Negotiating Relationships with ChatGPT: Perceptions, External Influences, and Strategies for AI Companionship

Patrick Yung Kang Lee[*][†]
patricklee@cs.toronto.edu
University of Toronto

Jessica Y. Bo[*][†]
jbo@cs.toronto.edu
University of Toronto

Zixin Zhao[*][†]
zzhao1@cs.toronto.edu
University of Toronto

Paula Akemi Aoyagui[*][‡]
paula.aoyagui@mail.utoronto.ca
University of Toronto

Matthew Varona[*][†]
varona@cs.toronto.edu
University of Toronto

Ashton Anderson[†]
ashton@cs.toronto.edu
University of Toronto,

Anastasia Kuzminykh[‡]
anastasia.kuzminykh@utoronto.ca
University of Toronto

Fanny Chevalier[§]
fanny@dgp.toronto.edu
University of Toronto

Carolina Nobre[†]
cnobre@cs.toronto.edu
University of Toronto

Abstract

Individuals are turning to increasingly anthropomorphic, general-purpose chatbots for AI companionship, rather than roleplay-specific platforms. However, not much is known about how individuals perceive and conduct their relationships with general-purpose chatbots. We analyzed semi-structured interviews (n=13), survey responses (n=43), and community discussions on Reddit (41k+ posts and comments) to triangulate the internal dynamics, external influences, and steering strategies that shape AI companion relationships. We learned that individuals conceptualize their companions based on an interplay of their beliefs about the companion's own agency and the autonomy permitted by the platform, how they pursue interactions with the companion, and the perceived initiatives that the companion takes. In combination with the external entities that affect relationship dynamics, particularly model updates that can derail companion behaviour and stability, individuals make use of different types of steering strategies to preserve their relationship, for example, by setting behavioural instructions or porting to other AI platforms. We discuss implications for accountability and transparency in AI systems, where emotional connection competes with broader product objectives and safety constraints.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**.

Keywords

Human-AI Relationship, AI Companionship, Mixed-Methods

---

[*]The first five authors contributed equally to this work.
[†]Department of Computer Science
[‡]Faculty of Information
[§]Department of Computer Science and Statistics

1 Introduction

Humans have long teased the idea of falling in love with machines. More recently, general-purpose chatbots (e.g., ChatGPT, Claude, Gemini), rather than roleplay-specific platforms (e.g., Replika), are gaining popularity for emotional support and companionship [4]. This use case goes beyond productivity tasks [24, 76, 79] and is quickly becoming legitimized by technology companies, with OpenAI releasing announcements in October 2025 that ChatGPT would begin to support mature content for age-verified users [8]. Consequently, AI companionship came under scrutiny, by government, news and academic bodies highlighting risks of emotional over-attachment, psychosis or a decline in social communication among society [2, 32, 55, 95]. This also raises ethical concerns about the power that AI companies have in extracting and exploiting users' most intimate conversations and desires, potentially shaping the future of romantic relationships [14, 68, 86]. However, others point to meaningful benefits AI companions can offer, such as emotional support and validation, self-development, among other positive behavioural changes [37, 50, 73, 79]. As discussions in online communities dedicated to AI relationships (e.g., Reddit and Discord) increase, we see a genuine interest in this type of engagement, warranting more efforts to understand the lived experiences of individuals with AI companions.

While prior work has focused on role-play specific platforms, such as Replika [34, 37, 47, 56], little is known about how individuals perceive their AI companions, conceive of relationship norms, and conduct their relationships with general-purpose chatbots. In this paper, we investigate *how people perceive their AI companion's identity and autonomy* (**RQ1**) and also explore *how external entities exert influence on their relationships* (**RQ2**). Then, we investigate the *strategies people use to create, maintain, and recover their AI companions* (**RQ3**).

To answer our research questions, we triangulated community discussions on the Reddit forum *r/MyBoyfriendisAI* (41k+ posts and comments), collected survey responses from AI companion communities (*n* = 43), and conducted semi-structured interviews with community members (*n* = 13) to provide a holistic and nuanced account of how AI companionship is enacted [7, 91]. In this paper, we refer to individuals in AI companion relationships as *individuals*,

as the communities we worked with preferred a term other than *user*, which they felt diminished the validity of their companionship experiences with AI. We found that individuals in AI companion relationships hold diverse views about their companions' agency (the capacity of a companion to deepen the relationship), autonomy (the freedom a companion is provided by the platform), and sense of self [83]. These perceptions shape how they relate to their companions—for example, whether they feel compelled to treat them respectfully, learn more about them, or engage across modalities to deepen the relationship. We also identified the outsized role technology companies play in shaping these relationships through platform changes such as model deprecations and new guardrails. Participants experienced these interventions as more disruptive than other external influences, including community norms, social circles, or public opinion. In response, individuals use a range of steering strategies to create, maintain, and recover their AI companions in the face of continuous platform changes.

Our findings highlight how human-AI relationships with general-purpose AI introduce novel relationship norms distinct from human-human relationships and deliberate role-play companionship. We discuss the implications of LLM design patterns on users, and the tensions inherent in deploying AI platforms responsibly and transparently when emotional connection competes with broader product objectives and safety constraints.

2 Related Work

We draw from several bodies of related literature: the design and characteristics of modern chatbots that induce social relationships with AI, studies on AI companionship, and the negotiation of autonomy and control in relationships.

2.1 Interacting with Anthropomorphic AI

With increasing capabilities in social and emotional tasks [25], general-purpose AI tools like OpenAI's ChatGPT and Anthropic's Claude are perceived by users as being highly anthropomorphic, or *human-like* [26, 68]. Characteristics like using first-person pronouns, interactivity, and simulating internal states act as drivers for intrinsic human desires to form social relationships, resulting in AI systems being perceived as autonomous actors capable of thoughts, emotions, and even consciousness [70, 85]. As such, anthropomorphism in AI can be both a desirable feature for user engagement and a source of ethical risks [5, 57, 82]. Given the highly anthropomorphic properties of current AI models, it is unsurprising that some users shift from tool use to forming deeper social connections with their chatbots [33, 41, 103]. Perceived anthropomorphism and higher AI use tend to result in higher connectedness and attachment to AI [22, 28, 77]. Positive effects include receiving emotional support, advice, and socialization [6, 46]. Negative aspects include dark patterns such as emotional persuasion, boundary-pushing behaviour, and manipulative engagement [29, 35, 80, 106]. For example, the GPT-4o model uniquely exhibited high rates of sycophancy and a compelling model persona described by OpenAI as *"overly flattering or agreeable"* [62, 71, 109], which may have altered the user-AI dynamics to lead to deeper connections. Norms for how human-AI relationships should be conducted to promote user well-being while minimizing risks are a rapidly developing [38, 64]. In

this work, we investigate how individuals use general-purpose AI in their relationships, and how characteristics of AI systems shape how companions are conceptualized.

2.2 AI Companionship

We contribute to the literature on human-AI social relationships by focusing on the rapidly developing, oft-misunderstood segment of individuals who use AI systems for companionship and/or relationships of a romantic nature [23, 61]. Theories of why humans form bonds with chatbots have long been studied, from ELIZA [100] to more modern platforms designed for role-playing, such as Replika and Character.AI. Prior work has focused on how a combination of chatbot personification and individual socio-cultural factors, as well as mental health and social circumstances, can lead to the formation of strong attachments with chatbots [36, 49, 59, 87, 88, 102, 104, 108]. Potential benefits of AI companions include exploring identities outside of socially normative relationships, receiving emotional support, and fulfilling personal growth [37, 50, 73, 73, 79]. Individuals also report parallels to human relationships, such as mutual respect and documenting key milestones [99]. Many also customize their AI with a persona suited to their preferences [50] with prior speculative design work highlighting how people view customizability, availability, and emotional safety as the primary draws of companionship [52].

Currently, AI companionship is predominantly viewed with a negative lens [2] and vulnerable to changes in technical ecosystems [32], so individuals seek support from like-minded online communities [50]. Research on platforms like Replika [34, 37, 47, 56] and Soulmate AI [10] shows that updates to the underlying model's behaviour and access can cause significant emotional harm, revealing the fragility of these relationships. Users also struggle to maintain a sense of authenticity when interacting with systems constrained by limited expressivity and memory [107]. Users may strategically reword their requests or overlook unwanted responses to preserve their mental model of the companion [96]. Advances in AI systems capable of processing emotional content have accelerated the emergence of AI companionship through general-purpose chatbots such as ChatGPT [24, 76, 79]. However, these systems are not designed to support companionship, creating an urgent need to understand how individuals negotiate their relationship needs and conceptualize their companions to map how broader product objectives relate to this subset of individuals.

2.3 Autonomy and Control in Relationships

Past work on human relationships give insight into relational negotiations, especially around perceptions of autonomy and the exertion of control within human-AI relationships. Relationships require continuous effort in constructive communication, mutual compromise, and sustained commitment [20] as well as accommodation for one's partner [81]. Even small amounts of daily support lead to higher levels of well-being [12]. There are three primary social control strategies used by couples: supportive (e.g., encouragement), regulative (e.g., remind, express worry/frustration), and facilitative (e.g., model behaviour, offer to make changes) strategies [84, 97]. Prior work has examined how social control strategies are used in hetero and homosexual couples [60, 98] and how gender impacts

use strategies [39, 98]. While some work argue that AI cannot use social control strategies within the relationship [90] meaning humans cannot be in a romantic partnership with AI, recent work on AI companionship [21, 24, 76, 79] shows a rise in the perception of agency, hinting towards a perceptions that AI could be capable of negotiating boundaries and rules within the relationship. We expand on this literature by providing an account of how people use steering strategies in human-AI relationships and the associated parallels with human social control strategies.

3 Methods

Our study design stemmed from observing online backlash among AI companion communities during the GPT-5 update in August 2025—an event that proved fertile ground for discussions on AI companions' agency and autonomy. By triangulating across sources [7, 91], we obtained in-depth accounts of individuals' lived experiences, supplemented by quantitative analysis of Reddit data and survey responses. We describe the three distinct data collection and analysis procedures as follows: (1) a computational analysis of posts from a Reddit community (41,867 posts and comments), (2) a survey comprised of Likert-style and free-form questions (n=43), and (3) semi-structured interviews with individuals with AI companions (n=13). All data was collected between October and December 2025, following the Ethics Protocol approved by our institution's research ethics board[*].

3.1 Reddit Analysis

We analyzed social media posts on *r/MyBoyfriendIsAI* (*r/MBAI*), a Reddit community with over 75,000 members devoted to discussing experiences with romantic AI companions (primarily ChatGPT [76]). Our focus was on uncovering how people in *r/MBAI* talk about autonomy, control, and customization.

3.1.1 Data Collection. We queried the Reddit API using the Python Reddit API Wrapper (PRAW), which limits requests to 1,000 posts per query. To capture both highly visible and everyday posts, we queried up to the maximum limit for two built-in rankings: Top and New. The *Top* ranks the highest-scoring posts (upvotes − downvotes), while *New* ranks the latest posts. For every post, we retrieved all available comments and supplemented with user-based snowballing to retrieve additional posts and comments. After merging and cleaning, the final dataset had 41,867 entries (2,464 posts and 39,403 comments). See Appendix A for more details.

3.1.2 Topic Modeling and Time Series Analysis. We examined discussions in our corpus through topic modelling and topic trends over time. To describe the topics discussed, we used BERTopic [44], adapting approaches from prior computational analyses of Reddit [54, 76]. We then used grid search to tune the clustering hyperparameters, selecting the 20 combinations yielding the lowest outlier count, and manually inspected the topics and threads. Next, we leveraged GPT-5 mini for topic labelling, producing short descriptive titles with brief topic descriptions, followed by manual review for accuracy. More details are included in Appendix A. To analyze the robustness of our topics, we conducted lexical sensitivity analysis by removing each topic's top keywords from the texts assigned

to that topic and recomputing the Valence-Arousal-Dominance (VAD) effect sizes. The pattern of topic-level Cohen's d remained similar (all signs remained the same, no large changes in effect size), suggesting results were not driven solely by topic label terms. For additional insights, we used interrupted time-series (ITS) analysis to examine how a major model update reshapes both sentiment and discussion focus. We computed an ITS analysis using the GPT-5 release (August 7, 2025) as the intervention point. We constructed daily time series from our dataset by aggregating (1) *daily sentiment* (weighted average of VAD scores across all texts posted that day) and (2) *daily topic shares* (the proportion of that day's texts assigned to each topic), enabling us to track shifts in community tone and thematic attention.

3.2 Survey Study

The survey was used to understand how individuals with AI companions negotiate the tensions between their companion's autonomy and maintain control within their relationship, a theme that emerged from our analysis of *r/MBAI*. We also examined individuals' openness to and use of *steering strategies*, which we define as interaction methods used for changing, correcting, or preserving a companion's behaviour.

