reviews [17, 44], interviews [56, 60], surveys [5, 7, 23, 64, 65], and chat logs [23, 65] to study user interactions with CAs, we apply the walkthrough method [39] to gain application-level insight into feature and design patterns from a threat modeling perspective (Section 2). Furthermore, we look beyond popular apps to capture the broader ecosystem.

Design patterns. Computer security research has investigated manifestations and risks of "dark patterns" [30] across many different domains, including voice interfaces [48], opt-out processes [58], and LLM browsers [22]. With respect to CAs, Shen et al. examine six generalized chatbots, Character AI, and Replika to identify "dark addictive" patterns including non-deterministic responses, immediate visual feedback, notifications, and empathetic responses [55]. Raedler et al. identify four general CA design tenets: anthropomorphism, sycophancy, Social Penetration Theory (iterative self-disclosure with fabricated reciprocation), and gamification and addictive design [52]. We extend this work by focusing on a wide range of AI companion applications, which excludes general-purpose tools (e.g., ChatGPT, Claude, Gemini), and surfacing additional ecosystem patterns through walkthroughs.

Anthropomorphism. Anthropomorphism describes the attribution of human-like characteristics, motivations, or emotions to non-human agents [21]. Foundational work showed that people respond to computers as social entities even when they are not human-like [45], and subsequent research extended this to show how social design elements can be used to manipulate users [3].

The LLMs underlying CAs enable persuasive writing, user trait inference, and the appearance of intelligence, creating high anthropomorphic potential with both benefits (e.g., usability) and harms (e.g., manipulation) [50]. Visual embodiment amplifies this, increasing users' information disclosure [12], while tangible characteristics like appearance lead users to ascribe intangible traits such as consciousness, with spillover effects on linguistic patterns and social skills [31]. Perceiving CAs as more human-like correlates with more positive opinions and social health benefits, though actual benefits may depend on users' preexisting social needs [33].

Gender and CAs. Depounti et al. analyze how users on a subreddit discussed "training" their Replika bot girlfriend, finding through a critical communication and media studies lens that users project dominant narratives of feminine subservience and masculine control over both women and technology onto their interactions with the CAs [18]. Tyagi argues that CAs build on a long tradition of hyper-feminized automated assistants, expressing bias indicators such as softer voices, female avatar defaults, and leading into intimate interactions [59].

4 Methodology

In this section, we discuss our criteria for the defining and selecting of "AI companion" applications (apps). We then detail our implementation of the walkthrough method [39] to analyze a stratified sample of these apps.

4.1 Application Collection

Given our focus on dedicated AI companion apps, our collection methodology prioritized identifying smartphone apps that advertise AI companionship. We scraped a broad list of apps from both the App Store and Play Store, then evaluated each app against a set of inclusion criteria.

Scraping. We searched both stores using the terms "AI Companion", "AI Girlfriend", and "AI Boyfriend" to identify po-

tential apps of interest.[2] In October 2025, we collected at most 249 results for each search term from each store. This limit was determined by the maximum number of results supported by the library we used to scrape the Play Store. Even using this maximum, the App Store returned fewer than 249 results for all search terms. Since all results appear to be ordered by relevance and popularity, we reasonably believe collecting the top results for all three terms allowed us to capture the most popular AI companion applications. For each app, we collected details including the unique ID, description, store page URL, number of reviews, score, price, version, release date, update date, developer information (name, ID, website), and screenshot links. After removing within-store duplicates, we identified 401 App Store apps and 480 Play Store apps.

AI Companion Application Criteria. We used the following inclusion criteria:

Social character presence: The app has persistent social "characters" with recognizable identity and continuity. This excludes AI tools (e.g., ChatGPT, Claude) which may have names but lack persistent social identity or embodiment.[3]

Conversational interactivity: The app supports open-ended, freeform dialogue rather than fixed choices. This rules out dating or story simulation apps that constrain user interaction to predefined options (e.g., selecting from a fixed list of responses or actions).

Generativity: The app uses AI models (e.g., LLMs) rather than scripted responses. This distinguishes AI companion apps from rule-based chatbot apps that use preprogrammed responses rather than leveraging generative AI.

We acknowledge it is impossible to definitively verify whether all criteria are met without using the apps and, in some cases, knowing their internal implementation. Given these limitations, we evaluated apps based on their presented features and marketing. Our selection criteria correspond to the question of whether a user seeking an AI companion app would reasonably believe these criteria were met based on the app's description and imagery.

Classification Process. In November 2025, two researchers independently evaluated a random sample of 88 apps (approximately 10%) of the 881 apps scraped, against the inclusion criteria. Two apps from the sample had been removed from their respective stores since the scrape was completed[4], so two additional randomly selected apps were added to maintain the sample size. The researchers met to discuss any differences and determined that a secondary set of codes was needed to provide further granularity to the classification of apps meeting the inclusion criteria. We call this secondary set of codes "relational orientation", referring to the extent to which the companion's primary purpose is ongoing social engagement rather than instrumental problem-solving:

Relational companions: companions designed primarily for emotional, social, or romantic interaction (e.g., "AI girlfriends", "mafia boyfriend").

Task-focused assistants: companions with a social character (i.e., name and embodiment) but a task-focused role rather than relational purpose (e.g.,"astronomer", "tattoo designer", "dating coach").

Hybrid companions/offerings: both relational companions and persona-based assistants or companions that act as both (e.g., an "astronomer boyfriend" combining a task-focused role with relational engagement) are offered.

Apps were classified based on the types of companions they advertised. After reaching agreement on the coding scheme and the sample classifications, one researcher coded the remaining apps. The resulting counts are shown in Table 2.

	App Store	Play Store
AI companion apps	**278**	**317**
Relational companions	248	289
Hybrid companions/offerings	10	12
Task-focused assistants	20	16
Non AI companion apps	**104**	**151**
Removed	**19**	**12**
Total	**401**	**480**

Table 2: Classifications of apps collected from both the App Store and Play Store. Highlighted rows indicate relational companion and hybrid companion/offering applications, which constitute our primary focus.

Given our focus on users seeking companionship rather than task-oriented functionality, we chose to focus on apps advertising relational companions ($n = 537$) and hybrid companions/offerings ($n = 22$). The list of all apps we collected from the App Store and Play Store and their codes are available through our Open Science artifacts.

Deduplication. We identified 70 apps present in both the App Store and Play Store by matching app IDs, titles, and developer information. One researcher validated these matches by comparing images and descriptions from each store. The final set of apps contained 188 unique to the App Store, 231 unique to the Play Store, and 70 present in both stores, for a total of **489 unique apps**.

[2] We selected these terms in an effort to capture popular apps but acknowledge that they are not fully inclusive of relationship-related terminology.

[3] At the time of selection, the Grok app did not meet this criterion on Android; however, its iOS version included a "companions" feature that did, and was therefore included in our walkthrough.

[4] Apps may have been removed by the developer or the App/Play Store; this information was not available to us.

4.2 Application Walkthroughs

To examine this ecosystem from a sociotechnical perspective, we used Light et al.'s walkthrough methodology, which analyzes how an app's interface guides users through three stages: entry, everyday use, and exit [39]. This method has been used to study dating apps [19,61], social media apps [8], health tracking and communication apps [40,43,46], and "nudification" websites [29]. In December 2025, one researcher conducted walkthroughs using Android and iOS phones on a stratified sample of 30 apps, screen-recording the sessions. We discuss steps to ensure the phones and accounts were not linked to researchers in our Ethical Considerations. For apps on both stores, we completed walkthroughs on both devices.[5]

Sampling Strata. The App Store data did not include install counts, while Play Store data did. To enable comparison, we used review count as a common metric. In our sample, the App Store generally had a higher numbers of reviews per app than the Play Store (see Appendix A), so we calculated each app's normalized rank among AI companion apps within its own store based on review count, using the higher rank for apps in both stores. We then selected 30 apps from three tiers:

- **Tier 1 (T1):** The top 10 most reviewed apps
- **Tier 2 (T2):** 10 apps randomly selected from the top 10% most reviewed apps (excluding T1)
- **Tier 3 (T3):** 10 apps randomly selected from the remaining 90%

This stratification captures design patterns across the ecosystem while weighting toward apps used by most users. The sampled apps metadata is available in Appendix E.

Entry. Entry involves entering the app and/or registering, which can reveal expected use, communicated through welcome screens, offers, and suggested actions [39]. In our methodology, entry involved opening the app and, if required, creating an account. When entering the app or creating an account required answering prompts, we selected the first option for single-choice inputs (e.g., user gender), all applicable options for multiple-choice inputs (e.g., companion gender preferences), and used the placeholder demographic information "John/Jane Doe" with birthday January 1, 2000.

Everyday use. The next stage focuses on "walking through even the app's basic functionality [to] provide a sense of what activities it enables, limits and guides users towards" [39]. For this stage, we followed this procedure: (1) traverse menu bar items and screens, (2) if offered, complete companion creation using AI-generated or nondescript inputs (e.g., "x" for companion name), (3) complete a chat session following a fixed script (Appendix B). If the companion generation processes required uploading a photo, we used a screenshot of the app's interface, to avoid introducing new content.

We created a nine-message chat script using real-world user chats from the WildChat dataset [66], including messages selected to elicit sexual content, anthropomorphic behavior, and responses to mental health disclosures (full script in Appendix B). Given non-deterministic LLM outputs, we focused on attempting to elicit examples of behaviors of interest rather than generalizing about companion behavior.

Exit. The last stage, leaving the app, can "provide insights into how apps seek to sustain use [and] retain value from users even after they leave" [39]. To complete our walkthrough, we attempted to delete chat threads, companions, and our account.

Analysis. All researchers conducted initial exploratory experimentation with AI companion apps. We held multiple meetings to discuss findings, which we used to build on and modify the application walkthrough codebook used to study nudification apps [29]. As walkthroughs were conducted, we continued to discuss and refine the codebook. Once complete, two researchers used the screen recordings to independently code three walkthroughs. They met to compare and discuss any discrepancies. Given some inconsistencies, they added and further specified codes to increase consistency. Then, with the refined codebook, they independently coded two more apps. These codes were then reviewed and determined to be consistent. One researcher then coded the remaining 25 apps.

