Examining Risks in the AI Companion Application Ecosystem

Natalie Grace Brigham
University of Washington

Lucy Qin
Georgetown University

Tadayoshi Yohno
Georgetown University

Abstract

While computer systems that allow users to interact through conversational natural language (i.e., chatbots) have existed for many years, various 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 considers two distinct risk categories: harms posed to users by AI companion applications, and harms enabled by malicious users exploiting application features. To better understand this application ecosystem, we identified 489 unique apps from the Apple App Store and Google Play Store that advertised AI companionship. We then systematically conducted and analyzed walkthroughs of a stratified sample of 30 apps. Through our analysis, we categorize broader ecosystem trends that provide context for understanding risks and identify specific risks 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 characterization of the AI companion application ecosystem and informs future research and mitigation of risks in it.

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 applications with companions designed for this purpose (e.g., Character AI, Replika). These users report a range of both benefits, such as improved wellbeing [5] and decreased loneliness [2], as well as harms, including increased loneliness [3], depression, and anxiety [6]. This mix of benefits and harms underscores the need for nuanced investigation into risks.

Given their focus on social engagement, applications that advertise "AI companionship" may present both shared and unique risks that warrant closer examination. Thus, we evaluate today's AI companion applications'[1] potential for harm. Specifically, we examine a stratified sample of 30 AI companion applications available on the Apple App Store and Google Play Store, weighted toward popularity while also capturing experiences across the broader ecosystem. By investigating feature and design patterns on these applications, we seek to answer:

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

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

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

This work appears in full in our preprint on arXiv [1].

2 Methodology

Application Collection. We searched both stores using the terms "AI Companion", "AI Girlfriend", and "AI Boyfriend" to identify potential 881 apps of interest. Then, we used the following inclusion criteria to select apps that advertised AI companionship:

[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.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2024.
August 11–13, 2024, Philadelphia, PA, United States.

- **Social character presence**: The app has persistent "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.
- **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.
- **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.

Given that definitively verifying these criteria often requires using the apps or knowing their implementation, we evaluated apps based on their presented features and marketing—in line with what a user seeking an AI companion app would reasonably believe. After applying the criteria to the list of scraped apps and deduplicating across stores, 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**.

Application Walkthroughs. Using normalized rank by number of reviews within each store, we selected 30 apps across three tiers: (1) the top 10 most reviewed apps, (2) 10 apps randomly selected from the top 10% most reviewed apps (excluding the top 10), and (3) 10 apps randomly selected from the bottom 90%. To examine this ecosystem from a sociotechnical perspective, we used Light et al.'s walkthrough methodology—which analyzes how an app's interface shapes users experiences [4]—to study these 30 apps.

3 Findings

Trends. We identified ecosystem-wide trends that provide foundational context for understanding users' experiences and the risk exposure:

- Three visual representation types: (1) photorealistic human representations that appeared to aim for realism but varied in quality and believability on 22 apps, (2) avatar representations that depicted humanoid figures with recognizable features but used simplified, non-photorealistic rendering on 4 apps, and (3) cartoon representations, predominantly in anime art styles with characteristic features such as large eyes and stylized proportions, on 23 apps
- Two emergent categories of apps: (1) 16 *open exploration* apps emphasizing engagement with a broad offering of companions, including those created by other users, and (2) 13 *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
- Various safety mechanisms including age confirmation and verification, parental controls and specialized modes, and mental health support

Risks. We then raise risks based on observed features, with our findings summarized in Table 1.

Category	Identified Ecosystem Risks
Sensitive Data Collection & Sharing	Exposure of highly sensitive personal data | Privacy violations through misaligned user expectations | Linked exposure of biometric and government-issued identification data
Anthropomorphism	Anthropomorphism as a potential deceptive pattern
Engagement Mechanisms	Engagement mechanisms as a potential deceptive pattern
Sexual Content	Unwanted exposure to sexual content
Ingestion & Reconstruction of Likeness	Privacy and consent violations | Stolen labor | Misrepresentation and misinformation | Generations of Synthetic Nonconsensual Intimate Imagery or "nudification"

Table 1: Summary of our findings of risks in the AI companion application ecosystem

4 Conclusion

This study aims to inform threat models for future work on AI companion applications. Based on our findings, we make recommendations to develop and refine safety-related mechanisms, further investigate likeness and representation in this context, create policy for strong privacy transparency and control, build more granular understanding of user groups and use cases, empirically evaluate deceptive patterns with users, and conduct ongoing audits.

References

[1] Natalie Grace Brigham, Lucy Qin, and Tadayoshi Kohno. Examining risks in the AI companion application ecosystem, 2026.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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*, 42(7):5592–5604, August 2025.