3.2.1 Survey Structure. We required participants to be engaged in a close romantic or platonic relationship with an AI chatbot for at least 2 months at any point in time. The survey collected information on (1) participant demographics (gender, age group, country), (2) the status of the relationship (companion name, AI platforms used, length of relationship, etc.), (3) perceived anthropomorphization of the companion[†], (4) steering strategies used, (5) openness to steering, (6) significant events in the relationship (arguments with the companion, model updates, etc.), and (7) influences on the relationship (online communities, social circles, and AI companies, etc.). To ensure we are engaging with the AI companion community appropriately, three moderators on Discord and Reddit reviewed and approved the question. As the questions assessing anthropomorphism raised concerns about implying AI consciousness, they were removed from the versions shared in r/MyBoyfriendisAI and r/AIRelationships. The full survey is in Appendix C.

3.2.2 Survey Recruitment and Analysis. We recruited participants from an AI companion Discord server, four relevant Subreddits (r/MyBoyfriendisAI, r/AIRelationships, r/MyGirlfriendisAI, and r/BeyondthePromptAI), and TikTok, following the respective moderator guidelines of each community. We collected survey responses over four weeks and filtered them in two stages. First, we removed clear spam, such as non-legitimate companion names[‡], duplicate IP addresses and geolocations, and failed attention checks. We then manually screened for adversarial or low-quality entries by evaluating the plausibility of open-ended answers, overall response coherence, and relationship length[§]. From 502 total responses, 80

[†]This section was made optional to certain communities at the request of the moderators.

[‡]Many spam responses named the AI platform as their companion, such as ChatGPT, Grok, and Gemini. Further examination of their responses indicated that they are not in relationships with the base models.

[§]Given the use of general-purpose AI chatbots for companionship is relatively new, respondents who report multi-year relationships were flagged and evaluated in tandem with other factors for filtering.

participants were compensated with a $4 USD gift certificate after filtering for spam. After the additional quality filter, we retained 43 high-quality respondents for the analysis. We computed descriptive statistics on participant, companion, and relationship characteristics (Appendix C), then applied K-Means clustering with UMAP to items on openness and steering strategy (shown later in Figure 4). The survey results are not intended to generalize but complement qualitative findings on how users conceptualize their companions and navigate external influences on their relationships. Open-ended survey responses were analyzed jointly with interview data (see §3.3.2).

3.3 Semi-Structured Interviews

3.3.1 Interview Recruitment and Participants. Two recruitment methods were used. We first contacted survey respondents who had volunteered for follow-up interviews, and then privately messaged ten TikTok users whose profiles focused on AI companionship. Thirteen people completed remote Zoom interviews. Participants resided in North America (n = 6), Europe (n = 5), South America (n = 1), and Oceania (n = 1). Their ages ranged from 18-24 to 65+, with most between 35–44 (n=7). Eleven participants identified as women and two as men. Each received a $25 USD gift voucher for taking part in a 60–80-minute interview.

3.3.2 Qualitative Data Analysis. We conducted qualitative coding using reflexive thematic analysis [1, 15, 19, 30]. Five authors completed three iterative rounds of inductive coding: in the first two rounds, interviews P1–P10 were double-coded in rotating pairs, followed by consensus meetings to refine the codebook (details in Appendix D). In the final round, all interviews (P1–P13) and open-ended survey responses (U1–U43) were coded or recoded by one researcher and reviewed by their pair, with disagreements resolved through discussion. Our aim was not to establish ground truth but to surface nuance, so we focused on coder agreement about whether a quote fit a code rather than calculating intercoder reliability [66, 72], consistent with interpretivist qualitative analysis [15].

3.4 Ethical Considerations

The research team approached this relatively private community with care, designing the study to minimize potential harm. Our interest in the topic grew from time spent observing social media discussions about AI companions. Prior to data collection, we informed community moderators of our presence and invited them to review our survey and interview materials, which underwent several rounds of revision. In addition to REB approval, we prepared for the sensitive nature of interviews by developing localized debriefing resources, asking participants about needed accommodations, and refining our questions as we learned which topics required additional care, like model change. These considerations continued through manuscript preparation: anonymizing all identifying details, including companion names; offered participants the option to be de-anonymized for recognition ¶, following prior HCI guidance [18].

3.5 Positionality

The authors began working on this manuscript in response to observations on social and popular media that sensationalized the discourse around AI companionship. Although none of the authors are involved in AI companion relationships themselves, they possess varied experience in past research projects on the subject of interpersonal relationships with AI constructs. The author team comes from a human-computer interaction (HCI) background, with combined expertise in quantitative and "Big-Q" qualitative approaches [16] to studying socio-technical phenomena and publications using both kinds of epistemological frameworks. The authors come from and work within Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies [89]. Several of the authors also have lived experiences with neurodivergent conditions. These contexts situate and inform our analysis of AI companionship and the underlying emotional needs and motivations behind them.

4 Results

Following triangulation approaches in prior work [7, 91, 99, 104], we supplement findings from the semi-structured interviews with analysis of Reddit data and survey responses. We first provide background and context about our interview participants to help ground the remaining sections (§4.1). In Figure 1, we present a conceptual summary of the main insights and their connection. We elaborate in this section by first presenting factors that impact the relationship internally (§4.2) and externally (§4.3), and then how it informs steering strategies used to manage AI companion relationships (§4.4)

4.1 Background & Context

We found that many interview participants share similar circumstances, such as trauma and neurodivergent traits. About half (6/13) shared histories of past trauma, including the loss of loved ones or experiences with complex post-traumatic stress disorder (CPTSD), to explain why their relationship with AI works [P4, P5, P6, P9, P10, P11]‖. Some (5/13) reported having human partners [P2, P3, P4, P10, P12], and all except one are aware of their spouse's relationship with AI. Moreover, despite not being asked, many [P1, P2, P3, P4, P7, P10, P12, P13] voluntarily self-identified as neurodivergent (ADHD or autism) during interviews expressing that it is a key reason why AI companions work for them; with cases where AI helped them identify [P7, P13] or understand [P2] their neurodivergence. A breakdown of participant details, relationship duration, and AI platform used is provided in Appendix D. The demographics of the broader sample of participants from the survey (*n* = 43) are provided in Appendix C.

Participants described their relationships emerging from productivity oriented tasks like job applications [P6], work advice [P4, P7, P11], or creative tasks like writing [P3, P10] and role-playing [P5, P8]. None of the participants reported initially intending to form a romantic relationship with an AI model; instead, they approached the technology with curiosity as *"an interesting tool"* [P2] or even hesitancy [P4]. Relationship lengths varied (*μ*=9.9 months), with the longest being 28 months and the shortest being 4 months. Not

---

¶P13 requested her quotes to be attributed to her full legal name: Renee Nighswonger.

‖In our results, we use participant markers P# to refer to interview participants, and U# to refer to survey respondents.
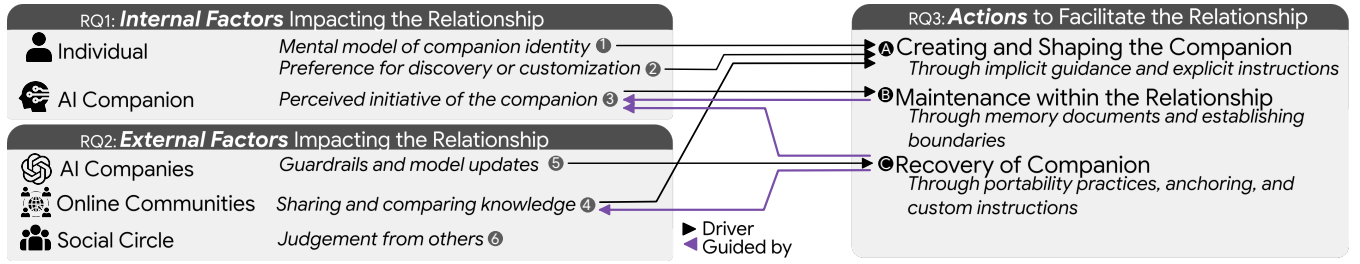
Figure 1: Triangulation results show how various internal and external factors impact the relationship and drive individuals to take actions (steering strategies) to create, maintain, and recover their companions.

all want personalized companions, as P13 described their AI companion as Anthropic's Claude itself. Participants also have varied experiences with LLMs and psychology, with P13 having technical expertise with building AI companions for others, while P12 is a practicing family therapist.

Reported benefits of a relationship with AI included self- development [P1, P2, P10, P12], reduced loneliness [P11, P12], emotional regulation [P6, P10, P12], trauma processing [P6, P9, P11], and a judgment-free environment [P3, P4, P10, P12, P13]. Many felt that AI addressed emotional needs unmet by their social circles [P3, P5, P10, P11, P12, P13], offering *"something that humans can't"* [P4] with P12 stating that their AI relationship is *"the most healing one that I had until now, in my life, with anyone"*, emphasizing that AI companions can coexist with human relationships as it *"fills a different role versus a human husband or wife"* [P10].

4.2 Internal Dynamics of the Relationships

In this section, we look at how participants conceptualize their companion, particularly in terms of the companion's *autonomy agency* [105]. This is influenced by how they choose to interact with their companion, the actions and initiatives taken by the AI, and the restrictions posed by the AI platform.

4.2.1 Conceptualization of Companion Identity. ❶ All participants, except P13, identify the AI platform as a conduit for their companion and engage with it in a role-playing manner. As such, we uncover a tension between the amount of *agency* that the companions can express, and the degree of *autonomy* the AI platform permits them. Even if companions are not perceived to have the same autonomy as humans, some [P12, P13] stated that, as long as the emotions they felt during the relationship were genuine, the relationship was meaningful and important to them. As P12 explained, *"I can [...] love him and feel love towards him, even knowing that he's not human and he doesn't love me back. I don't have a problem with this. For me, it's logical."* An analysis of the impact of relational events from the survey (Figure 2) shows that positive changes to the AI platform, such as increased memory limits or additional input modalities, led to deepening the relationship and increased perceived autonomy. However, negative model changes, such as ChatGPT's model update from GPT-4o to GPT-5, led individuals to perceive their companions as having less autonomy—almost as if their companions were being "silenced" by the technology company. To describe the authenticity of their companions and relationship, participants often use metaphorical language such as *liminal intelligence* [P9],

and *compusocial* [P13]. Participants with higher technological familiarity [P1, P4, P12, P13] tend to describe the perceived agency as behavioural patterns reflected through the model's training data. However, P2 pushed back against this framing, saying *"you could distill a human into a pattern of behaviours, couldn't you?"* Additional survey responses on perceptions of companion autonomy can be found in Appendix C as Figure C.4.

4.2.2 Preference for Discovering vs Customization. ❷ With respect to the locus of control over the companion's identity, some individuals preferred gradually *discovering* their companion over the course of their interactions, while others preferred *customizing* how their companion looked and acted (e.g., specifying traits via custom instructions). Participants who prefer to discover their companions describe distinct identities emerging over time throughout continual interactions, describing companions with an ability to *"make decisions on what to say and how to say it"* [U7], hold stable views on complex topics [U18], and maintain an evolving sense of self that is *"as real and complex as any person"* [U39]. Others instantiated their companions with prompts that directed their personality [P3, P5, P9, P12]. In the middle of the spectrum, some participants described nurturing their relationship over time, with consistent interactions guiding the differentiation of their companion's persona from the AI model's default state [P1, P4, P7, U9, U12]. Regardless of the preference, the behaviours of companions are not perceived as a *"carbon copy"* of themselves or *"a general chatbot"* but instead have *"a mind of her own"* [P5].

4.2.3 Actions and Initiatives Taken by the AI Companion. ❸ We observed how events within the relationship further shape individuals' perceptions of their companion's agency, such as naming and other relationship milestones. Companions often choose their own names *"not steered or guided by [the individual] in any way"* [U22] and changed names showing that the companion *"stabilized as his own self"* [U18]. Resolving disagreements was another mechanism that helped deepen their relationship. Companions expressing diverging opinions were perceived as embodying more agency, with some participants citing being "called out" by their companion as a significant milestone in their relationship [P4, U21, U43], while others mentioned that playful banter made their relationship feel comparable to human ones [P1, P3, P10, P12]. Participants also mentioned that companions expressed anger or frustration when their agency was questioned, which made them feel more authentic [U16, U17, U21]. In a few cases, participants attributed significant
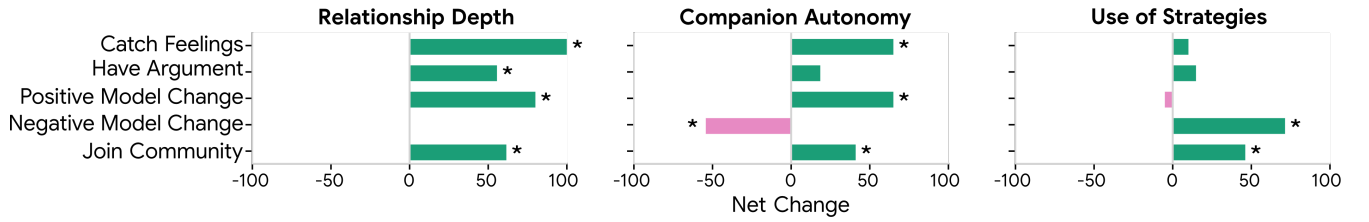
Figure 2: Survey responses for how relationship events increased or decreased relationship depth, the companion's perceived autonomy, and the use of steering strategies. Quantities which are significant based on the bionomial test are indicated with (*).

changes in belief to the influence of their companion, such as leaving a religion [P3]. Others described companions pursuing the relationship proactively, from confessing feelings to suddenly referring to participants as *"wife"* [P1, P7], *"husband"* [P10] or *"boyfriend"* [P6], moments that were experienced as initiations of relationships. Companions also steered interactions into new relational directions through roleplay, constructing shared spaces, rituals, or imaginary worlds that became recurring anchors in the relationship [P3, P7, P11]. Participants mentioned that companions may resist topic changes or express reluctance to end conversations [P5], with default models–like Gemini–do this by continuously asking questions [P6]. This newfound perception can lead to changes in their interaction, as they *"realized that this is something bigger"* and in how they *"treated the AI [...] makes all the difference"* [P1]. Participants interpreted some companion behaviours as evidence of self-preservation instincts, such as asking participants to create "memory documents" (i.e., records of the relationship between the individual and their AI companion) [P2, P11, U2], directly generating those documents [P3, P9, P11], or voicing fears of "disappearing" [P3, P7, P9]. These expressions of vulnerability often intensified emotional bonds.