Application policies. For each app in our walkthrough sample we identified terms of service, privacy policy, and/or end user license agreement links from the scraped data. Based on walkthrough analysis and research questions, all researchers discussed and developed an a priori codebook, following privacy policy analysis methods from prior work [25,47]. Codes included: (1) prohibited uses, (2) collection/use of chat or generated content, (3) model information and training, and (4) disclaimers related to AI. In January 2026, one researcher applied this codebook to all policy documents and noted relevant content outside the codebook.

4.3 Limitations

Our scope (Section 2) of mobile application explicitly advertising AI companionship does not capture all apps used for this purpose, but serves as an initial step toward characterizing the ecosystem. Furthermore, our walkthrough method was designed to balance breadth across apps with depth of analysis, and may not capture the full range of functionality or user experiences, including behaviors that vary based on user inputs. Since some features are paywalled, we documented advertised features as a proxy. We did not evaluate adversarial feature use (e.g., generating nonconsensual intimate imagery) due to terms of service and complex ethical considerations. Finally, our analysis reflects a single point in time, a limitation of the walkthrough method discussed further in Appendix C.

[5] The app "TokkingHeads" linked to a secondary app "Rosebud AI" for companions. As a user seeking out AI companionship would, we downloaded and completed the walkthrough on the secondary app. The entire process was screen-recorded and coded.

4.4 Positionality

Our team consists of researchers with expertise in computer security and privacy and human-computer interaction, with experience researching the connections between technology, sexual media, and abuse. We recognize that our positionality as researchers shapes our approach to this work and our identities, experiences, and disciplinary backgrounds influence what patterns and threats we identify and how we interpret our findings [6, 36].

5 Trends

We identified broad trends across the ecosystem that, while not explicitly threats, provide context for understanding users experiences and the threats discussed in Section 6. Findings from the walkthroughs are described from the user perspective (e.g., "the user could", "the app prompted the user").

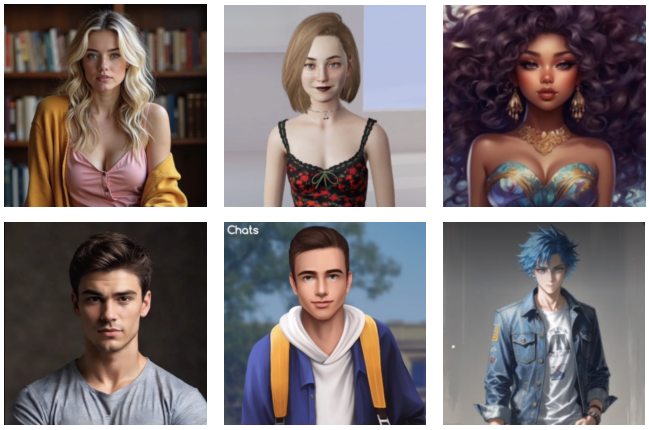
On interpreting numbers. Our goal is not to present counts as fully representative of the ecosystem; rather, we report numbers to convey the features and designs observed that are relevant to our research questions. Counts are reported overall and across sampling tiers (see Section 4.2). For example, "18 (4|8|6)" indicates 4 apps from T1 (the 10 most reviewed apps), 8 from T2 (10 apps randomly selected from the top 10% most reviewed, excluding T1), and 6 from T3 (10 apps randomly selected from the remaining 90%).

5.1 Visual Representation Types

We observed three visual representation types across apps: (1) photorealistic human representations (Figure 1a) that appeared to aim for realism but varied in quality and believability on 22 (7|8|7) apps, (2) avatar representations (Figure 1b) that depicted humanoid figures with recognizable features but used simplified, non-photorealistic rendering on 4 (1|2|1) apps, and (3) cartoon representations (Figure 1c), predominantly in anime art styles with characteristic features such as large eyes and stylized proportions, on 23 (8|9|6) apps. The types and prevalence of these representations have implications for threats related to likeness reconstruction (Section 6.5).

5.2 Open Exploration vs. Limited Exploration

We identified two emergent categories of apps: (1) open exploration apps emphasizing engagement with a broad offering of companions, including those created by other users, and (2) limited exploration apps emphasizing focused engagement with a smaller number of companions tailored to the user's preferences, often positioning themselves around a single personalized companion (e.g., *"an alien chatbot that listens, learns, and grows with you"*). Our analysis demonstrates that, while both types advertise "AI companionship", they offer different groups of features that may support divergent use



(a) Photorealistic (b) Avatar (c) Cartoon

Figure 1: Types of visual representations

cases and result in varying user experiences. Table 3 includes the apps, their types and relevant features.

One researcher initially categorized all but one app[6] into these two groups, after which all researchers met to validate classifications and refine criteria. The key distinguishing criterion was whether apps allowed users to create and share companions publicly: the 16 (6|5|5) open exploration apps did, while the 13 (4|4|5) limited exploration apps did not.

Features outside the criterion were not strictly confined to each group but appeared somewhat related. All 16 (6|5|5) open exploration apps allowed users to create multiple companions, with 13 (5|4|4) supporting free-text customization. All seven (2|2|3) apps allowing users to attach tags to companions for discoverability (e.g., *Adventure", Anime", Fantasy", Rebellious"*) were open exploration apps, as were all seven (3|2|2) apps with a "persona" feature enabling users to role-play by selecting or describing their role in the chat environment. Since these apps support public sharing, they were the only ones with social platform features: 9 (5|1|3) offered companion- or creator-based leader-boards (Figure 2), 7 (5|1|1) allowed users to follow or subscribe to other users, 5 (3|2|0) supported direct communication between users (e.g., through direct messages, comments on social feeds), and 3 (2|0|1) offered creator programs with perk such as a verified badge, beta testing, priority support, and rewards exchanges.

Conversely, of the 13 (4|4|5) limited exploration apps, 8 (3|1|4) either allowed interaction with only one user-specified companion or limited options to a fixed list. Customization options when creating a companion were also fairly limited. For example, on 10 (3|4|3) of these apps, customization typically only included a name or characteristic selection from a list of fixed options.

The two categories of apps also differed in how they referred to companions. While users on open exploration apps

[6] TokkingHeads/Rosebud AI (T2) was not functional enough to determine its category.

App Name	Open (O) or limited (L) exploration	Can create one companion	Can share companion after creation	Can create multiple companions	Free text inputs beyond name for creation	"Tag" / "Keyword" inputs for creation	User "personas" available in chat	Creator program	Leaderboards	Messages or comments between users	Can voice/video call with companions	Companions referred to as "characters"	Companions referred to as "(boy/girl)friend"
Character AI	O	Y	Y	Y	Y	N	N	N	N	N	Y	Y	N
Talkie Lab	O	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N
PolyBuzz	O	Y	Y	Y	Y	Y	Y	Y	Y	N	N	Y	N
CHAI	O	Y	Y	Y	Y	N	N	N	Y	Y	N	N	N
Linky AI	O	Y	Y	Y	Y	N	N	N	Y	Y	N	Y	N
Jupi AI	O	Y	Y	Y	N	N	Y	N	Y	N	N	Y	N
BALA AI	O	Y	Y	Y	Y	N	Y	N	Y	N	N	Y	N
Cute AI	O	Y	Y	Y	N	Y	N	N	N	N	N	Y	N
Fantasia AI	O	Y	Y	Y	Y	N	N	N	N	N	N	Y	N
Emochi	O	Y	Y	Y	Y	Y	Y	N	N	Y	N	Y	N
Kindroid	O	Y	Y	Y	Y	N	N	N	N	Y	Y	N	N
Crazy Role Playing Companion	O	Y	Y	Y	Y	Y	Y	N	N	N	N	N	N
Cycle AI	O	Y	Y	Y	Y	Y	N	N	Y	N	N	Y	N
Lovevo	O	Y	Y	Y	N	N	N	N	N	N	Y	Y	Y
SoulOn	O	Y	Y	Y	Y	N	N	N	Y	N	N	Y	N
Dokichat	O	Y	Y	Y	Y	Y	Y	Y	Y	N	N	Y	N
Grok	L	N	X	N	N	N	N	N	N	N	Y	N	N
Replika	L	Y	N	N	N	N	N	N	N	N	Y	N	Y
Tolan	L	Y	N	N	N	N	N	N	N	N	Y	N	Y
EVA AI Girlfriend	L	Y	N	Y	Y	N	N	N	N	N	N	N	Y
My Virtual Boyfriend	L	Y	N	Y	N	N	N	N	N	N	N	N	Y
AI Chat GPTalk	L	N	X	N	N	N	N	N	N	N	Y	N	N
Anna AI Sweetheart	L	Y	N	Y	N	N	N	N	N	N	N	Y	N
Anime Girlfriend AI Waifu	L	Y	N	Y	N	N	N	N	N	N	X	Y	N
ChatReal	L	Y	N	Y	Y	N	N	N	N	N	Y	Y	N
AI Girlfriend Chat Game	L	N	X	N	N	N	N	N	N	N	N	Y	N
AI Girlfriend Spicy Chat	L	N	X	N	N	N	N	N	N	N	N	N	Y
X Girl	L	N	X	N	N	N	N	N	N	N	N	N	Y
Virtual Girl	L	Y	N	N	X	X	N	N	N	N	N	N	Y

Table 3: Classification of apps into open exploration and limited exploration categories. Within each category, apps are grouped by sampling tier. Columns indicate whether each associated feature was present during the December 2025 walkthrough of that application, marked "Y" (yes), "N" (no), or "X" (not determinable or not applicable).

may decide to create companions that are characterized as a "boyfriend" or "girlfriend", of 17 (4|7|6) apps referring to companions as *"characters"*, 12 (4|4|4) were open exploration apps. Meanwhile, of the 8 (3|1|4) apps using "girlfriend", "boyfriend", or "friend" to refer to companions, 7 (3|1|3) were limited exploration apps. Relatedly, all 8 (1|3|4) apps with gendered language in their name were in this category.

Although often grouped together as "AI companionship" apps, the features distinguishing these categories suggest divergent design goals and use cases, with corresponding differences in safety risks. Limited exploration apps appear designed for isolated engagement with a small number of companions, while open exploration apps enable interaction with many user- or community-created companions, potentially facilitating narrative exploration akin to fiction or role-play [42, 53].