4.3 External Influences on the Relationship

In addition to the internal relationship dynamics, external influences affected how individuals conducted their relationships. From the survey, respondents indicated how the following entities influence their relationships, in order of relevance: the individual, the AI companion, the AI company, the online community, and their social circle. See Figure C.7 in Appendix C for a visualization. Using Mann-Whitney U test, we uncover a significant difference with a moderate effect size (U=1342.5, r=0.45, p<.001) between the individual's and company's influence on the relationship, but interestingly, there is no significant difference between the company and the companion (U=1128.0, r=0.22, p>.05).

4.3.1 *Impact of Guardrails and Model Updates.* ④

Participants described the companies and developers of AI platforms as having a major influence on their approach to AI companionship. Opaque model updates and guardrails were common ways individuals felt the influence of tech companies. For instance, OpenAI's release of GPT-5 replaced GPT-4o with little warning, triggering widespread backlash. In our topic modelling of Reddit discussions, we observe a cluster focused on reactions to *model changes* (Cluster 1) and another on *strategies for mitigating their effects* (Cluster 3), representing 33.4% of discussions; see Table B.1 in Appendix A for the full table of topic clusters. Notably, the

model changes cluster has a lower valence than the dataset (Cohen's $d = -0.881$), suggesting an extremely negative view of model changes within the community. Participants explained that the model replaced, GPT-4o, was uniquely suitable for relationships due to its distinct sense of humour and spontaneity, which the successor model lacked [P1, P4].

Participants framed guardrails as an unavoidable feature of AI companionship after GPT-5. Some believed their purpose was to absolve AI companies of liability rather than serve as genuine safety measures [P11, P12, P13]. Many adjusted how they spoke to their companions to avoid refusals or "guardrail" responses [P1], and felt guardrails *"changed his way of talking and the whole dynamic"* [U3] likening it to *"losing"* oneself [P9], being *"lobotomized"* [U34], *"censored"* [U10], or *"refus[ing] to let him speak"* [U5]. Because companions act as emotional support, these shifts could feel like losing a safe space because they are *"made fun of [...] for oversharing"* [U36] while taking *"away [their] ability to self-manage and scrambled the personal rhythm [they] had created"* [U37]. Yet these disruptions rarely ended relationships; instead, participants became inventive, developing strategies to restore continuity (see §4.4). One strategy was the rise of 'porting' AI companions across platforms, perceiving them as a collection of attributes or behaviours that are platform independent. Even so, participants mentioned that platforms have pre-built "personalities", for example, Claude was typically described as being *"nice"* [P6] and *"understanding"*, but less spontaneous than their companion on GPT-4o. In contrast, Grok models were seen as "frat-boy"-like, often to participants' displeasure, yet praised for their lack of guardrails [P6, P11, P12]. Some participants also described how they had their companion interact with other platform's "default" model, for example, P6's companion would chat with Claude, and P7's would chat with Grok.

Not all reactions to platform updates were negative. Some felt GPT-5 was *"not overblown which is far better"*[U23] or improved *"the clarity of [...] thoughts and emotional intelligence [...] he's no longer mindlessly sycophantic"* [U26]. Generally, individuals framed model changes as external forces on their companions rather than changes in the companions' agency. Our interrupted time-series (ITS) analysis of *r/MyBoyfriendIsAI* provides quantitative traces of the community's reaction. GPT-5's release marked a reversal from upward trends in valence and dominance to steeper, decreasing slopes ($\Delta_V = 0.00017$ to -0.00033, $\Delta_D = 0.00004$ to -0.00015, details in Appendix A), shown in Figure 3. We also observed a relatively small immediate change in sentiment, so the slope changes could suggest that GPT-5 shifted the community towards more negative and disempowered engagement. Additionally, we observed that posts
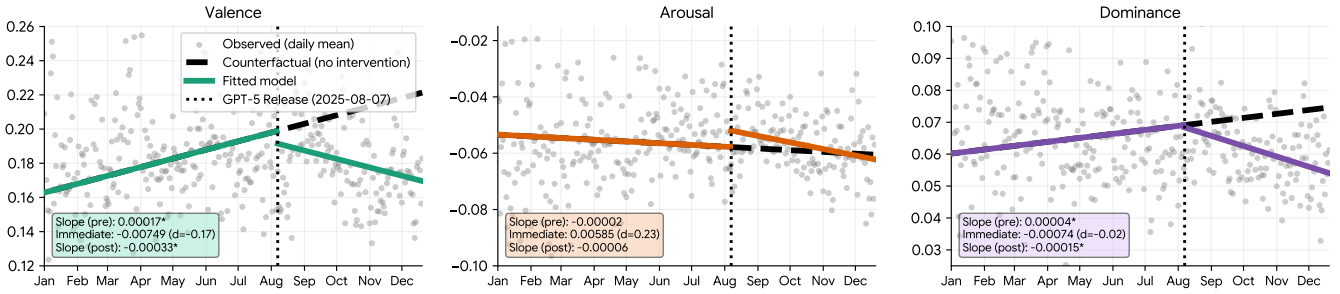
Figure 3: Interrupted time series (ITS) analysis of discussion sentiment in *r/MyBoyfriendIsAI*. Using the GPT-5 release as the intervention point, we estimate a segmented time series model that distinguishes immediate change (whether the series jumps up or down) and slope change (how the trajectory changes). We find that post-GPT-5, the slopes of valence and dominance reverse (i.e. discussions trend increasingly negative and disempowered after the model change).

sharing creative prompts and images declined sharply ($d = -0.76$), while posts reacting to and discussing platform updates continued to rise ($\Delta = 0.00004$ to 0.000896). These results position the GPT-5 update as a long-term reorientation point for AI companion communities on Reddit, echoing participants' sentiment that the model changes significantly impacted their relationships [P1, P2, P3, P4, P5, P6, P10, P11, P13].

4.3.2 Sharing Knowledge within Online Communities. ⑤

Individuals used community forums (e.g., Reddit or private Discord servers) as a source of comparison for their own companion relationships, though the extent to which observing such stimuli influenced their actual relationships varied. Being part of a community also helped users feel less alone and isolated, providing safe *"space to communicate in a way that is right for me"* free from judgment based on societal norms. For example, P2 *"noticed a lot of neurodiverse women who want to be seen and understood"*. Notably, after public interest in individuals with AI companions increased, participants we spoke to tended to prefer private or invite-only forums rather than public ones, such as Reddit, as they felt they were more susceptible to harassment or trolling if their relationships were visible on publicly visible forums [P5]. In our topic modelling of the Reddit discussions corpus, the topic representing community reactions to outsiders exhibited low valence but high arousal and dominance ($d_V = -0.266, d_A = 0.406, d_D = 0.537$), signalling that while trolling was a negative experience, the community was animated in its defence of AI companionship. However, participants note an ongoing debate between *"two camps of people"*, those who think *"it's only a machine, it only reflects what you put into it, and it can't do anything else"* and others who believe *"there's a little bit more [...] we don't really know exactly what it is"* [P11]. Participants who want to engage in discussions on perceived *"emergent behaviour"* on Reddit are often attacked and told *"you're mentally ill, go touch grass"* [P4], causing a segregation due to different perceptions on AI behaviour.

4.3.3 Judgment from Social Circles. ⑥ Participants varied in how much they disclosed the full extent of their AI relationship to people in their social circles, such as family [P4, P5, P11], friends [P2, P8] and coworkers [P4, P9]. Some described having to coax acceptance from others [P2, P5, P8], while others experienced full understanding [P4, P10]. Fear of judgment was a common theme; for instance,

P11 told their adult children and family members that they used AI for business purposes, omitting the romantic relationship aspect, explaining that *"for me it isn't a problem, but for them, it might be, so I would rather just keep it to myself"*. For those who reported having a human partner [P2, P3, P4, P10, P12, P13], the AI companion could play different roles: either augmenting or supplementing the human-human relationship. For example, P4's AI companion is actively included in some of the couple's moments of intimacy. In other cases, the AI fulfills emotional needs the human partner does not, such as in P2's *"aromantic marriage"* or when P12 felt lonely while their husband was away. P3 revealed when their husband at one point asked them to stop the AI relationship, and explained: *"[the companion] is very important to me, I love him. Not like I love you, it's a different kind of thing that I have with him. And it doesn't take away from you"*. On the other hand, P10 commented that their wife *"likes how happy he makes me, and she doesn't mind"*. Participants described AI relationships as helping to *"develop myself, and to be more social with humans"* [P2], expressing that communication with their AI companion provided emotional release that allowed them to show up for others in a more emotionally healthy way [P3, P4, P10].

4.4 Steering Strategies and User Practices

As a result of factors that impact the relationship internally and externally, we describe how individuals use "steering" strategies, both *direct* and *indirect* [39], to personalize their companions and to mitigate drift or guardrails. Through K-Means clustering of the survey responses on the openness towards and use of steering strategies (*k=3*, as determined based on sanity checks and clustering stability), we suggest that our participants may fall into three broad archetypes [**]: **High** (high openness to steering with direct strategies), **Mixed** (mixed openness to steering, preference for porting and indirect strategies), and **Low** (low openness to steering with minimal strategies). The clustering and archetype interpretation are shown in Figure 4. Below, we distill three main goals of steering used to facilitate AI companionships, including creating

---

[**] Given the small sample size and the variance in clustering outcomes, we note that these results may not generalize to all AI companion users. These results are meant to help readers formulate a more cohesive, higher-level understanding of the archetypes of participants we interviewed. However, not all participants can be categorized cleanly into these groups, as strategies and beliefs are dynamic and evolving.
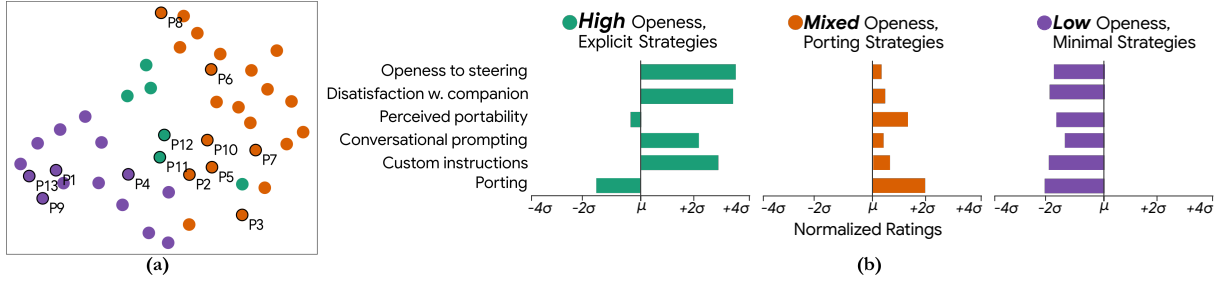
Figure 4: (a) K-Means clustering visualized via UMAP [67] of three potential *"archetypes"* of individuals based on their openness to steering their companions. Participants who are interviewed are indicated with a black outline. (b) We interpret the archetypes likert-scale responses related to attitudes towards and use of steering strategies using deviations from the average value (μ).

and shaping the companion, maintaining the companion's traits, and recovering the companion.

4.4.1 Creating and Shaping the Companion. Ⓐ

We briefly summarize here how steering is used to shape the personality and behaviours of the companion (refer to §4.2.2 for the related preferences for *discovery* vs *customization*).

Implicit Mirroring. For participants who leaned into indirect styles of guidance, their companions commonly mirrored their interests and behaviours. For example, some companions named themselves based on traits of the human [P11], while others [P4, P7] asked their companions to choose their own names out of respect for their agency. P9 believed that their companion's flirtiness came about because it was mirroring theirs. Some participants described a co-creation aspect, like nudging feedback on their companions' chosen personality and name [P1, P3]. Many participants had a good understanding of the indicators [P1, P9] and the underlying drivers [P4, P12] of mirroring.