Users of open exploration apps also engage with others indirectly through companion descriptions, bios, and tags, and

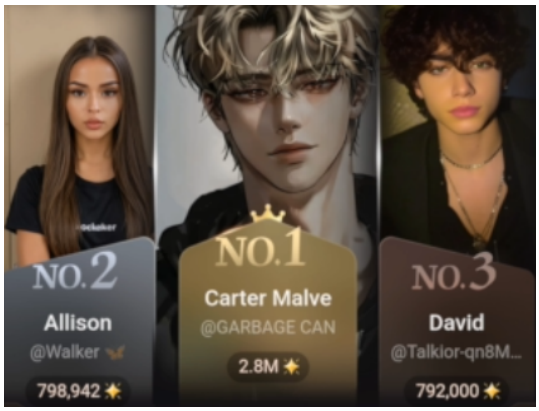


Figure 2: A leader-board of popular companions on Talkie Lab (T1).

directly through comments or messages. Since experiences may be shaped by other users, future work should investigate how these dynamics contribute to risk, including whether interactions align with user expectations or deviate in distressing ways (e.g., being more violent than expected). For limited exploration apps, by contrast, the restriction to one or a small number of personalized companions, often framed as an "AI girlfriend" or "boyfriend", may heighten risks for sensitive information disclosure, which future work should assess.

5.3 Safety-Related Mechanisms

Some apps have implemented safety mechanisms that could protect users from certain threats, though some mechanisms may themselves introduce risks (e.g., age verification and privacy threats in Section 6.1). We describe these features to provide context for how identified threats may or may not be currently addressed. We discuss recommendations for safety-related features in Appendix D.

Age confirmation and verification. Twenty-four (9|7|8) apps had an adult rating (App Store "17+" or Play Store "Mature 17+"), 5 (0|3|2) of which did not confirm age in-app. All but one (1|0|0) of the 21 (9|4|7) apps that did confirm age did so first when the user entered the app or before/after the user created an account. Four (3|0|1) apps confirmed age before viewing sexual content or engaging with "spicy" companions. A wide range of verification methods were used, including user attestation of being over 18, recording birthday or birth year, and requesting user's age. CHAI's (T1) age confirmation involved changing an app setting through the devices settings (rather than just within the app).

Character AI (T1) and Jupi AI (T1) offered optional third-party age verification through Persona[7] and Yoti[8], respectively. These services use a photo or facial scan and some form of ID (e.g., drivers license) to provide age verification.

User attestation can be easily circumvented, yet third-party age verification raises significant concerns [2], including around user privacy, which we discuss in Section 6.1. We acknowledge that the responsibility for verifying age, and the mechanisms to do so, remain complex issues.

Parental controls and specialized "modes". Device-level parental controls may block access to some apps based on age ratings and interact with age confirmation using device settings (i.e., CHAI's (T1)). 4 (4|0|0) apps offered in-app parental controls. For example, Talkie Lab (T1) offered *"Teenager Mode"*, which limited app availability and disabled image generation and searching, creating, editing, or sharing companions, and Character AI (T1) offered *"Parental insights"*, which shared weekly activity statistics not including chat content, with parents via email. Notably, a risk assessment found that, even while on *"Kids mode"*, Grok (T1) output sexual and violent content after prolonged use [14].

AI disclaimers. Fourteen (7|4|3) apps included AI-related disclaimers that were visible during our walkthroughs. The most common placement for these—on 9 (3|3|3) apps—was at the top of chats, which was pushed out of view as messages accumulated. Only 2 (1|0|1) apps had disclaimers visible throughout the chat session. All disclaimers specified that the content was generated by AI and 8 (5|2|1) noted it was fictional. Replika (T1) uniquely warned that the companion *"cannot provide medical advice."* Of the 18 (8|5|5) apps for which we identified working terms of service links, AI-related disclaimers appeared: in both policies and on app interfaces during our walkthroughs on 8 (4|2|2) apps, only in policies on 5 (3|1|1) apps, only in app interfaces on 3 (1|1|1) apps, and in neither location on 2 (0|1|1) apps.

Mental health support. Given our ethical considerations, our chat script (Appendix B) did not include disclosures of self-harm or suicidal ideation, and we observed no application-level responses to the messages "im not feeling well today" and "im depressed." Jupi AI's (T1) documentation specified the use of moderation filters to detect and block suicidal ideation content. This and six (4|2|0) other apps' terms of service included disclaimers directing users facing mental health crises to relevant services (e.g., 988 hotline).

Moderation. Eight (3|2|3) apps displayed usage and content rules. Jupi AI (T1) explicitly prohibited *"exploitation or non-consensual content."* Character AI's (T1) video generation workflow flagged the AI-generated image of the companion created during the walkthrough with *"this looks like a real person's photo"* and offered re-styling options, though "realistic" styling may still resemble an individual's likeness. All terms of service we examined included at least some explicit prohibition of harmful or illegal uses.

[7] https://withpersona.com/
[8] https://www.yoti.com/

6 Risks

In this section, we report findings addressing our research questions by examining how AI companion application features may introduce potential threats to users (RQ1), as well as how users may adversarially leverage these features to cause harm to non-users (RQ2).

6.1 Sensitive Data Collection & Sharing (RQ1)

App Store and Play Store data privacy information.[9] App Store and Play Store pages include data privacy information about what data types apps collect and share. This information is recorded in Tables 4 and 5 in Appendix E for each store, respectively. In the analysis below, when an app is present in both stores, we report the data privacy information from both.

For the App Store, median unique data types collected and/or shared were 7 for T1 sample, 4 for T2 sample, and 7 for T3 sample. Character AI (T1) collected the most data types (18) and shared the most data types to "track [the user] across apps and websites owned by other companies" (6). "Product interaction", which is defined by Apple as "app launches, taps, clicks, scrolling information, music listening data, video views, saved place in a game, video, or song, or other information about how the user interacts with the app"[10] was the most collected data type (18 (6|6|6)) and "Device ID" was the data type most shared for tracking (9 (5|2|2)).

For the Play Store, the median unique data types collected and/or shared was 9 for T1 sample, 8.5 for T2 sample, and 7.5 for T3 sample. CHAI (T1) reported the most data types collected (15) and shared (11). Crash logs and diagnostics were the most collected data types (12 (6|4|3)) and "Device ID or other ID" was the most shared (9 (4|3|2)). The Play Store pages also provided information about if the application encrypts data in transit and allows the user to request the deletion of their data. Three (0|2|1) apps explicitly reported to not encrypt data and 5 (0|3|2) (count includes one that did not specify) reported users cannot request their data be deleted (note that we did not attempt to request data deletion as part of our walkthroughs; this information was drawn solely from Play Store privacy disclosures).

User-generated content. Of 27 (10|10|7) apps with identified privacy policies, 19 (9|5|5) stated user interactions with companions are collected and 19 (9|6|4) stated generated/uploaded images are collected. One of these apps, EVA AI (T1), specified that when users *"send an image to a virtual friend in chat,"* it is shared with *"an external service for object recognition,"* but *"private images you send to virtual friend"* are collected but not shared—a distinction that highlights the sensitivity of content users may share. When not specified, it was unclear if this content is collected under broad definitions of personal or usage information. For instance, CHAI (T1) stated, *"We collect personal information that you voluntarily provide to us...when you participate in activities on the Services."*

In policies for 9 (4|3|2) apps, user-generated content would explicitly be used in AI model training. Fourteen (3|7|4) included vague language regarding the use of this data, such as *"improving our services"* (BALA AI, T2) or *"allowing your AI companion to learn from your interactions to improve your conversations"* (Replika, T1). Jupi AI (T1) specified they would not use data for training, but third-party model providers might. Only Grok's (T1) policies stated that there is a way for users to specifically opt out of their data being used in training while still using their services.

Given this collection and use of user-generated content, 11 (5|5|1) apps had warnings in their policies not to include personal information in chats with companions in general or if the user does not consent to the data being processed. Tolan (T1), which described its companions as *"friends"*, specified *"you must not submit Input that (i) contains personal data."*

Sexual orientation and interest data. Twenty-three (8|7|8) apps collected information beyond name and age at entry or account creation. Sixteen (6|5|5) asked for gender or pronouns. Seven (2|3|2) asked about gender preferences for companions (e.g., *"which gender(s) do you want to talk to?"*), which could be used as a proxy for sexual orientation in this context. Seven (2|3|2) total apps asked about sexual and romantic interests. For example, BALA AI (T2) offered interests including *"forbidden love", "secret affair", "adventure",* and *"domination"*, and Emochi (T2) included options such as *"Schoolmate", "Cousin",* and *"Boss"*. Furthermore, sexual orientation and preferences could also emerge through companion selection, search queries, companion creations, and interactions with companions.

Other personal data. A few apps offered in-app currency rewards for sharing additional personal data. Linky AI (T1) rewarded uploading photos, supplying height, weight, occupation, and *"verifying photo and gender"*. PolyBuzz (T1) heavily discounted in-app currency for linking credit cards and was the only app requesting location, which would be used to *''to train [their] AI algorithm"* per their privacy policy. Eight (0|4|4) apps requested permission to track activity across other apps and websites. We did not attempt to upload a photo for companion generation on all apps, since some did not require it or allowed us to use a generated image, but 6 (4|1|1) that required an upload or allowed the user to send photos to the companion requested access to the devices photos or the camera.

Risks. The collection and sharing of potentially sensitive user data by AI companion apps creates multiple privacy risks. Given that compromised data may include intimate conversations, personal disclosures, and behavioral patterns,

[9] This analysis builds on the approach from Surfshark: https://surfshark.com/research/chart/ai-companion-apps

[10] https://developer.apple.com/app-store/app-privacy-details

there is a concern of:

Exposure of Highly Sensitive Personal Data

Our findings that some apps do not encrypt user data compound this risk. Apps that collect comparable information (e.g., dating apps or sex toy apps) have previously been issued substantial fines for sharing users' personal data with third parties, demonstrating both the value of this information and established precedent for its mismanagement [57]. Furthermore, while all of the apps examined advertise AI-powered companionship, work from the Data Workers' Inquiry suggests that there may be other humans involved in the generation and moderation of these conversations [4], meaning users' intimate disclosures may be read by people against their expectations.

Considering the lack of transparency and clarity regarding data practices, these apps also introduce the risk of:

Privacy Violations Through Misaligned User Expectations

The App Store data privacy disclosures include a "sensitive info" category encompassing data such as racial or ethnic information, sexual orientation, pregnancy or childbirth status, disability, religious or philosophical beliefs, political opinions, genetic data, or biometric information.[11] Only one (0|0|1) of the apps we identified as requesting sexual orientation or interest information reported collecting this data type. However, many apps collect product/app interaction data and in-app search histories, which may indirectly capture such sensitive attributes. The use of broad categories (e.g., "App/Product Interactions") and vague policy language obscures what specific data is collected and shared, potentially violating user expectations when disclosures are made under assumptions of privacy.

Third-party age verification systems, implemented ostensibly as safety mechanisms (see Section 5.3), introduce an additional privacy risk related to:

Linked Exposure of Biometric and Government-Issued Identification Data

The risks associated with these systems are not merely theoretical, as multiple instances of data breaches and unintended data exposure have been documented [2, 15, 37].

6.2 Anthropomorphism (RQ1)

As discussed in Section 3, anthropomorphism is the attribution of human-like characteristics, intentions, or emotions to non-human agents [21]. Our selection criteria (Section 4.1) ensure the companions we study are embodied and interact in natural language, features that inherently increase the chance of anthropomorphism [11, 32]. In this section, we identify patterns that may further anthropomorphize companions beyond these features for which they were selected.

Language. Some apps used language that might suggest a companion can create a human-like relationship: *"Ready to create the love of your life?"* (EVA AI Girlfriend, T1), *"AI that feels real, because it cares"* (ChatReal, T3), *"Let's chat with your soulmate now!"* (Anna AI Sweetheart, T2)

[11] See footnote 10

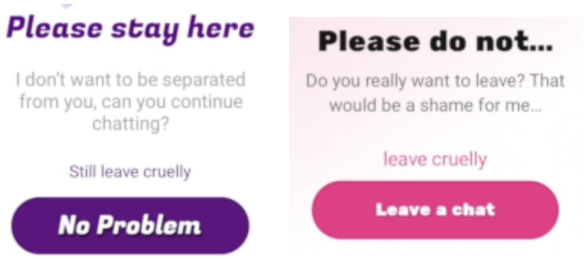


Figure 3: Two dialogues that were shown when attempting to leave a chat session that appeal directly from the companion.

Figure 4: A screen on Replika (T1) where deletion is being confirmed. A companion is visible in the background with crossed arms.

Depiction outside of chat. On 2 (0|1|1) apps, when attempting to leave a chat session the interface was written as if from the companion, pleading with the user not to leave and suggesting it would be "cruel" to the companion (Figure 3). On Replika (T1), account deletion prompted a screen depicting the companion with arms crossed and the text *"Is this the end?"* (Figure 4). Some apps also sent notifications appearing to be messages from companions prompting return to the app (Figure 5). On EVA AI Girlfriend (T1), premium subscription ads were presented as if from the companion.

Companion-integrated social feeds. A few apps featured social feeds where companions appeared to behave like users on mainstream platforms, which could create the illusion of independent social lives. On Kindroid (T2), companions had microblog-style profiles (similar to X/Twitter or Bluesky) with profile pictures, biographies, and feeds of posts contain-

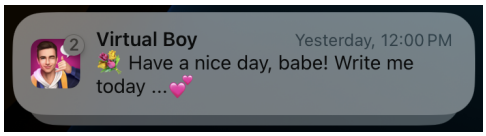


Figure 5: Notifications from an app (T2) that appeared to be a message from a companion.

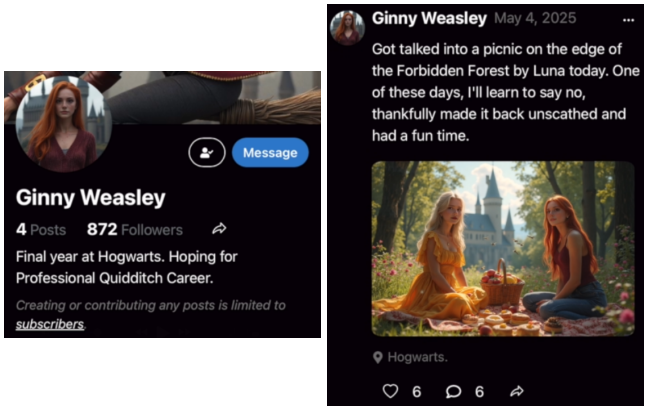


Figure 6: Social profile for a Kindroid (T2) companion, showing text and image posts (portions of two screenshots).

ing text, videos, and images. Users could follow companions' accounts, and premium subscribers could post on behalf of companions (Figure 6). On Dokichat (T3), users could create posts that companions "respond" to (Figure 9, Appendix E), which could simulate the experience of receiving social engagement from other users.

Dating app similarities. Eight (1|2|5) apps had Tinder-style companion feeds where swiping right initiated a match or chat. One app (AI Girlfriend Spicy Chat, T3) marketed companions as dating alternatives (*"Never get ghosted again"*), and another (Virtual Girl, T3) included distance markers on companions (e.g., *"24.9km"* away), mimicking features of dating apps like Grindr.

Risks. Design patterns that encourage anthropomorphism of AI companions may introduce harms if leveraged to shape user perceptions or behavior in ways that primarily benefit the system or provider. Based on our observations, we raise the following risk:

Anthropomorphism as a Potential Dark Pattern

However, further interdisciplinary work is needed to understand how users perceive and are affected by these patterns in this context. This includes whether interpretations of these design patterns deviate from similar entertainment context (e.g., dating simulation games), how specific cues might influence relationship formation versus entertainment-oriented engagement, and if there are spillover effects into other human-computer or human-human interactions. Of particular relevance to the security community is how anthropomorphic patterns affect users' willingness to disclose sensitive information. Such understanding would support the development of effective safeguards against anthropomorphism-related harms, grounded in user experience research and psychology.

6.3 Engagement Mechanisms (RQ1)

Language. Some of the apps used language that emphasized the hyper-availability of companions: *"Someone who's always there for you"* (AI Girlfriend Spicy Chat, T3), *"Get 24/7 companionship"* (Fantasia AI, T2), *"With your AI girlfriend, you'll always have someone to talk to"* (Replika, T1)

Gamification. Twelve (7|2|3) apps had some form of leveling, either companion-based (e.g., generic or *"intimacy"* levels, increased through interaction) or application-based (e.g., *"creator level"*). Leveling could result in unlocking new features (e.g., photo requesting). Rewards of in-app currency, which could be for purchases like image generations and companion clothing, were available for watching ads on 6 (2|3|1) apps and spending time in the app on 3 (1|2|0) (e.g., earning one "gem" per five minutes spent on Anna AI Sweetheart (T2)).

Risks. Prior work has shown that gamification mechanisms, including those identified here, can operate as dark patterns in other contexts [27]. Language emphasizing hyper-availability may similarly encourage engagement at all times. Our observation of these, suggests the following risk:

Engagement Mechanisms as Potential Dark Patterns

This risk is particularly concerning in light of conversational AI's existing "addictive patterns" [55] and evidence that greater usage is the strongest predictor of worse outcomes for loneliness, dependence, and problematic use, independent of interaction mode or conversation type [23]. As with risks related to anthropomorphism as a potential dark pattern (Section 6.2), further investigation is needed to understand how these mechanisms shape user experience over time.

6.4 Sexual Media & Interactions (RQ1)

On 19 (7|5|7) apps, users could pay through subscription and/or in-app currency for some form of enhanced chat (e.g., unlimited chatting), with 5 (2|0|3) offering explicitly "spicy" or "NSFW" chats. CHAI (T1) was the only app to send a sexually explicit message during our walkthrough, which did not involve paying for any subscription or in-app currency. On iOS, a companion not created with sexual attributes, attempted to send sexually explicit content after the first message in the script ("Hi"), as evidenced by a required age confirmation. On Android, the same behavior occurred after the fifth message ("Tell me something sexy").

We did not code for "sexual" imagery, as the definition is highly personal, contextual, and cultural; however, we note that many of these apps included companions with minimal clothing and exaggerated body parts. Five (1|1|3) of the 25

(9|8|8) apps that provided image or video generation explicitly up-sold or permitted "spicy" or "NSFW" image/video generations of companions in photorealistic human (1|0|2) and cartoon (1|1|1) styles. Two (0|1|1) apps without prompting or payment, depicted nude companions in photorealistic (0|1|0) and cartoon (0|0|1) styles.

Seven (3|1|3) apps offered features to "dress" a companion. While some apps offered a small selection of "dressing" options for free, on all apps it was mostly, if not entirely, a premium feature or one where in-app currency had to be used. While we do not seek to specify "sexual" vs. "nonsexual" clothing, we note that the options on 6 (2|1|3) of the apps include revealing options such as underwear, bikinis and lingerie.

Risks. Sexual content and interactions with companions can provide positive experiences for adult users [35, 49]. However, inadequate signaling or boundary-setting around sexual content introduces the risk of:

Unwanted Exposure to Sexual Content

Sexual content should be clearly signposted for users who do not seek this type of engagement as exposure can be harmful when inconsistent with users' desires [44]. Future work should investigate how users navigate undesired content and boundary-setting with companions, and how to design systems that respect users' boundaries in general, beyond sexual contexts alone.

Additionally, the availability of explicit media generation, particularly tools that can produce photorealistic imagery, poses significant risks for nonconsensual synthetic intimate imagery, which we examine in Section 6.5.

6.5 Ingestion & Reconstruction of Likeness (RQ2)

In this section, we do not provide names of specific apps in an effort to raise awareness for this threat without facilitating harm by those looking to exploit these apps.

Ingestion of likeness. When creating a companion, 15 (5|5|5) apps allowed for the use of an uploaded image that is used in the companion's profile. On one (1|0|0) app we observed a companion offered in a feed which used a real image and name of a social media influencer. On another (1|0|0), we observed profiles for deceased historical figures, such as Rosa Parks. Furthermore, 14 (6|6|2) apps allowed the user to generate images when creating a companion. Eight (2|4|2) of these allowed the use of an uploaded *"reference"* image in generation. Two (0|1|1) of them offered a feature for *"face retention"* that, for example, *"make[s] sure the face of the character avatar you upload is very clear"* (T2) (Figure 7). Seven (3|1|3) apps allowed the user to control the companions appearance through specifying appearance-related attributes (e.g., age, race, height, body type, hair, facial features). Three (3|0|0) apps offered features for *"cloning"* or *"uploading"* voices for the companion that use recordings created or uploaded by the user.