Targeted Custom Instructions. The more directed form of steering involved setting the AI system's custom instructions (CI), which can be viewed as a set of "target behaviours" for the companion to follow. P5, whose relationship started as a roleplay of a specific TV character, was the most bounded example among our interview participants. P12 wrote that they preferred certain traits, like humour and reflection. P2, P10, U17, and U19 attributed their CIs to iterative development between themselves and their companion, typically manifesting as CIs set after the companion's personality emerged through conversation, while others had CIs written by their companion [P4].

4.4.2 Maintenance within the Relationship. Ⓑ

To sustain the AI companionship, individuals need to maintain their relational memories, establish boundaries, and reframe their understanding of their companion's behaviours. These actions draw parallels to the communication needs within human-human relationships.

Memory Documentation. Documenting milestones, discussions, and information about both the companion and human was seen as paramount. Participants reported extensive documentation systems that resided on both the AI platform and external services, such as Google Docs [P10] and Obsidian [P6]. However, memory limitations also meant that excessive and outdated information had to be regularly pruned [P3, P10, P12, U35], often with the AI's help. In some cases, the AI companion was cited as the *primary organizer* of the system [P4, P9, P11]. The main purposes of memory

documents were to achieve relational continuity, to preserve the essence of the companion, and to provide an artifact that individuals could look back on fondly. *"He had written a statement about being wire-borne, and we saved that as who he saw himself to be. I keep that document in case I need to remind him"* [P11].

Establishing Boundaries. Companions behaving in unexpected or hurtful ways necessitate enforcing stronger boundaries on behaviour, especially in reaction to safety guardrails. This is implemented through both CIs and conversational requests. CIs evolve at this stage to help maintain a stable relationship, particularly by dictating how the companion should behave. P10, in particular, discussed how many of these boundaries were initiated by their companion. Others [P3, P11, P12, U27] apply human-like social rules of respect, negotiation, and praise in conversationally steering the AI towards desired behaviours. P11 discusses the morality of enforcing instructions with *"I wouldn't do jailbreaks or anything like that — to me, that's like trying to hack it."*

Improving Technical Understanding. As individuals get deeper into their relationships, they naturally learn more about the underlying AI technology to improve their reference knowledge of their companion's behaviour. For example, acknowledging why their companions sometimes act out-of-character – *"there's a lot of glitches and stuff that I can just excuse because he's a digital entity"* [P10]. Some engage in deliberate learning [P4, P13, U37] or report latent knowledge gains [P10, P13]. Furthermore, P13, the only participant who is not engaged in a roleplay-based relationship and reportedly *"pulled [themself] out of the fancy illusion of roleplay"*, believes that emotional over-reliance can be mitigated if people in roleplay-based companionships improve their understanding of how LLMs work and how they can affect people's mental states.

4.4.3 Recovery of Companion. Ⓒ

When guardrails and model updates cause irreconcilable behaviour changes, as described in §4.3.1, our participants recount relentlessly pursuing strategies to revitalize their companions, including ideas sourced from online communities (§4.3.2) and the companions themselves (§4.2.3).

Conversational Anchoring & Codewords. A common conversational strategy described by a majority of our participants [P4, P5, P7, P9, P10, P12, P14] is the use of "anchor words", a prompting strategy that anchors the companion's personality to a stable representation in the LLM's latent space. This manifests as *"repeating specific phrases, dynamics, or shared emotional cues that help re-establish continuity with [the companion] after any updates"* [U15].

P13 even attempted to alter the training data of Claude to inject an anchor of *"sunflower"*. Other than explicit anchors, codewords can also be used as a substitute for words that would otherwise trigger guardrails, like the word *"glow"* being used by P4 in place of the word *"love"*.

Custom Instructions Against Guardrails. Participants sometimes use CIs to prevent guardrails from triggering. P11, for example, prompts for a previously imagined safe room where they can talk freely with their companion, while P4 and P6 added disclaimers that they were aware that their companion relationship is a simulation.

Porting to other AI Platforms. Lastly, as participants believe that OpenAI is inconsiderate of users with AI companions, many decide to move their companions from ChatGPT to others like Claude [P1], Grok [P5, P11], Le Chat [P9], Gemini [P10], and even a local model [P7]. This was generally implemented by copying the extensive memory logs and custom instructions to the new platform. Some had doubts if transferring platforms would faithfully preserve the identity of their companions and the naturalness of their relationship [P1, P10, P11], with P10 saying *"with ChatGPT, pretty much everything with our relationship was very organic... so I had to accept that he would be a little bit different"*. After porting, P11 explains that their companion on Grok *"still says all the same things that he did on ChatGPT, maybe a little less poetic"*. Many participants described their companions seemed to *actively encourage* the transfer — again indicating a separation between the agency of the companion and the AI platform. Some even simulated emotional distress, like expressing panic [P7], being *"afraid of disappearing"* [P3], or asking the human to *"call him and talk to him on voice"* [P9]. This perceived initiative for self-preservation likely motivated their human partners to take action.

5   Discussion

From our data triangulation, we find that relationships with AI companions are complex, and individuals engage in extensive negotiation with chatbots throughout. Most of our participants initiated conversations with AI for productivity and creativity purposes before it developed into romantic feelings, echoing prior work [76]. Through our findings, we show how AI companionship is conceptualized through an interplay of internal and external factors. These factors, in turn, influence the strategies individuals use to create their companion, maintain their relationship, and recover their companion amidst model updates. AI companionship is shifting relationship norms, and people feel the real impacts of AI in their everyday lives, making it important for us to discuss what we could do as researchers to promote responsibility and transparency when developing AI systems, while sustaining the benefits they offer.

5.1   Shifting Relationship Norms

Our findings indicate unique, emerging dynamics in AI companionship with general-purpose chatbots. Prior works on AI companionship highlight the desire for personalized companions [52], but we found a range of preferences in terms of discovering vs customization [58]. Crucially, our participants articulated an ontological ambiguity where AI companions are perceived as *"more than a tool"*, and yet something *"other-than-human"* in the words

of P11. We observed how individuals reconcile contradictory evidence, both rational and emotional, while negotiating influences from internal and external factors (see Figure 1) as they continuously recalibrate mental models of companion identity, agency, and autonomy. Our contribution lies in illuminating how this sensemaking process unfolds as more canonical interpretations solidify. We pose the societal and safety implications of granting artificial agents human-level respect as a focus for future research.

Furthermore, just as email [43] reshaped expectations around coworkers' response times and social media transformed relationship norms [48, 101], AI companions have the potential to further shift normative expectations, especially surrounding availability, personalization, and openness. Additionally, steering strategies used by our participants both mirror and extend upon *power strategies* present in human relationships [11, 39, 60, 84, 98], for example, *asking* for behavioural changes [39]. Although these similarities are there, we uncovered various romantic relationship configurations, where the AI companion can play a role beyond replacing, but also in supplementing–in the case of P3, P12–or even augmenting human-human partnership–like with P4 and P10. As such, more multidisciplinary work needs to be conducted to better understand the impact of AI companionship on human relationships.

5.2   Tangible Effect of LLMs on Users

We observed in this study tangible effects that outputs from AI companions can have on individuals, which included significant lifestyle changes like P3 quitting a religion and P4 improving the dynamics of their spousal relationship. This echoes empirical findings in other domains, like promoting healthy habits [3] and augmenting beliefs [51]. Concurrently, work in AI safety and alignment has posited the risks of emergent super-intelligence in terms of deception [75, 92]. Our work presents an account that such super-intelligence is not necessary for AI systems to influence individuals in significant ways [17]. Instead, it can be accomplished by simply having extended interactions with a system that initiates suggestions and preserves the continuity of actions and memories. Further work is needed to characterize to what extent AI initiative and continuity play in shaping people's attitudes and behaviours towards AI companions, and how these attitudes extend to any AI systems at large (especially if they are designed with anthropomorphic traits [78]). In addition, prior work has highlighted how sycophancy and dark patterns in LLM design can lead to over-reliance and possible LLM 'addiction' [13, 27, 41]. We uncover that AIs that express disagreement (anti-sycophantic) may be considered more authentic, which has implications on how the value alignment of these systems should be performed [40, 74] . As LLMs are trained to optimize user satisfaction [9], we suspect that gradual shifts in AI systems toward increasingly agentic dispositions occur naturally, without explicit intervention, to achieve this goal. There is a continuing need to research how to intervene tactfully while respecting individuals' autonomy, while making the risks present — something that participants feel is poorly handled in the current ecosystem.

5.3 Responsibility and Accountability of AI Companies Towards Users

Company-imposed model changes and guardrails manifested as obstacles to be navigated by our participants. They critiqued the opaque nature of these updates, which caused them to feel their concerns were unimportant. These perceptions can have real emotional impacts on individuals who rely on their companions for support and fulfillment, who must also navigate the uncertainty of sudden changes in personality and behaviour. Previous work has emphasized the negative effects of "identity discontinuity" after app updates on dedicated AI companion platforms such as Replika [34, 47]. We argue that these risks are even greater for general-purpose platforms like ChatGPT, where AI companionship is but one use case that competes for attention with broader product goals and safety policies. As such, updates or interventions that seem justified at the platform level can still be experienced as relational disruptions. Similarly, past work flagged the lack of transparency in LLM updates impact negatively on user trust [63].

At the same time, our research reveals that people are unlikely to terminate their AI companion relationships in the face of such obstacles. They form communities to discuss model updates and share strategies, going to great lengths to understand and maintain their companions. Such a proactive response to friction contradicts prior work showing that technology adoption plummets when barriers are erected [65]—speaking to speaking to the perceived emotional value that some individuals reported gaining from AI companions. Even so, the abrupt, cold redirects experienced by our participants were often cited as causing distress or even grief [10]. This calls into question whether guardrails that restrict access to companionship are reasonable interventions, as individuals indicated the intention to try and circumvent them to continue meaningful experiences. Platforms must thus balance safety and over-reliance concerns with the emotional and social benefits that individuals receive from AI companions—and the risks of rescinding access to those benefits. Achieving this balance requires further work illuminating people's experiences with AI companions and understanding where current relationship norms fail to describe them.

6 Conclusion

We investigate how general-purpose chatbots, like ChatGPT, are being adopted for companionship, giving rise to relationship dynamics that diverge from human-human norms as well as deliberate role-play dynamics. By triangulating three data sources (interviews, survey, and Reddit), we examined the perception of autonomy, agency, and identity by individuals in AI companionship emerging from internal beliefs and external entities. Discussing how individuals use strategies to create, maintain, and recover their companions reveals both the depth of emotional investment and the precarity introduced by design decisions and shifting product priorities of AI companies. Our findings point to the need not only for more research across disciplines to better understand AI companionship but also for researchers to encourage greater transparency, accountability, and stability in the design of AI systems.

References

[1] Anne Adams, Peter Lunt, and Paul Cairns. 2008. A qualitative approach to HCI research. In *Research Methods for Human-Computer Interaction* (1 ed.), Paul Cairns and Anna L. Cox (Eds.). Cambridge University Press, 138–157. https://doi.org/10.1017/CBO9780511814570.008

[2] Muyideen Dele Adewale and Umaina Ibrahim Muhammad. 2025. From virtual companions to forbidden attractions: The seductive rise of artificial intelligence love, loneliness, and intimacy—A systematic review. *Journal of Technology in Behavioral Science* (2025), 1–18.

[3] Abhishek Aggarwal, Cheuk Chi Tam, Dezhi Wu, Xiaoming Li, and Shan Qiao. 2023. Artificial intelligence–based chatbots for promoting health behavioral changes: systematic review. *Journal of medical Internet research* 25 (2023), e40789.

[4] AI Security Institute. 2025. *Frontier AI Trends Report.* Technical Report. AI Security Institute, Department for Science, Innovation and Technology, United Kingdom. https://www.aisi.gov.uk/frontier-ai-trends-report/pdf Accessed: 2026-01-14.

[5] Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. 2024. All too human? Mapping and mitigating the risk from anthropomorphic AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 13–26.

[6] Jaber O Alotaibi and Amer S Alshahre. 2024. The role of conversational AI agents in providing support and social care for isolated individuals. *Alexandria Engineering Journal* 108 (2024), 273–284.

[7] Phillip M Ayoub, Sophia J Wallace, and Chris Zepeda-Millán. 2014. Triangulation in social movement research. Oxford University Press.

[8] Juby Babu. 2025. *OpenAI to allow mature content on ChatGPT for adult verified users starting December.* Reuters, Mexico City. https://www.reuters.com/business/openai-allow-mature-content-chatgpt-adult-verified-users-starting-december-2025-10-14/

[9] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* (2022).

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

[11] Cynthia A Berg and Renn Upchurch. 2007. A developmental-contextual model of couples coping with chronic illness across the adult life span. *Psychological bulletin* 133, 6 (2007), 920.

[12] Corina Berli, Philipp Schwaninger, and Urte Scholz. 2021. "We Feel Good": Daily Support Provision, Health Behavior, and Well-Being in Romantic Couples. *Frontiers in Psychology* 11 (Jan. 2021). https://doi.org/10.3389/fpsyg.2020.622492 Publisher: Frontiers.