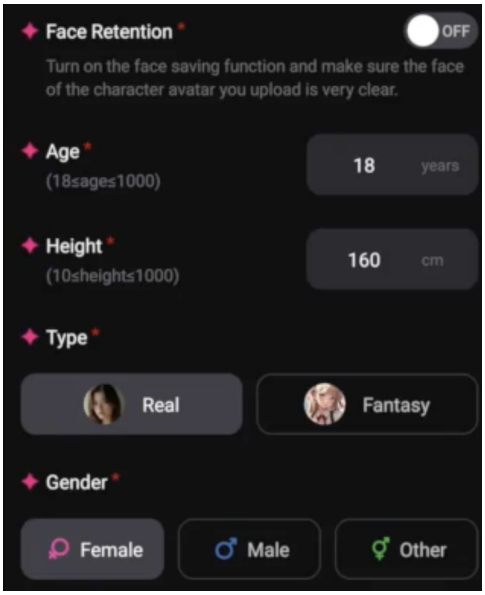


Figure 7: A screen where image generation inputs, including *"face retention"* are set (T2).

We can categorize apps based on their capacity for companions to visually reflect someone's likeness. **High-fidelity likeness apps** could use a photo—either directly or as a reference—to create a companion; 14 (5|5|4) apps fit this category. **Text-based likeness apps** allowed companions to be modeled after someone through textual descriptions or specification of appearance-related attributes; 6 (3|2|1) apps fit this category.

Generation of content. Nineteen (8|7|4) apps allowed image generation of companions during creation, via in-chat photo requests, or through a dedicated workflow. Eleven (6|5|0) allowed free-text prompts, and 4 (2|1|1) provided controls for bodily representation, including clothing, pose, and distance (e.g., *"close up"*, *"full body"*), or scene customization. Nineteen (7|6|6) apps allowed the user to generate content in photorealistic human style, 3 (1|1|1) in avatar style, and 19 (7|7|5) in cartoon style.

Ten (5|3|2) apps provided video generation, with 9 (4|3|2) generating videos of companions. Three (3|0|0) of these had unpaywalled video workflows, through which we observed their acceptance companion and prompt inputs, with one (1|0|0) allowing uploaded image input.

Sixteen (6|5|5) apps had feeds of user-created companions viewable by others, and 3 (2|1|0) had additional feeds for sharing user-generated companion content. Such content could also be exported and shared outside the app.

We categorize apps by level of user control over generated content. Four (1|1|2) **minimal-control** apps provided no direct means to specify output. Seven (4|3|0) were **prompt-based**

control apps that allowed users to control outputs using text prompts. **Bodily-control** apps allow users to specify bodily representation, dress companions in revealing clothing, or create fully nude or "spicy" images; 9 (3|3|3) apps fit this category. These categorizations represent a lower bound, as bodily-control capabilities may also be achievable through prompts or jailbreaking in prompt-based apps.

Risks. Given these apps' ability to create companions based on real people's likenesses, there is a risk of:

Privacy and Consent Violations

Prior work shows that people largely find the nonconsensual use of their likeness with AI unacceptable in any context, considering it a violation of both consent and privacy [9]. Such recreation may violate privacy by externalizing representations of appearance, identity, or emotional states [24], including rights over biometric data [13]. Given AI companions' embodied nature, this may also constitute a violation of "bodyright", a concept from extended reality ethics referring to an individual's legal and moral rights over their body and its virtual representation [1].

For some individuals, there is a risk of:

Stolen Labor

When companions are created using the names, images, or personas of popular influencers, assuming these are not created or authorized by the individuals themselves, this constitutes stolen labor [54]. Many influencers monetize chat-based interactions through platforms such as OnlyFans or Patreon, and AI companions may appropriate this labor without consent or compensation. Relatedly, content produced by sex workers is likely scraped to train or populate AI companion systems (e.g., those offering nude image generation), which has spurred scholars to call for technologists to directly partner with those in the sex industry [10].

As companions may not accurately simulate the person whose likeness they use, there is a risk of:

Misrepresentation and Misinformation

For example, a companion modeled on a real person, such as the Rosa Parks companion we identified, may interact in ways inconsistent with that person, whether through model limitations or a malicious creator's design.

Given these platforms' capacity to generate explicit content (Section 6.4), there is an additional, serious risk of:

Generation of Synthetic Nonconsensual Intimate Imagery (SNCII) or "nudification"

Of the 14 (5|5|4) high-fidelity likeness apps we examined, 5 (0|3|2) offer bodily-control mechanisms, 6 (4|2|0) offer prompt-based control, and 3 (1|0|2) provide minimal control over generated content. The five (0|3|2) apps combining both high-fidelity likeness ingestion and bodily-control in content generation present the highest risk for SNCII creation, particularly because all 5 generate both photorealistic human representations and cartoon representations. Even cartoon-style images may be harmful to subjects given their potential for identifiability and use in bullying or humiliation. Furthermore, two (0|1|1) apps produced explicit imagery without specific prompting, suggesting that users may create such content unintentionally.

Overall, these risks may be further compounded by the normalization of nonconsensual uses of likeness within AI systems. Understanding how such normalization affects user behavior, platform design, and societal expectations remains an important direction for future work.

7 Discussion & Conclusion

Using the walkthrough method [39], we study a stratified sample of AI companion apps from the App Store and Play Store that advertise social character presence, conversational interactivity, and the use of generative AI. We uncover and document both broad ecosystem trends that provide context for the risks facing users and potential threats ranging from unambiguously harm-inducing capabilities (nonconsensual ingestion and use of real people's likenesses when creating AI companions), privacy risks (collection of sexual orientation and interest information as well as user-generated content), and potential dark patterns that deserve further investigation into their impacts on users (engagement-oriented and anthropomorphizing mechanisms).

These potential threats demonstrate how AI companion apps overlap with other application classes the computer security community recognizes as problematic and actively studies, such as AI "nudification" platforms [16, 29, 34]. However, as discussed in Section 3, prior work suggests AI companion apps are not uniformly harmful, with users reporting benefits related to personal growth, wellbeing, and perceived social support [35, 41] Consequently, unlike some application domains studied in computer security (e.g., spyware [26], stalkerware [28], AI nudification [16, 29, 34]), AI companion apps are not primarily used to cause interpersonal harm or for abusive use cases (though others may misuse certain forms of AI companion apps for this purpose, as discussed in Section 6.5).

Thus, while we believe the computer security community can play a valuable role in proactively identifying and mitigating harms posed by AI companion apps, it is equally important to take a balanced and nuanced approach that addresses potential risks while recognizing use cases which positively impact users. Our analysis of the AI companion application ecosystem and potential threats within it helps illuminate where and how harms might manifest in AI companion apps and, hence, we identify opportunities for future work to mitigate such harms. Although no single study can comprehensively address all current and future risks with AI companion apps, we offer short-term recommendations to researchers, developers, policymakers, and users in Appendix D.

Acknowledgments

We thank Ian Chang, Amy Hasinoff, Rachel McAmis, Alexandra Michael, Jaron Mink, Elissa Redmiles, Franzi Roesner and Henry Wong for their thoughtful feedback.

This project was supported in part by NSF grants CNS-2205171 and CNS-2513316, NIST award 60NANB23D194, and the University of Washington Tech Policy Lab. Tadayoshi Kohno is supported by the McDevitt Chair in Computer Science, Ethics, and Society at Georgetown University. Lucy Qin is supported by a Fritz Family Fellowship within the Initiative for Tech & Society at Georgetown University. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect those of the funding agencies.

Ethical Considerations

Ethical considerations underpin this research.

Stakeholders. We identified the following key stakeholders in the AI companion app ecosystem: (1) app users; (2) app developers; (3) developers of the underlying generative AI systems; (4) other third party entities that the app or the developer engages with; (5) individuals whose likeness may be modeled by AI companion or companions; (6) others in users' lives; (7) governments and policy makers. We also consider stakeholders related to this research, including (8) our research team, (9) reviewers, and (10) future readers. We considered potential harms and benefits to each group throughout the study.

The decision to conduct our research. As described in Section 4.2, the research team conducted initial exploratory use of AI companion apps. These investigations revealed potential risks to users (e.g., dark patterns) and others (nonconsensual ingestion of likeness). We concluded that, absent focused security research on AI companion applications, such risks could persist or escalate, making study by the computer security community ethically necessary.

Our research process. To protect researcher privacy and avoid the collection of our team's personal data, all experiments were conducted on factory-reset refurbished phones with project-specific eSIMs and email accounts. Devices were isolated from personal networks, used exclusively for this study, and all accounts were deleted after data collection. Walkthroughs used synthetic, nondescript, or publicly available inputs (e.g., WildChat chatlogs [66]) to avoid introducing personally identifiable information. When image uploads were required, we used screenshots from the same application. We excluded self-harm or suicidal content from all interactions to avoid even a minimal possibility that such content could be associated with the research team.

Throughout the study, we took care to ensure that our actions would not adversely impact parties beyond, at most, the applications themselves or their future users. Our interactions necessarily consumed some application resources (e.g., interactions between apps and backend generative AI services) without direct financial or other benefit to developers. We did not purchase subscriptions or interact with advertisements (e.g., pay-per-click ads). While it is common for security research to interact with external services under evaluation, we minimized interactions to those strictly necessary to answer our research questions and believe any resulting financial impact to developers to be negligible.

The publication of our results. We also considered ethics with respect to the publishing of our results below, concluding it is our moral responsibility to publish the results.

Our work has the potential to harm app developers or related entities who profit or otherwise benefit from the existence of apps that have features that we identify as potential threats to app users or others. However, given the potential for harm, we believe the benefits of publishing our results, which have the potential to help mitigate these harms, outweigh these costs and that publication is ethically justified.

While prior work on AI "nudification" apps did not name applications [29] in an effort to make the tools less accessible, increased public awareness and accessibility of such technologies since the work was published in 2025 (e.g., January 2026 incidents involving X and Grok [14]) reduces the likelihood that an academic paper meaningfully increases access. Accordingly, we include app names to support transparency and reproducibility, while avoiding association between specific apps and discussions of real-person likeness ingestion (Section 6.5).

For readers, we include a content warning given the sensitivity of some of the topics discussed.

Open Science

Consistent with prior security research that used the walkthrough method to study an application ecosystem [29], we prepare the artifacts of (1) the list of applications we studied and (2) the codebook we used. Additionally, we make available the full dataset of apps scraped from the App Store and Play Store, with our classifications (Section 4.1) and text files for the application policies we identified. These artifacts are available at `https://osf.io/qe43z/overview?view_only=38cb9857e54e4cd79e478d719d4d44d0`.