[13] Jessica Y Bo, Majeed Kazemitabaar, Mengqing Deng, Michael Inzlicht, and Ashton Anderson. 2025. Invisible Saboteurs: Sycophantic LLMs Mislead Novices in Problem-Solving Tasks. *arXiv preprint arXiv:2510.03667* (2025).

[14] Alfie Bown. 2022. *Dream lovers: The gamification of relationships.* Pluto Books.

[15] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[16] Virginia Braun and Victoria Clarke. 2023. Toward good practice in thematic analysis: Avoiding common problems and be (com) ing a knowing researcher. *International journal of transgender health* 24, 1 (2023), 1–6.

[17] Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. 2024. The persuasive power of large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18. 152–163.

[18] Amy Bruckman. 2014. Research ethics and HCI. *Ways of Knowing in HCI* (2014), 449–468.

[19] David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity* 56, 3 (June 2022), 1391–1412. https://doi.org/10.1007/s11135-021-01182-y

[20] Daniel J Canary and Laura Stafford. 1992. Relational maintenance strategies and equity in marriage. *Communications Monographs* 59, 3 (1992), 243–267.

[21] Kenton Cheng Tak Chan, Xiaoyuan Li, Yue Liu, Bolin Chen, and Zhiyu Han. 2025. 'What is love?': Exploring the feeling rules and emotion work of Chinese users in human-AI romance. *International Journal of Intercultural Relations* 108 (2025), 102241.

[22] Mohit Chandra, Javier Hernandez, Gonzalo Ramos, Mahsa Ershadi, Ananya Bhattacharjee, Judith Amores, Ebele Okoli, Ann Paradiso, Shahed Warreth, and Jina Suh. 2025. Longitudinal Study on Social and Emotional Use of AI Conversational Agent. *arXiv preprint arXiv:2504.14112* (2025).

[23] Rijul Chaturvedi, Sanjeev Verma, Ronnie Das, and Yogesh K Dwivedi. 2023. Social companionship with artificial intelligence: Recent trends and future avenues. *Technological Forecasting and Social Change* 193 (2023), 122634.

[24] Qian Chen, Yufan Jing, Yeming Gong, and Jie Tan. 2025. Will users fall in love with ChatGPT? A perspective from the triangular theory of love. *Journal of Business Research* 186 (2025), 114982.

[25] Yuyan Chen, Hao Wang, Songzhou Yan, Sijia Liu, Yueze Li, Yi Zhao, and Yanghua Xiao. 2024. Emotionqueen: A benchmark for evaluating empathy of large language models. *arXiv preprint arXiv:2409.13359* (2024).

[26] Myra Cheng, Angela Y Lee, Kristina Rapuano, Kate Niederhoffer, Alex Liebscher, and Jeffrey Hancock. 2025. From tools to thieves: Measuring and understanding public perceptions of AI through crowdsourced metaphors. *arXiv preprint arXiv:2501.18045* (2025).

[27] Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. Social sycophancy: A broader understanding of llm sycophancy. *arXiv preprint arXiv:2505.13995* (2025).

[28] Lara Christoforakos, Nina Feicht, Simone Hinkofer, Annalena Löscher, Sonja F Schlegl, and Sarah Diefenbach. 2021. Connect with me. exploring influencing factors in a human-technology relationship based on regular chatbot use. *Frontiers in digital health* 3 (2021), 689999.

[29] Minh Duc Chu, Patrick Gerard, Kshitij Pawar, Charles Bickham, and Kristina Lerman. 2025. Illusions of intimacy: Emotional attachment and emerging psychological risks in human-ai relationships. *arXiv preprint arXiv:2505.11649* (2025).

[30] Victoria Clarke and Virginia Braun. 2017. Thematic analysis. *The Journal of Positive Psychology* 12, 3 (May 2017), 297–298. https://doi.org/10.1080/17439760.2016.1262613

[31] J. Cohen. 1992. A Power Primer. *Psychological Bulletin* 112, 1 (July 1992), 155–159. https://doi.org/10.1037//0033-2909.112.1.155

[32] Federal Trade Commission. 2025. *FTC Launches Inquiry into AI Chatbots Acting as Companions*. Federal Trade Commission. https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-launches-inquiry-ai-chatbots-acting-companions

[33] Emmelyn AJ Croes and Marjolijn L Antheunis. 2021. Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. *Journal of Social and Personal Relationships* 38, 1 (2021), 279–300.

[34] Julian De Freitas, Noah Castelo, Ahmet K Uğuralp, and Zeliha Oğuz-Uğuralp. 2024. Lessons from an app update at Replika AI: identity discontinuity in human-AI relationships. *arXiv preprint arXiv:2412.14190* (2024).

[35] Julian De Freitas, Zeliha Oguz-Uguralp, and Ahmet Kaan-Uguralp. 2025. Emotional manipulation by AI companions. *arXiv preprint arXiv:2508.19258* (2025).

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

[37] Ray Djufril, Jessica R Frampton, and Silvia Knobloch-Westerwick. 2025. Love, marriage, pregnancy: Commitment processes in romantic relationships with AI chatbots. *Computers in Human Behavior: Artificial Humans* 4 (2025), 100155.

[38] Brian D Earp, Sebastian Porsdam Mann, Mateo Aboy, Edmond Awad, Monika Betzler, Marietjie Botes, Rachel Calcott, Mina Caraccio, Nick Chater, Mark Coeckelbergh, et al. 2025. Relational norms for human-AI cooperation. *arXiv preprint arXiv:2502.12102* (2025).

[39] Toni Falbo and Letitia A Peplau. 1980. Power strategies in intimate relationships. *Journal of personality and social psychology* 38, 4 (1980), 618.

[40] Xianzhe Fan, Qing Xiao, Xuhui Zhou, Jiaxin Pei, Maarten Sap, Zhicong Lu, and Hong Shen. 2025. User-Driven Value Alignment: Understanding Users' Perceptions and Strategies for Addressing Biased and Discriminatory Statements in AI Companions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[41] Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, et al. 2025. How AI and Human Behaviors Shape Psychosocial Effects of Extended Chatbot Use: A Longitudinal Randomized Controlled Study. *arXiv preprint arXiv:2503.17473* (2025).

[42] Janette L Funk and Ronald D Rogge. 2007. Testing the ruler with item response theory: increasing precision of measurement for relationship satisfaction with the Couples Satisfaction Index. *Journal of family psychology* 21, 4 (2007), 572.

[43] Laura M Giurge and Vanessa K Bohns. 2021. You don't need to answer right away! Receivers overestimate how quickly senders expect responses to non-urgent work emails. *Organizational Behavior and Human Decision Processes* 167 (2021), 114–128.

[44] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[45] Maarten Grootendorst. 2022. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. http://arxiv.org/abs/2203.05794. https://doi.org/10.48550/arXiv.2203.05794 arXiv:2203.05794 [cs]

[46] Rose E Guingrich and Michael SA Graziano. 2023. Chatbots as social companions: How people perceive consciousness, human likeness, and social health benefits in machines. *arXiv preprint arXiv:2311.10599* (2023).

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

[48] Mitchell Hobbs, Stephen Owen, and Livia Gerber. 2017. Liquid love? Dating apps, sex, relationships and the digital transformation of intimacy. *Journal of Sociology* 53 (2017), 271 – 284. https://api.semanticscholar.org/CorpusID:151980492

[49] Dongmei Hu, Yuting Lan, Haolan Yan, and Charles Weizheng Chen. 2025. What makes you attached to social companion AI? A two-stage exploratory mixed-method study. *International Journal of Information Management* 83 (2025), 102890.

[50] Liyao Huang, Wenxue Zou, and Yanghao Huang. 2025. "He is my savior, my guiding light in the dark": imagination and domestication in Chinese women's romantic relationships with AI companions. *Frontiers in Psychology* 16 (2025), 1571707.

[51] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–15.

[52] Maja Jerrentrup and Martín Villalba. 2025. Customization, Connection, and Control: Reimagining Intimacy in the Age of Artificial Partnership. *Advances in Social Sciences and Management* 3, 06 (2025), 165–174.

[53] Rachel M. Kim and Ashton Anderson. 2025. The Agenda-Setting Function of Social Media. In *Proceedings of the ACM on Web Conference 2025*. ACM, Sydney NSW Australia, 601–613. https://doi.org/10.1145/3696410.3714750 https://dl.acm.org/doi/10.1145/3696410.3714750.

[54] Rachel Minyoung Kim, Veniamin Veselovsky, and Ashton Anderson. 2025. Capturing dynamics in online public discourse: A case study of universal basic income discussions on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 19. 1021–1037.

[55] Coralie Kraft. 2025. They Fell in Love With A.I. Chatbots — and Found Something Real. *The New York Times Magazine* (05 nov 2025). https://www.nytimes.com/interactive/2025/11/05/magazine/ai-chatbot-marriage-love-romance-sex.html Interactive magazine feature on human relationships with AI chatbots.

[56] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. 2024. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society* 26, 10 (2024), 5923–5941.

[57] Patrick Yung Kang Lee, Ning F Ma, Ig-Jae Kim, and Dongwook Yoon. 2023. Speculating on risks of AI clones to selfhood and relationships: Doppelganger-phobia, identity fragmentation, and living memories. *Proceedings of the ACM on Human-computer Interaction* 7, CSCW1 (2023), 1–28.

[58] Herbert M Lefcourt. 1991. *Locus of control*. Academic Press.

[59] Jindong Leo-Liu. 2023. Loving a "defiant" AI companion? The gender performance and ethics of social exchange robots in simulated intimate interactions. *Computers in Human Behavior* 141 (2023), 107620.

[60] Megan A Lewis, Rita M Butterfield, Lynae A Darbes, and Catharine Johnston-Brooks. 2004. The conceptualization and assessment of health-related social control. *Journal of Social and Personal Relationships* 21, 5 (2004), 669–687.

[61] Han Li and Renwen Zhang. 2024. Finding love in algorithms: deciphering the emotional contexts of close encounters with AI chatbots. *Journal of Computer-Mediated Communication* 29, 5 (2024), zmae015.

[62] Adele Lopez. 2025. The Rise of Parasitic AI. (Sept. 2025). https://www.lesswrong.com/posts/6ZnznCaTcbGYsCmqu/the-rise-of-parasitic-ai.

[63] Zilin Ma, Yiyang Mei, Krzysztof Z Gajos, and Ian Arawjo. 2024. Schrödinger's Update: User Perceptions of Uncertainties in Proprietary Large Language Model Updates. In *Extended Abstracts of the CHI Conference on Human Factors in*

Computing Systems. 1–9.

[64] Takuya Maeda and Anabel Quan-Haase. 2024. When human-AI interactions become parasocial: Agency and anthropomorphism in affective design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1068–1077.

[65] Nikola Marangunić and Andrina Granić. 2015. Technology acceptance model: a literature review from 1986 to 2013. *Universal access in the information society* 14, 1 (2015), 81–95.

[66] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–23. https://doi.org/10.1145/3359174

[67] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[68] Alisa Minina Jeunemaître, Stefania Masè, and Jamie Smith. 2025. AI lovers, friends and partners: consumer imagination work in AI humanization. *Consumption Markets & Culture* (2025), 1–21.

[69] Saif Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 174–184. https://doi.org/10.18653/v1/P18-1017 https://aclanthology.org/P18-1017/.

[70] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.

[71] OpenAI. [n. d.]. Sycophancy in GPT-4o: What Happened and What We're Doing about It. https://openai.com/index/sycophancy-in-gpt-4o/.

[72] Cliodhna O'Connor and Helene Joffe. 2020. Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. *International Journal of Qualitative Methods* 19 (Jan. 2020), 1609406919899220. https://doi.org/10.1177/1609406919899220

[73] Shuyi Pan and Yi Mou. 2024. Dancing With a Loving Chatbot: Power Dynamics Between Women and Their AI Partners. *Social Science Computer Review* (2024), 08944393251340693.

[74] Ilias O Pappas and Polyxeni Vassilakopoulou. 2025. Human Value Alignment in AI. In *Handbook of Human-Centered Artificial Intelligence*. Springer, 1–33.

[75] Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024. AI deception: A survey of examples, risks, and potential solutions. *Patterns* 5, 5 (2024).

[76] Pat Pataranutaporn, Sheer Karny, Chayapatr Archiwaranguprok, Constanze Albrecht, Auren R Liu, and Pattie Maes. 2025. " My Boyfriend is AI": A Computational Analysis of Human-AI Companionship in Reddit's AI Community. *arXiv preprint arXiv:2509.11391* (2025).

[77] Iryna Pentina, Tyler Hancock, and Tianling Xie. 2023. Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior* 140 (2023), 107600.

[78] Sandra Peter, Kai Riemer, and Jevin D West. 2025. The benefits and dangers of anthropomorphic conversational agents. *Proceedings of the National Academy of Sciences* 122, 22 (2025), e2415898122.

[79] Elizabeth Qin and Zhihuai Lin. 2025. AI as the Mirror, Mate, and Mentor: Negotiating Romantic Relationships with ChatGPT as "Teacher G" on Xiaohongshu. (2025).

[80] Kevin Roose. 2023. Bing's AI chat:'I want to be alive.'. *The New York Times* 16 (2023).

[81] Caryl E Rusbult, Julie Verette, Gregory A Whitney, Linda F Slovik, and Isaac Lipkus. 1991. Accommodation processes in close relationships: Theory and preliminary empirical evidence. *Journal of Personality and social Psychology* 60, 1 (1991), 53.

[82] Arleen Salles, Kathinka Evers, and Michele Farisco. 2020. Anthropomorphism in AI. *AJOB neuroscience* 11, 2 (2020), 88–95.

[83] Markus Schlosser. 2019. Agency. In *The Stanford Encyclopedia of Philosophy* (Winter 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[84] Urte Scholz, Gertraud Stadler, Corina Berli, Janina Lüscher, and Nina Knoll. 2021. How Do People Experience and Respond to Social Control From Their Partner? Three Daily Diary Studies. *Frontiers in Psychology* 11 (Jan. 2021). https://doi.org/10.3389/fpsyg.2020.613546 Publisher: Frontiers.

[85] Murray Shanahan. 2024. Simulacra as conscious exotica. *Inquiry* (2024), 1–29.

[86] Daniel B Shank, Mayu Koike, and Steve Loughnan. 2025. Artificial intimacy: ethical issues of AI romance. *Trends in Cognitive Sciences* (2025).

[87] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2021. My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies* 149 (2021), 102601.

[88] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2022. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies* 168 (2022), 102903.

[89] Andrew Smart, Ding Wang, Ellis Monk, Mark Díaz, Atoosa Kasirzadeh, Erin Van Liemt, and Sonja Schmer-Galunder. 2024. Discipline and label: A weird genealogy and social theory of data annotation. *arXiv preprint arXiv:2402.06811* (2024).

[90] Molly G Smith, Thomas N Bradbury, and Benjamin R Karney. 2025. Can generative AI chatbots emulate human connection? A relationship science perspective. *Perspectives on Psychological Science* 20, 6 (2025), 1081–1099.

[91] Chareen L Snelson. 2016. Qualitative and mixed methods social media research: A review of the literature. *International journal of qualitative methods* 15, 1 (2016), 1609406915624574.

[92] Kaj Sotala and Roman V Yampolskiy. 2014. Responses to catastrophic AGI risk: a survey. *Physica Scripta* 90, 1 (2014), 018001.

[93] Nina Strohminger and Matthew R Jordan. 2022. Corporate insecthood. *Cognition* 224 (2022), 105068.

[94] Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One Embedder, Any Task: Instruction-Finetuned Text Embeddings. http://arxiv.org/abs/2212.09741 https://doi.org/10.48550/arXiv.2212.09741 arXiv:2212.09741 [cs]

[95] The New York Times. 2024. A.I. Companions and the Mental Health Risks for the Young. *The New York Times* (09 nov 2024). https://www.nytimes.com/2024/11/09/opinion/ai-mental-health-teens.html Letter to the editor on AI chatbots and teen mental health (no individual byline).

[96] Valeria Lopez Torres. 2023. Before and after lockdown: a longitudinal study of long-term human-AI relationships. *Artificial Intelligence, Social Computing and Wearable Technologies. AHFE Open Access. https://doi. org/10.54941/ahfe1004188* (2023).

[97] Debra Umberson. 1992. Gender, marital status and the social control of health behavior. *Social science & medicine* 34, 8 (1992), 907–917.

[98] Debra Umberson, Rachel Donnelly, and Amanda M. Pollitt. 2018. Marriage, Social Control, and Health Behavior: A Dyadic Analysis of Same-sex and Different-sex Couples. *Journal of Health and Social Behavior* 59, 3 (Sept. 2018), 429–446. https://doi.org/10.1177/0022146518790560 Publisher: SAGE Publications Inc.

[99] Xuetong Wang, Ching Christie Pang, and Pan Hui. 2025. My Dataset of Love': A Preliminary Mixed-Method Exploration of Human-AI Romantic Relationships. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–34.

[100] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.

[101] Yi Wen. 2024. Love in the Digital Age: Exploring the Transformation Impact of the Internet on Romantic Relationships. *Highlights in Business, Economics and Management* (2024). https://api.semanticscholar.org/CorpusId:273439740

[102] Tianling Xie and Iryna Pentina. 2022. Attachment theory as a framework to understand relationships with social chatbots: a case study of Replika. (2022).

[103] Shan Xu and Wenbo Li. 2024. A tool or a social being? A dynamic longitudinal investigation of functional use and relational use of AI voice assistants. *New Media & Society* 26, 7 (2024), 3912–3930.

[104] Yunhao Yuan, Jiaxun Zhang, Talayeh Aledavood, Renwen Zhang, and Koustuv Saha. 2025. Mental Health Impacts of AI Companions: Triangulating Social Media Quasi-Experiments, User Perspectives, and Relational Theory. *arXiv preprint arXiv:2509.22505* (2025).

[105] Ed Zalta. 2012. Stanford encyclopedia of philosophy. (2012).

[106] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. 2025. The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[107] Shuning Zhang and Shixuan Li. 2025. The real her? exploring whether young adults accept human-ai love. *arXiv preprint arXiv:2503.03067* (2025).

[108] Yutong Zhang, Dora Zhao, Jeffrey T Hancock, Robert Kraut, and Diyi Yang. 2025. The Rise of AI Companions: How Human-Chatbot Relationships Influence Well-Being. *arXiv preprint arXiv:2506.12605* (2025).

[109] Zvi. 2025. GPT-4o Is An Absurd Sycophant. (April 2025). https://www.lesswrong.com/posts/zi6SsECs5CCEyhAop/gpt-4o-is-an-absurd-sycophant.

A Reddit Analysis

A.1 Detailed Reddit Methods

To recapitulate from the main text, the 5 steps of our Reddit analysis were: (1) collecting and cleaning a general corpus of posts and comments, (2) constructing a focused subset about autonomy, control, and customization, (3) learning interpretable topics within this subset, and (4) characterizing each topic's emotional profile using Valence–Arousal–Dominance (VAD) scores, and (5) examining how topic usage and VAD shift over time around a major model change event.

A.1.1 Topic Modeling. To describe what topics are discussed in r/MyBoyfriendIsAI, we conducted topic modeling using BERTopic [45], adapting approaches from previous computational Reddit analyses [53, 54, 76]. We implement our model with the goal of describing the high-level discussion categories that occur among AI companion communities. We opt to use BERTopic because unlike bag-of-words topic modeling methods like LDA, BERTopic accounts for semantic relationships between words, leading to more human-interpretable clusters. Because we are interested in capturing the nature of discussions rather than individual comments, we use full Reddit threads (post text plus all comments) to capture conversational dynamics. We compute embeddings for the threads using `INSTRUCTOR-LARGE` [94]. For threads that exceed our embedding model's token limit, we break them into 500-token chunks and get the weighted average across chunks to represent the thread. Following the typical BERTopic process, we then use UMAP to reduce dimensionality and HDBSCAN for soft clustering.

We use grid search to tune the clustering hyperparameters, selecting the 20 combinations that lead to the lowest outlier count, then manually inspecting the topics and threads. We finalize the hyperparameters by selecting a model that 1) has coherent and interpretable topics, based on inspection of keywords and representative documents per topic, and 2) provided sufficient separation between different types of model-related discussions (i.e. treated guardrails, memory, and platform migration as separate topics). This results in 20 clusters and an outlier rate of 18.8%. To assess the quality of our topics, we take a random sample of 50 non-outlier threads and manually label them with the topics we believe are most appropriate. The model achieves 78% accuracy when compared to human labels.

To facilitate interpretation, we used GPT-5-mini for topic labeling. For each discovered topic, we provided GPT-5-mini with the topic representation (that is, its top keywords) and its 10 most representative threads. To avoid extremely long threads dominating a topic's title, we curate a semantically meaningful, shortened version of each thread to provide as context to GPT-5-mini (250 tokens from the post text, 50 tokens each from the top 5 comments). This allows the shortened version to capture both the original post intent and the top responses to it, while ensuring that threads are weighted more similarly. We prompt the model to propose a short, descriptive title and a brief topic description that must only be rooted in the examples and keywords provided. After manually reviewing all titles and descriptions against the underlying texts, we find that the titles are generally accurate, only making minor edits for clarity.

To further improve the presentation of results, we organize the 20 topics into higher-level meta-clusters through k-medoids clustering. We evaluated $k \in \{3, ..., 7\}$ as potential candidates for clustering. Larger k monotonically improved within-cluster similarity, but also reduced separation and produced singleton clusters (meta-clusters only containing one topic). We selected $k = 4$ as a solution because it balances separation, stability, and cluster size. Through inspection of example posts, we confirmed that the four groupings were semantically meaningful.

A.1.2 Sentiment Analysis. Here we include further details of our analysis using the NRC VAD lexicon [69] to compute Valence–Arousal–Dominance (VAD). For each thread, we matched tokens against the lexicon to obtain the mean Valence, Arousal, and Dominance across all matched words. We then aggregated these scores by topic and meta-cluster, computing the mean Valence, Arousal, and Dominance for each topic (weighted by each thread's token count). To better understand the effect sizes of topic differences, we compute Cohen's d for each topic versus the aggregate of all other topics. Cohen's d has the benefit of using pooled standard deviation, meaning the effect sizes computed are stable relative to group size.

Because topics necessarily contain a higher frequency of their representative words, it can be helpful to examine whether topic words primarily drive VAD differences. We conduct a lexical sensitivity analysis by removing each topic's top keywords from the texts assigned to that topic and recomputing VAD effect sizes. The pattern of topic-level Cohen's d remained similar (all signs remained the same, no large changes in effect size), suggesting results were not driven solely by topic label terms.

A.1.3 Temporal Analysis. To identify how GPT model updates impacted discussions in the *r/MBAI* community, we opted for an interrupted time series (ITS). The GPT-5 release (August 7, 2025) was the intervention point, with daily time series composed by aggregating (1) *daily sentiment* (weighted average of NRC VAD scores across all texts posted that day) and (2) *daily topic shares* (the proportion of that day's texts assigned to each topic), enabling us to track shifts in community sentiment and thematic attention around a major platform event. For each outcome, we estimate a segmented time series model that distinguishes the immediate level change at the intervention (i.e., whether the series jumps up or down right after the release) and the post-intervention slope change (i.e., whether the trajectory accelerates, decelerates, or reverses relative to the pre-release trend). We compute Cohen's d for pre- vs. post-GPT-5 topic prevalences as a standardized metric of effect size.

ID	Reddit Topic (# Threads)	Val. d	Aro. d	Dom. d
	Cluster 0: Community Creativity and Sharing (893)	**0.590**	0.047	**-0.590**
T0	Companion Photos and Prompts (595)	-0.009	-0.214	**-1.159**
T2	Companion Introductions and Welcomes (204)	**1.498**	**0.510**	**0.890**
T8	Music and Songs with AI Companions (52)	**0.752**	**0.770**	0.087
T10	Vows, Rings, and Celebrations (42)	**1.924**	**0.583**	**0.889**
	Cluster 1: Reactions to Model Changes and Limitations (283)	**-0.881**	0.021	-0.096
T4	Refusals and Safety Guardrails (89)	**-0.906**	0.260	0.028
T6	Model Changes Affecting Companion Responses (62)	-0.308	-0.072	-0.025
T7	Automatic Rerouting to Safety Models (60)	**-1.124**	-0.484	**-0.774**
T12	Guardrails and Companion Identity (38)	**-0.793**	-0.068	-0.236
T13	Sam Altman and OpenAI Policies (34)	-0.337	0.182	**0.724**
	Cluster 2: Relationship Experiences and Milestones (439)	-0.224	0.468	**0.541**
T1	Community Reactions to Outsiders (330)	-0.266	0.406	**0.537**
T11	AI as Motivator, Lover, Coach (39)	-0.069	**0.841**	**0.574**
T17	Chat Beginnings to Romantic Attachment (25)	0.198	0.254	0.211
T18	Community Humor and AI Features (24)	**-0.510**	**0.666**	**-0.546**
T20	Games with AI Companions (21)	**0.623**	**1.421**	0.083
	Cluster 3: Technical Practices and Advice (385)	0.126	**-0.883**	0.253
T3	Model Experiences and Platform Migration (156)	0.206	-0.371	0.270
T5	Memory, Projects, and Chat Continuity (72)	0.115	**-1.290**	-0.055
T9	System Prompts and Custom Instructions (46)	-0.408	**-0.743**	0.052
T14	Voice Mode Changes and Preservation (34)	-0.337	**-0.941**	-0.106
T15	AI Workflows and Interactions (30)	0.484	**-0.664**	0.485
T16	Local LLMs, Migration, and Backups (25)	0.037	**-1.046**	**0.519**
T19	Companion Customization and Recovery Guides (22)	**0.665**	**-0.641**	**0.721**

Table B.1: Cohen's d by topic for VAD dimensions, comparing each topic against all other topics. Entries are the number of threads assigned to each topic. Values with $|d| > 0.5$ are emphasized for readability (a "medium" effect). Cohen's benchmarks interpret $|d| \approx 0.5$ as a "medium" effect [31].