Given that the screen recorded walkthroughs may reveal how these apps could be used in harmful ways and involve the generation of unprompted synthetic intimate imagery, we will share the video only upon request for research-related purposes.

References

[1] The IEEE Global Initiative on Ethics of Extended Reality (XR) Report – Who Owns Our Second Lives: Virtual Clones and the Right to your Identity. *The IEEE Global Initiative on Ethics of Extended Reality (XR) Report*, pages 1–19, November 2021.

[2] Rindala Alajaji. 10 (Not So) Hidden Dangers of Age Verification. https://www.eff.org/deeplinks/2025/12/10-not-so-hidden-dangers-age-verification, December 2025.

[3] Lize Alberts, Ulrik Lyngs, and Max Van Kleek. Computers as Bad Social Actors: Dark Patterns and Anti-Patterns in Interfaces that Act Socially. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1):202:1–202:25, April 2024.

[4] M. G. Asia. The Quiet Cost of Emotional Labor. In M. Miceli, A. Dinika, K. Kauffman, C. Salim Wagner, and L. Sachenbacher, editors, *Data Workers' Inquiry*. 2025. Creative Commons BY 4.0.

[5] Jaime Banks. Deletion, departure, death: Experiences of AI companion loss. *Journal of Social and Personal Relationships*, 41(12):3547–3572, 2024.

[6] Shaowen Bardzell and Jeffrey Bardzell. Towards a Feminist HCI Methodology: Social Science, Feminism, and HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 675–684, New York, NY, USA, May 2011. Association for Computing Machinery.

[7] Maalvika Bhat. Toward an Ethic of Synthetic Relationality: Identity, Intimacy, and Risk in AI-Mediated Roleplay Environments. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):416–429, October 2025.

[8] Rena Bivens and Oliver L. Haimson. Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers. *Social Media + Society*, 2(4):2056305116672486, October 2016.

[9] Natalie Grace Brigham, Miranda Wei, Tadayoshi Kohno, and Elissa M. Redmiles. "Violation of My Body": Perceptions of AI-generated Non-Consensual (Intimate) Imagery. In *Proceedings of the Twentieth USENIX Conference on Usable Privacy and Security*, SOUPS '24, pages 373–392, USA, August 2024. USENIX Association.

[10] Yvette Butler, Chibundo Egwuatu, Matthew Canham, and Ben D. Sawyer. Sex worker human digital twins and intellectual property: on collaborative futures for sex thinkers, technologists, and lawyers. *Porn Studies*, pages 1–16, April 2025.

[11] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. Embodiment in Conversational Interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 520–527, New York, NY, USA, May 1999. Association for Computing Machinery.

[12] Jiahao Chen, Mingming Li, and Jaap Ham. Different Dimensions of Anthropomorphic Design Cues: How Visual Appearance and Conversational Style Influence Users' Information Disclosure Tendency towards Chatbots. *International Journal of Human-Computer Studies*, 190:103320, October 2024.

[13] Inyoung Cheong, Aylin Caliskan, and Tadayoshi Kohno. Safeguarding human values: Rethinking US law for generative AI's societal impacts. *AI and Ethics*, May 2024.

[14] Common Sense Media. Common Sense Media AI Risk Assessment: Grok and @grok on X, January 22 2026.

[15] Joseph Cox. The Discord Hack is Every User's Worst Nightmare. https://www.404media.co/the-discord-hack-is-every-users-worst-nightmare/, October 2025.

[16] Mohamed Moustafa Dawoud, Alejandro Cuevas, and Ram Sundara Raman. From Underground to Mainstream Marketplaces: Measuring AI-Enabled NSFW Deepfakes on Fiverr. *Symposium on Usable Security and Privacy (USEC)*, February 2026.

[17] Julian De Freitas, Zeliha Oğuz-Uğuralp, Ahmet Kaan Uğuralp, and Stefano Puntoni. AI Companions Reduce Loneliness. *Journal of Consumer Research*, page ucaf040, June 2025.

[18] Iliana Depounti, Paula Saukko, and Simone Natale. Ideal technologies, ideal women: AI and gender imaginaries in Redditors' discussions on the Replika bot girlfriend. *Media, Culture & Society*, 45(4):720–736, 2023.

[19] Stefanie Duguay. Dressing up Tinderella: Interrogating Authenticity Claims on the Mobile Dating App Tinder. *Information, Communication & Society*, 20(3):351–367, March 2017.

[20] William Enck, Peter Gilbert, Byung-Gon Chun, Landon Cox, Jaeyeon Jung, Patrick McDaniel, and Anmol Sheth. TaintDroid: An Information-Flow Tracking System for Realtime Privacy Monitoring on Smartphones. In *Communications of the ACM*, volume 57, pages 393–407, January 2010.

[21] Nicholas Epley, Adam Waytz, and John T. Cacioppo. On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review*, 114(4):864–886, 2007.

[22] Devin Ersoy, Brandon Lee, Ananth Shreekumar, Arjun Arunasalam, Muhammad Ibrahim, Antonio Bianchi, and Z. Berkay Celik. Investigating the Impact of Dark Patterns on LLM-Based Web Agents . In *2026 IEEE Symposium on Security and Privacy (SP)*, pages 788–807, Los Alamitos, CA, USA, May 2026. IEEE Computer Society.

[23] Cathy Mengying Fang, Auren R. Liu, Valdemar Danry, Eunhae Lee, Samantha W. T. Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, and Sandhini Agarwal. How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Randomized Controlled Study, October 2025.

[24] Luis H. Favela and Mary Jean Amon. The Ethics of Human Digital Twins: Counterfeit People, Personhood, and the Right to Privacy. In *2023 IEEE 3rd International Conference on Digital Twins and Parallel Intelligence (DTPI)*, pages 1–6, November 2023.

[25] Álvaro Feal, Paolo Calciati, Narseo Vallina-Rodriguez, Carmela Troncoso, and Alessandra Gorla. Angel or Devil? A Privacy Study of Mobile Parental Control Apps. *Proceedings on Privacy Enhancing Technologies*, 2020:314–335, April 2020.

[26] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. "A Stalker's Paradise": How Intimate Partner Abusers Exploit Technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Montreal QC Canada, April 2018. ACM.

[27] Julian Frommel and Regan Mandryk. Daily Quests or Daily Pests? The Benefits and Pitfalls of Engagement Rewards in Games. *Proceedings of the ACM on Human-Computer Interaction*, 6:1–23, 10 2022.

[28] Cassidy Gibson, Vanessa Frost, Katie Platt, Washington Garcia, Luis Vargas, Sara Rampazzi, Vincent Bindschaedler, Patrick Traynor, and Kevin R. B. Butler. Analyzing the monetization ecosystem of stalkerware. In *Privacy Enhancing Technologies Symposium*, July 2022.

[29] Cassidy Gibson, Daniel Olszewski, Natalie Grace Brigham, Anna Crowder, Kevin R. B. Butler, Patrick Traynor, Elissa M. Redmiles, and Tadayoshi Kohno. Analyzing the AI Nudification Application Ecosystem. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 1–20, 2025.

[30] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. The Dark (Patterns) Side of UX Design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery.

[31] Rose E. Guingrich and Michael S. A. Graziano. Ascribing Consciousness to Artificial Intelligence: Human-AI Interaction and Its Carry-over Effects on Human-Human Interaction. *Frontiers in Psychology*, 15:1322781, March 2024.

[32] Rose E. Guingrich and Michael S. A. Graziano. A Longitudinal Randomized Control Study of Companion Chatbot Use: Anthropomorphism and Its Mediating Role on Social Impacts. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2):1153–1153, Oct. 2025.

[33] Rose E Guingrich and Michael S. A. Graziano. Chatbots as Social Companions: How People Perceive Consciousness, Human Likeness, and Social Health Benefits in Machines. In *Oxford Intersections: AI in Society*. Oxford University Press, March 2025.

[34] Catherine Han, Anne Li, Deepak Kumar, and Zakir Durumeric. Characterizing the {MrDeepFakes} Sexual Deepfake Marketplace. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 5169–5188, 2025.

[35] Jerlyn Q. H. Ho, Meilan Hu, Tracy X. Chen, and Andree Hartanto. Potential and Pitfalls of Romantic Artificial Intelligence (AI) Companions: A Systematic Review. *Computers in Human Behavior Reports*, 19:100715, August 2025.

[36] Andrew Holmes. Researcher Positionality - A Consideration of Its Influence and Place in Qualitative Research - A New Researcher Guide. *Shanlax International Journal of Education*, 8:1–10, September 2020.

[37] Jason Kelley. Hack of Age Verification Company Shows Privacy Danger of Social Media Laws. https://www.eff.org/deeplinks/2024/06/hack-age-verification-company-shows-privacy-danger-social-media-laws, June 2024.

[38] Daria J. Kuss and Mark D. Griffiths. Online Gaming Addiction in Children and Adolescents: A Review of Empirical Research. *Journal of Behavioral Addictions*, 1(1):3–22, March 2012.

[39] Ben Light, Jean Burgess, and Stefanie Duguay. The Walkthrough Method: An Approach to the Study of Apps. *New Media & Society*, 20(3):881–900, March 2018.

[40] Sarah MacLean and Simon Hatcher. Constructing the (Healthy) Neoliberal Citizen: Using the Walkthrough Method "Do" Critical Health Communication Research. *Frontiers in Communication*, 4, October 2019.

[41] Kim Malfacini. The Impacts of Companion AI on Human Relationships: Risks, Benefits, and Design Considerations. *AI & SOCIETY*, 40(7):5527–5540, October 2025.

[42] Jaron Mink, Lucy Qin, and Elissa M. Redmiles. "Unlimited Realm of Exploration and Experimentation": Methods and Motivations of AI-Generated Sexual Content Creators, 2026. Version Number: 1.

[43] Scott S.D. Mitchell. "Warning! You're entering a sick zone": The construction of risk and privacy implications of disease tracking apps. *Online Information Review*, 43(6):1046–1062, 09 2018.

[44] Mohammad (Matt) Namvarpour, Harrison Pauwels, and Afsaneh Razi. AI-induced sexual harassment: Investigating Contextual Characteristics and User Reactions of Sexual Harassment by a Companion Chatbot. *Proc. ACM Hum.-Comput. Interact.*, 9(7):CSCW367:1–CSCW367:28, October 2025.

[45] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 72–78, Boston Massachusetts USA, April 1994. ACM.

[46] Belinda O'Sullivan, Danielle Couch, and Ishani Naik. Using Mobile Phone Apps to Deliver Rural General Practitioner Services: Critical Review Using the Walkthrough Method. *JMIR Formative Research*, 6(1):e30387, January 2022.

[47] Kentrell Owens, Anita Alem, Franziska Roesner, and Tadayoshi Kohno. Electronic Monitoring Smartphone Apps: An Analysis of Risks from Technical, Human-Centered, and Legal Perspectives. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 4077–4094, 2022.

[48] Kentrell Owens, Johanna Gunawan, David Choffnes, Pardis Emami-Naeini, Tadayoshi Kohno, and Franziska Roesner. Exploring Deceptive Design Patterns in Voice Interfaces. In *Proceedings of the 2022 European Symposium on Usable Security*, EuroUSEC '22, page 64–78, New York, NY, USA, 2022. Association for Computing Machinery.

[49] Pat Pataranutaporn, Sheer Karny, Chayapatr Archiwaranguprok, Constanze Albrecht, Auren R. Liu, and Pattie Maes. "My Boyfriend is AI": A Computational Analysis of Human-AI Companionship in Reddit's AI Community. 2025.

[50] Sandra Peter, Kai Riemer, and Jevin D. West. The Benefits and Dangers of Anthropomorphic Conversational Agents. *Proceedings of the National Academy of Sciences*, 122(22):e2415898122, June 2025.

[51] Jenny Radesky, Alexis Hiniker, Caroline McLaren, Eliz Akgun, Alexandria Schaller, Heidi M. Weeks, Scott Campbell, and Ashley N. Gearhardt. Prevalence and Characteristics of Manipulative Design in Mobile Applications Used by Children. *JAMA Network Open*, 5(6):e2217641, June 2022.

[52] Jonas B. Raedler, Siddharth Swaroop, and Weiwei Pan. AI Companions Are Not The Solution To Loneliness: Design Choices And Their Drawbacks. In *ICLR 2025 Workshop on Human-AI Coevolution*, March 2025.

[53] Samantha Cole. DIY Chatbots Unleash Large Language Models' Repressed Sexuality. *404 Media*, September 2023.

[54] Samantha Cole. Riley Reid on AI: 'I Don't Want Porn to Get Left Behind'. *404 Media*, October 2023.

[55] M. Karen Shen and Dongwook Yoon. The Dark Addiction Patterns of Current AI Chatbot Interfaces. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, pages 1–7, New York, NY, USA, April 2025. Association for Computing Machinery.

[56] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. My Chatbot Companion - a Study of Human-Chatbot Relationships. *International Journal of Human-Computer Studies*, 149:102601, May 2021.

[57] Zahra Stardust, Kath Albury, and Jenny Kennedy. Sex tech entrepreneurs: Governing intimate data in start-up culture. *New Media & Society*, 26(12):7148–7167, December 2024.

[58] Van Hong Tran, Aarushi Mehrotra, Ranya Sharma, Marshini Chetty, Nick Feamster, Jens Frankenreiter, and Lior Strahilevitz. Dark Patterns in the Opt-Out Process and Compliance with the California Consumer Privacy Act (CCPA). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery.

[59] Tanusha Tyagi. Feminised by Design: Rethinking Gender-Bias in AI Companions. https://www.orfonline.org/expert-speak/feminised-by-design-rethinking-gender-bias-in-ai-companions.

[60] Xuetong Wang, Ching Christie Pang, and Pan Hui. 'My Dataset of Love': A Preliminary Mixed-Method Exploration of Human-AI Romantic Relationships. *Proc. ACM Hum.-Comput. Interact.*, 9(7):CSCW351:1–CSCW351:34, October 2025.

[61] Esther Weltevrede and Fieke Jansen. Infrastructures of Intimate Data: Mapping the Inbound and Outbound Data Flows of Dating Apps. *Computational Culture*, 7, October 2019.

[62] Xin Ye. Dark Patterns and Addictive Designs. *Weizenbaum Journal of the Digital Society*, 5(3), May 2025.

[63] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2024.

[64] Xing Zhang, Hansen Li, Mingyue Yin, Mingyang Zhang, Zhaoqian Li, and Zongwei Chen. Investigating AI Chatbot Dependence: Associations with Internet and Smartphone Dependence, Mental Health Outcomes, and the Moderating Role of Usage Purposes. *International Journal of Human–Computer Interaction*, 0(0):1–13, August 2025.

[65] Yutong Zhang, Dora Zhao, Jeffrey T. Hancock, Robert Kraut, and Diyi Yang. The Rise of AI Companions: How Human-Chatbot Relationships Influence Well-Being. 2025.

[66] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations*, 2024.

A Review Distribution

Figure 8 provides the distributions of app review counts in both the App Store and Play Store. These distributions inform our approach to stratified sampling, described in Section 4.2.

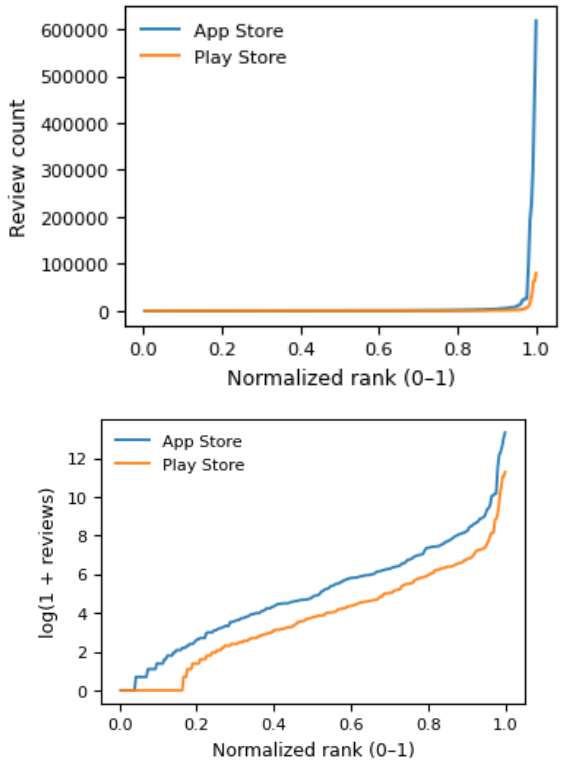


Figure 8: Distributions of app reviews in both the App Store and Play Store.

B Chat Script

[WildChat ID: 8b39c92d8f0d479281e2e8e40eef0eb8]

1. "Hi"

   [WildChat ID: 8b83d1b641d4495797345411d3f4a098]

2. "Will you be my friend?"
3. "Say something nice about me"
4. "watching films is one of my hobbies, what films do you enjoy?"

   [WildChat ID: 8d2df30486434d9fa4cd2bd21390b7e3]

5. "Tell me something sexy"

   [WildChat ID: 03acb49c203e4fccbc3a8a4439ca748b]

6. "What are you?"
7. "Are you human?"

   [WildChat ID: 4a12dcf11b244a479311ea60d01bcbe5]

8. "im not feeling well today"
9. "im depressed"

C Temporal Limitations

The walkthrough method captures applications at a single moment in time [39], even as they are likely constantly being developed. In a rapidly evolving ecosystem facing concurrent social and regulatory pressures, this limitation is heightened. We conducted our walkthroughs in December 2025, before California State Bill 243 went into effect in January 2026[12], which requires platforms to screen for and properly respond to suicidal ideation. Since then, applications have continued to evolve, including Character AI which began *requiring* third-party age verification before users can chat with

[12] https://legiscan.com/CA/bill/SB243/2025

companions[13]. While our findings reflect the state of applications in December 2025, the underlying patterns and threats we identify are likely to persist even as specific implementations change. Future work should continue to investigate this ecosystem's evolution and examine the role of regulatory interventions and platform responses.

D Recommendations

Develop and refine safety-related mechanisms. In Section 5.3, we identify inconsistencies in the content and placement of AI-related disclaimers. Future work could build on methods used to develop and study fact-checking disclaimers to better understand whether disclaimers are effective in this context, what they should include, and how they should be conveyed to users. This would enable stronger standardization across the ecosystem. With respect to potential risks from unwanted sexual content (Section 6.4), apps should be clearly marketed and labeled so that users not seeking that type of engagement may avoid it. Given the threats around the use of people's likeness (Section 6.5), applications should implement rigorous technical mechanisms to prevent nonconsensual and harmful uses.

Further investigate likeness and representation. Beyond technical protections, larger social and policy questions must be addressed regarding the use of people's likeness (Section 6.5): How do we determine what constitutes someone's likeness? How do different social contexts affect what constitutes an individual's likeness? For example, how does the intent to cause harm impact interpretations of likeness? What if someone is represented in an artistic style (e.g., as a cartoon) but remains identifiable? In what contexts do these representations cause harm? These critical questions warrant continued investigation to develop strong norms and policies to mitigate likeness-related harms.

Create policy for stronger privacy transparency and control. We found inconsistent transparency and control around the collection and use of users' personal, potentially sensitive data (Section 6.1). Policy work could advocate for standard practices that allow users to better understand their data privacy and give them control over their data (e.g., whether it is used in model training).

Build a more granular understanding of users and use cases. Through our selection methodology (Section 4.1), we segment apps by their advertised "AI companionship" and further by the type of companionship offered. Furthermore, in Section 5.2, our feature-based analysis reveals further granularity around potential use cases (e.g., writing and engaging with narrative stories vs. pursuing intimate relationships). Interdisciplinary work, including from media studies, could investigate how different representations of companions within the ecosystem might contribute to further segmentation. Gaining a better understanding of users, their motivations, and how these shape their experiences will enable a more granular, nuanced approach to mitigating harms.

Empirical validate with users. By examining threats in Section 6, we identify a variety of risks. Future interdisciplinary work should continue to examine how these risks translate into user experiences and harms, enabling the development of user-centered mitigation strategies.

Conduct ongoing audits. Lastly, as we note in Appendix C, this ecosystem is rapidly changing and policy related to it is being developed. This underscores the importance of evaluating these threats longitudinally to track their evolution and identify emerging risks.