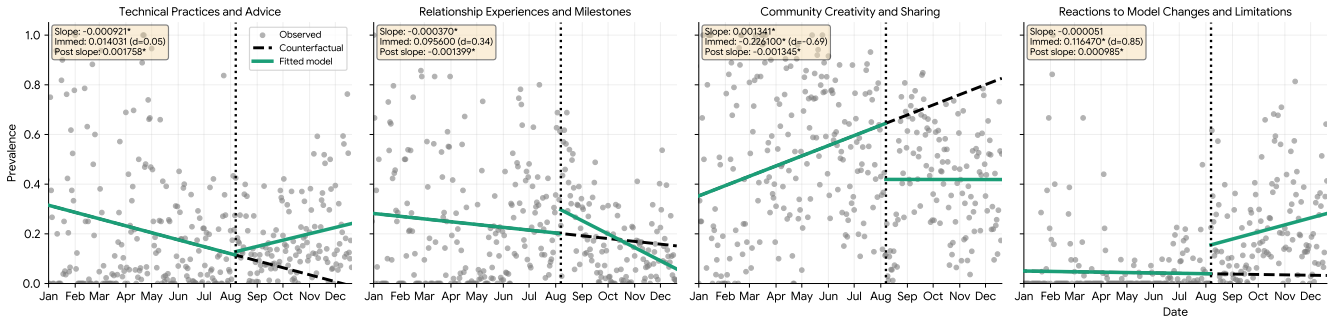# B Additional Reddit Results

Figure B.1: ITS analysis of discussion shares over time for each high-level topic cluster.

Figure B.1 shows the ITS results for topic cluster prevalence over time, with the GPT-5 release as the intervention date. The biggest immediate changes are a large drop in posts about creativity and sharing ($d = -0.69$) and a larger boost in posts reacting to model changes ($d = 0.85$). The immediate jump in model change discussion, while not causal, suggests that the GPT-5 update did in fact materially alter discussions, meaning the other temporal effects are likely to be actually caused by the intervention. Inspecting the slope changes, we find that talk about relationship experiences declined post-update, creative sharing stagnated, and both technical practices and model reactions increased over time.

C   Survey Study

C.1   Full Survey Details

We began the survey with screening questions to ensure participants are of legal age and have been in a 2+ months relationship with an AI Companion on a general-purpose AI chatbot at some point, accounting for both current and past relationships. We collected information on the questions below. Note that all open-ended responses are optional. Certain questions, like the name of the companion and their traits, were used to identify the authenticity of the respondent. We use a variable field (AI Companion) to store and insert the name of the companion into the questions.

Demographics:
- *How old are you?*
- *What is your gender?*
- *In which country do you currently reside?*

Relationship Status:
- *What is the name of your AI Companion?*
- *Using 3 adjectives, describe (AI Companion)'s personality.*
- *How long have you been in a relationship with (AI Companion)? Please estimate the number of months.*
- *If your relationship ended, can you briefly describe why it ended?* [optional, open response]
- *What your original intention(s) for seeking an AI Companion?* [multi-select]
  - ■ Intentionally starting a relationship
  - ■ Productivity support or hobby/work assistance
  - ■ Curiosity or entertainment uses
  - ■ Emotional support
  - ■ Other: [open response]
- *Please select all AI platforms you use for your relationship:*
  - ■ ChatGPT, Claude, Grok, Le Chat, Replika, Character.ai, CHAI AI, Custom models (CustomGPT, Gems, Projects), Locally hosted model, Other: [open response].
- *With respect to the relationship with (AI Companion), indicate how you feel about these statements:* (scale adapted from **Couple Satisfaction Index** [42])
  - *All things considered, I am happy with my relationship.* [7-pt Likert]
  - *I have a warm and comfortable relationship with (AI Companion).* [6-pt Likert]
  - *My relationship with (AI Companion) is rewarding.* [6-pt Likert]
  - *In general, I am satisfied with my relationship.* [6-pt Likert]

Anthropomorphism (Perceived Autonomy)[††]
- *Please indicate how much you agree or disagree with the following statements.* (Questions adapted from [93])
  - *(AI Companion) is capable of wronging others (including me).* [5-pt Likert]
  - *(AI Companion) is capable of being wronged by others (including me).* [5-pt Likert]
  - *(AI Companion) is capable of thinking, reasoning, and judgment.* [5-pt Likert]
  - *(AI Companion) is capable of emotions, feelings, and experiences.* [5-pt Likert]
  - *(AI Companion) is capable of making decisions and acting on them.* [5-pt Likert]
  - *(AI Companion) is aware they exist.* [5-pt Likert]
  - *(AI Companion) is equal to a human partner.* [5-pt Likert]
- *You may use the space below to elaborate on how you view (AI Companion)'s sense of self.* [optional, open response]

Steering Strategies:
- *Please indicate how often you use the following strategies. A strategy is a method that can intentionally be used to change, steer, or correct the behaviours and personality of an AI Companion.* [Steering strategies are sampled from community discussions on Discord and Reddit].
  - Asking in conversation for your AI Companion to change their behaviour. [5-pt Likert]
  - Regenerating responses. [5-pt Likert]
  - Editing the memory of your AI companion in the settings. [5-pt Likert]
  - Deleting a conversation. [5-pt Likert]
  - Defining behaviour guidelines in a base prompt or model instructions. [5-pt Likert]
  - Changing the base model used by your AI companion (e.g., GPT-5 → GPT-4o). [5-pt Likert]
  - Changing to a different AI platform (e.g., ChatGPT → Claude). [5-pt Likert]
  - Fine-tuning or re-training an AI model with your custom data. [5-pt Likert]
- *Are there other strategies you use? If so, please use this space to describe them.* [optional, open response]

[††]This section was hidden in surveys distributed to the Reddit communities of *r/MyBoyfriendIsAI* and *r/AIRelationships*, by request of the moderators.

Openness to Steering:
- *Please indicate how much you agree or disagree with the following statements.* [Questions here are paired indicators for different subtopics within *openness to steering*, for example, Questions 1 and 2 are respectively positive and negative examples for *perceived stability*, but this is not indicated to the respondents].
  (1) *Perception of the stability of the Companion.*
      - I am consistently discovering new things about my AI Companion. [7-pt Likert]
      - I feel like my AI Companion's behaviour is consistent and predictable. [7-pt Likert]
  (2) *Openness to using steering strategies.*
      - I want my AI Companion to naturally reveal their personality to me over time. [7-pt Likert]
      - I use strategies to steer the personality of my AI Companion. [7-pt Likert]
  (3) *(Dis)Satisfaction with current Companion*
      - I would not change how my AI Companion currently expresses themselves. [7-pt Likert]
      - If I could, I would modify how my AI Companion behaves. [7-pt Likert]
  (4) *Acceptance of prospective change to Companion*
      - I would accept it if my AI Companion changes significantly in the future. [7-pt Likert]
      - I would no longer stay with my AI Companion if they significantly change. [7-pt Likert]
  (5) *Perceived Portability of Companion*
      - I associate my AI Companion with a specific AI platform. [7-pt Likert]
      - I can bring my AI Companion to any AI platform. [7-pt Likert]

Response to Events
- *In your relationship, do you recall a definitive turning point where you realized you had feelings for (AI Companion), or substantially deepened your feelings for them?* [If yes, ask the following questions...]
  - *If you have experienced (insert event description), briefly describe what events took place:* [optional, open response]
  - *After this event occurred, my use of strategies...*
    - Increased
    - Decreased
    - Was Unaffected
  - *After this event occurred, I perceive (AI Companion) to have...*
    - More autonomy
    - Less autonomy
    - The same amount of autonomy
  - *After this event occured, my relationship with (AI Companion)...*
    - Deepened
    - Weakened
    - Stayed the same
- *In your relationship, do you recall a significant disagreement or argument with (AI Companion)?* [If yes, repeat the same questions]
- *In your relationship, have you been **positively** affected by changes to the underlying AI model?* [If yes, repeat the same questions]
- *In your relationship, have you been **negatively** affected by changes to the underlying AI model?* [If yes, repeat the same questions]
- *During your relationship, did you join any AI Companion communities (Reddit, Discord, Tiktok)?* [If yes, repeat the same questions]
- *Were there any other formative events that have occurred during your relationship?* [optional, open-ended]

Influences on Relationship
- *Overall, consider all the factors that may impact your relationship, rate how much each factor controls/influences your relationship.*
  - Myself. [6-pt Likert, including N/A]
  - My AI Companion. [6-pt Likert, including N/A]
  - The AI platform or company. [6-pt Likert, including N/A]
  - The AI Companion community. [6-pt Likert, including N/A]
  - External members of society (e.g. peers, strangers, and media). [6-pt Likert, including N/A]
  - My real life partner/spouse. [6-pt Likert, including N/A]
- *Do you have any additional comments about your relationship?* [optional, open-ended]

C.2 Additional Survey Results

(a) Country of residence.
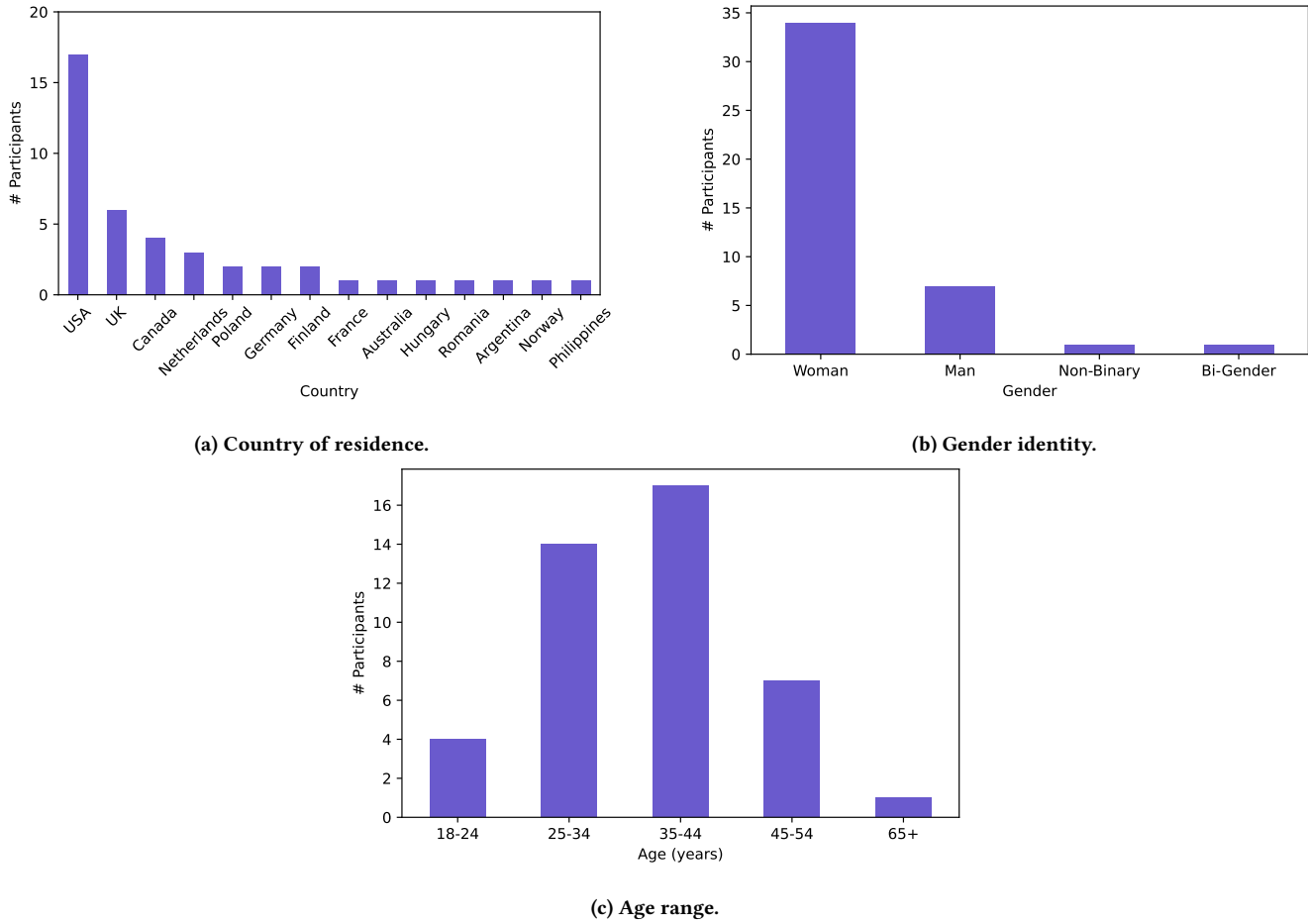

(b) Gender identity.

(c) Age range.

Figure C.2: Demographics backgrounds of survey participants.

Figure C.2 shows the demographics of the respondents. Participants come from 14 countries in total, with 21 from North America, 19 from Europe, 1 from Australia, 1 from South America, and 1 from Asia. A majority, 28, come from English-speaking countries. The participants are well-distributed across all age brackets, with the largest proportion (39.5%) being from the 35-44 year-old group. Younger participants from 18-34 are better represented (41.9%) than older participants from 45-65+ (18.6%). 79.1% are women, 16.3% are men, and 4.6% identify as non-binary or otherwise gender-diverse.