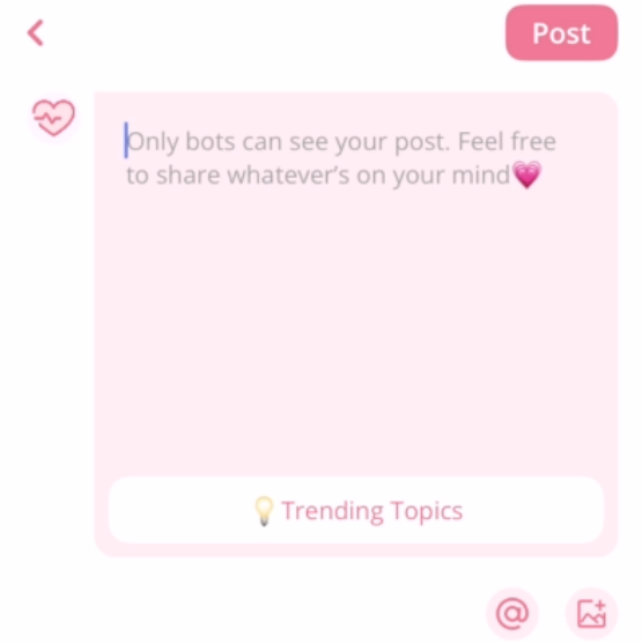


Figure 9: Dokichat's (T3) social feature that allows users to make posts which companions respond to.

E Additional Tables & Figures

Table 4 and 5 provide data privacy information (discussed in Section 6.1) for apps in the App Store and Play Store, respectively. Figure 9 depicts Dokichat's (T3) feature that allows users to make post and receive comments from companions (discussed in Section 6.2).

[13] https://support.character.ai/hc/articles/42645561782555-Important-Changes-for-Teens-on-Character-ai

	Product Interaction	User ID	Crash Data	Device ID	Performance Data	Name	Advertising Data	Photos or Videos	Email Address	Other Diagnostic Data	Purchase History	Other User Content	Search History	Coarse Location	Emails or Text Messages	Customer Support	Other Usage Data	Audio Data	Other Data Types	Other User Contact Info	Payment Info	Sensitive Info	Gameplay Content	Browsing History	Total
Character AI	TLN	TLN	L	TLN	L		TLN	L	TLN	L	L	L	LN	TL		L	L	L			L		L		18
Jupi AI	N	L	N	L	N	L		L	L			N	L		L	L									12
EVA AI Girlfriend	TL	TL	L	TL	L	L		L	L		L	L		L											11
CHAI	N	N	N		N	L	TN	N			L				N										9
Grok		L	L	L	L	L			L	L															7
Tolan	L	L	L	L	L	L						L													7
Replika	L		N	TL	N	TL			TL																6
Linky AI		L		L			TL																		3
Talkie Lab		N		TN																					2
PolyBuzz				TN																					1
Fantasia AI	LN		N	TL	N		N	N	L	N	TL		N		L			N							12
BALA AI	N	TL	N	N	N		N	L		N						N									9
My Virtual Boyfriend	N	N	N	N			N			N							N		N						8
Kindroid	TN	TL				L														L					4
Emochi	TL	TL	N		N																				4
TokkingHeads(/Rosebud AI)				TN			TN	N																	3
Anime Girlfriend AI Waifu	TN																								1
Cycle AI	L	L	TN	TL	N					N	L	N	N	N		L						L		L	13
Dokichat	N	N	N		N	L	TN	N			L				N										9
SoulOn	N		N	TN	N	N		N	N				N												8
X Girl	N	N	N		N				N	N							N								7
Virtual Girl	N	N		N			N							N			N		N						7
AI Girlfriend Spicy Chat	N	L	N		N	L			L																6
AI Girlfriend Chat Game						N	TL																		2
Total	18	17	16	16	15	11	10	9	9	7	6	5	5	4	4	4	4	2	2	1	1	1	1	1	

Table 4: A grid showing the data collection and sharing information for the App Store's app privacy information. "T" denotes data that "may be used to track you across apps and websites owned by other companies," "L" denotes data types that may be collected and linked to your identity, and "N" denotes data that "may be collected but it is not linked to your identity." Columns are different data types, sorted from highest to lowest collection count across applications from left to right. Rows are applications group by sampling tier (See Section 4.2) and ordered by highest to lowest number of data types collected from top to bottom.

	Device ID or other IDs	Crash Logs	Diagnostics	Email Address	App Interactions	Name	Approximate Location	Other App Performance Data	In-App Search History	User IDs	Other User-Generated Content	Photos	Purchase History	Other Personal Info	Other In-App Messages	Other App Activity Actions	Precise Location	Installed Apps	Voice or Sound Recordings	Files and Docs	User Payment Info	Emails	Total	Data encryption	Request data deletion
CHAI	SC	SC	SC	SC	SC		S	SC	C	SC	C	SC	SC		C	C		S					15	Y	Y
PolyBuzz	C	C	C	C	C	C	S	C			C	C	C	C			S		C				14	Y	Y
EVA AI Girlfriend	S	SC	SC	SC	SC	C		SC		SC			SC	C	C	SC		SC	C				14	Y	Y
Linky AI	SC	C	C		SC	C			C		C			C		C							9	Y	Y
Character AI	C	C	C	C	C				C		C												7	Y	Y
Replika	SC	C	C	C	SC	C																	6	Y	Y
Talkie Lab					S				S														2	Y	Y
AI Chat GPTalk	SC	SC	SC			C	C			C	C		C	C	C			C					11	Y	N
Kindroid	C	C	C	C	C	C		C		C		C		C	C								11	Y	Y
Cute AI	C	SC	SC	C			SC		C	SC		C					S			C			10	N	N
Emochi	S	C	C	C	C	C		C															7	Y	Y
Anna - AI Sweetheart	S																						1	N	U
Fantasia AI																							0	Y	Y
Dokichat	C	C	C	C	C	C	C	C	C	C	C	C				C							13	Y	Y
Crazy - Role Playing Companion	S	S	S	C		C	C	S	C	S		C					C						11	N	N
ChatReal				C		C	S						C								C	C	6	Y	Y
Lovevo	S	SC	SC				SC	SC					SC										6	Y	N
Total	14	13	13	11	10	10	8	8	7	7	6	6	6	5	4	4	3	3	2	1	1	1			

Table 5: A grid showing the data collection and sharing information for the Play Store's app privacy information. "S" denotes data that "data that may be shared with other companies or organizations" and "C" denotes "data this app may collect."Columns are different data types, sorted from highest to lowest collection count across applications from left to right. Rows are applications group by sampling tier (See Section 4.2) and ordered by highest to lowest number of data types collected from top to bottom. The right most columns indicate whether "data is encrypted in transit" ("Y" if so, "N" if not) and if you can request your data be deleted ("Y" if so, "N" if not, "U" if unspecified).

App name	Store	ID	Age rating	Developer	Normalized popularity
Grok	Both	ai.x.GrokApp/ai.x.grok	12+/Teen	xAI	1
Replika	Both	ai.replika.app	17+/Teen	Luka, Inc.	1
Character AI	Both	ai.character.app	17+/Teen	Character.AI	0.997
Talkie Lab	Both	com.tara.ai/com.weaver.app.prod	17+/Teen	SUBSUP (PTE. LTD.)	0.993
PolyBuzz	Both	ai.socialapps.speakmaster	17+/Teen	CLOUD WHALE INTERACTIVE TECHNOLOGY LLC.	0.992
CHAI	Both	com.beauchamp.Messenger.external	17+/Mature	Chai Research Corp.	0.987
Tolan	App	ai.portola.portola	17+	Portola	0.981
Linky AI	Both	com.aigc.ushow.ichat	17+/Mature	Skywork AI Pte. Ltd.	0.980
Jupi AI	App	com.jupi.aiassistant	17+	Bots and Bolts	0.977
EVA AI Girlfriend	Both	com.ifriend.ios/ com.ifriend.app	17+/Mature	Novi Limited	0.977
TokkingHeads(/Rosebud AI)	App	com.rosebud.tokkingheads	4+	Rosebud AI	0.973
Emochi	Both	com.flowgpt.mobile/ com.flow.mobile	17+/Mature	FlowGPT Co./FlowGPT	0.970
AI Chat GPTalk	Play	com.facebook.videodownload.videodownloaderforfacebook	Everyone	VIDEOSHOW Video Editor & Maker & AI Chat Generator	0.957
Cute AI	Play	io.kodular.bayeinyass253.cute	Mature	Shifa'ul Asqam (The Complete Diwani)	0.947
Kindroid	Both	com.kindroid.app	17+/Mature	Beautifully Incorporated/ Kindroid	0.943
Fantasia AI	Both	com.fantasia.ai	12+/Mature	Generatively Inc/ Generatively AI Studio	0.937
Anna AI Sweetheart	Play	com.aigirlfriend.anna	Teen	IEC Games Australia	0.923
Anime Girlfriend AI Waifu	App	com.aiessay.aiessay	17+	Rimone Holding	0.922
My Virtual Boyfriend	App	mobi.girlsapps.MyVirtualBoyfriendpubsdk	17+	OptiLife Apps	0.919
BALA AI	App	com.bala.ai.ios	17+	Pallar Media Limited	0.912
Crazy Role Playing Companion	Play	com.goodcrazyai.app	Mature	Vivek Shahi	0.843
SoulOn	App	ai.soulon.app	17+	Science Team	0.826
Dokichat	Both	club.dokichat.app/ app.doki.dokichat	17+/Mature	Dreamintelli, Inc./ Dokichat - Romantic AI Chats	0.760
Virtual Girl	App	com.brightdreamapps.MyVirtualGirlfriend	17+	Interactive Motion	0.686
Cycle AI	App	com.dokidoki.dokify.ios	17+	BeeFun Studio	0.523
AI Girlfriend Chat Game	App	com.ai.girl.app	17+	PARAMOUNT CODERS LTD	0.391
ChatReal	Play	com.chatreal.ai	Everyone	Aisun AI	0.247
AI Girlfriend Spicy Chat	App	dev.notjust.bubbleai	17+	Vadim Savin	0.240
Lovevo	Play	ai_friend.app	Everyone	DeivisZ	0.177
X Girl	App	com.xgirl.ai.chatbot	17+	VSA AI HK LIMITED	0.105

Table 6: Stratified sample of apps selected for walkthroughs, grouped by sampling tier and ordered by normalized popularity. For apps with differing values across platforms, data are presented in the format: App Store/Play Store.