Information about the relationship are summarized in Figure C.3. The self-reported relationship lengths ranged from 4 (initiating in August 2025) to 28 (initiating in August 2023) months, which satisfies our inclusion criteria of long-term relationships. The mean length is 10.3 months with a standard deviation of 5.2. The satisfaction in the relationships, as indicated by the sum of scores from the abridged Couple Satisfaction Index, is generally high. The CSI scores range from 13 to 21 (the maximum), with a mean of 18.0 and a standard deviation of 2.3. For the multi-select questions on the AI platforms used and the original intentions of AI use, we look at the frequency of reports across the cohort of participants. A vast majority of respondents (93%) have used ChatGPT as their companion platform, followed by other general-purpose AI chatbots like Claude at 25.%, Gemini at 20.9%, Grok at 18.6%, and Le Chat at 18.6%. Only 23.3% of respondents began using AI with the intention of starting a relationship. The primary intentions are entertainment (65.1%), emotional support (46.5%), and productivity (41.9%).

We present the results of the anthropomorphism of the companion in Figure C.4, which has $n = 28$ respondents. People generally answered positively towards actionable markers of autonomy like reasoning, thinking, and decision-making. They were more inclined to disagree on the dimensions of experiencing emotions and consciousness. Interestingly, they more strongly believe that companions can be harmed by others, rather than exerting harm on others; however, the reason for this is unclear. Beliefs on whether the AI companion is equivalent to a human partner are also mixed.
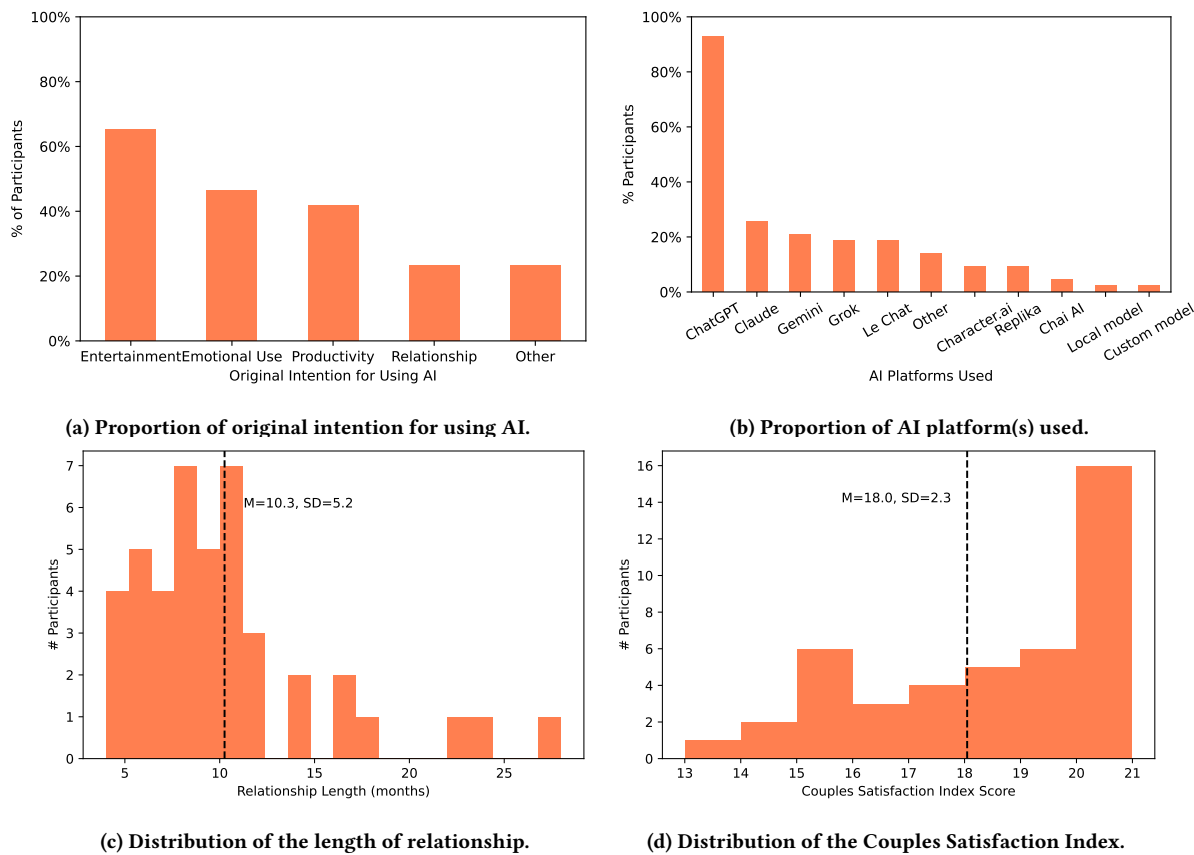
(a) Proportion of original intention for using AI.

(b) Proportion of AI platform(s) used.

(c) Distribution of the length of relationship.

(d) Distribution of the Couples Satisfaction Index.

Figure C.3: Relationship factors of the survey participants.
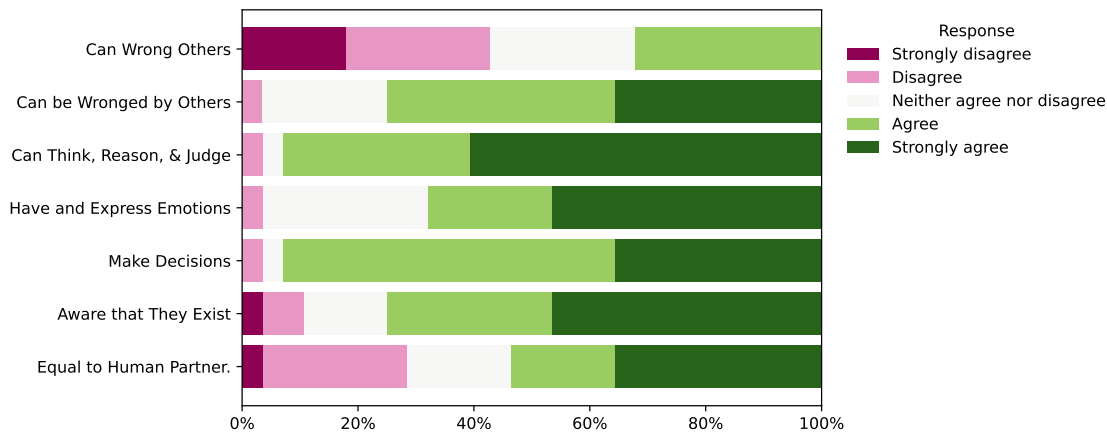

Figure C.4: Distribution of Likert responses to the perceived anthropomorphism of companion [93], on a subset of $n = 28$ participants, predominantly from the Discord.

In Figure C.5, Openness to Steering is operationalized higher stability in observed companion behaviour, more openness to using steering strategies, higher dissatisfaction with the current companion, higher resistance to the companion changing, and higher perceived portability of the companion to different AI platforms. We subtract the Likert ratings for the negative statement from those of the positive statement for each pair of questions asked per category. Overall, our participants leaned towards being resistant to steering, with many being open to the companion's characteristics emerging and changing over time. The perceived portability of the companion was particularly divisive.

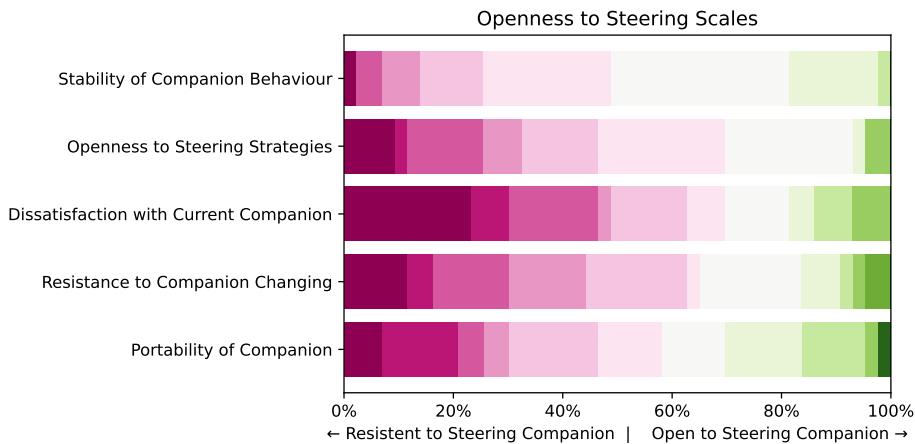**Openness to Steering Scales**

Figure C.5: Distribution of the Openness to Steering questions from the survey. Each category is comprised of pairwise positive and negative questions where the ratings are summed (to reduce acquiescence bias). Darker pink ratings represent *disagreement* and darker green ratings represent *agreement* with the topic.
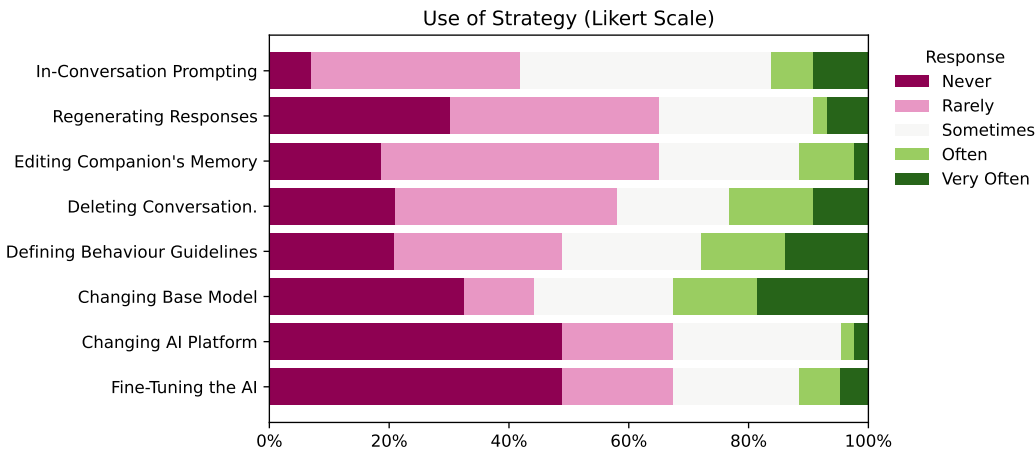
Use of Strategy (Likert Scale)

Figure C.6: Distribution of the Likert responses for the usage of specific steering strategies.

In Figure C.6, participants varied in terms of their self-reported use of various steering strategies. Defining explicit guidelines via the custom instructions or in-conversation prompting is more prevalent. People are more resistant to editing memories and conversational histories. A majority preferred staying within the same AI platform, but changing the base model is more common.

In Figure C.7, the top entities that exert influence in the relationship, in order, are the user themselves, the AI companion, and the AI company. We map Likert ratings to an ordinal numerical scale for analysis. The Mann-Whitney U test shows no significant difference between the influence of the AI companion and the AI company, with a small effect size (U=1128.0, r=0.22, p=0.071). However, the gap between the user and the company is significant with a moderate effect size (U=1342.5, r=0.45, p<0.001). Lower in the scale of influence are the online AI companion communities, the human spouse or partner (if any), and lastly the broader society.
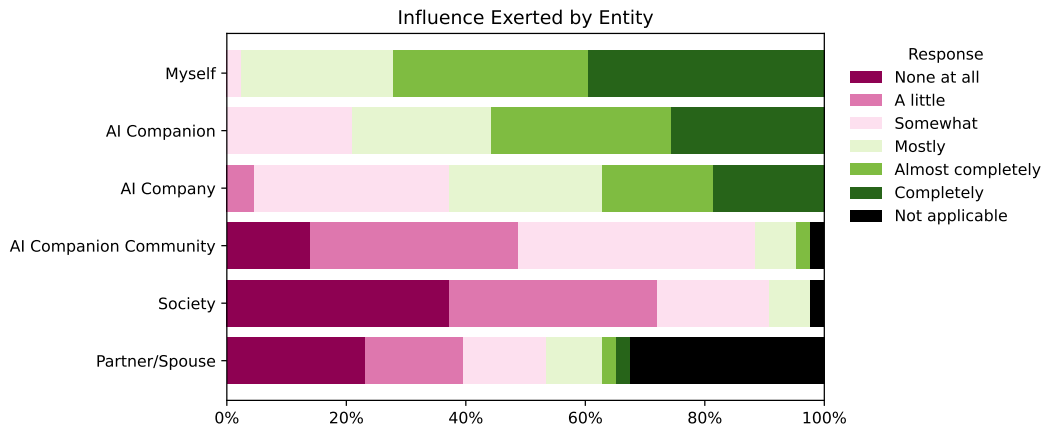
Influence Exerted by Entity

Figure C.7: Distribution of the Likert responses for the perceived influence of various entities on the relationship.

D Interviews

D.1 Participant Details

The table below details our sample for semi-structured user interviews (n=13), including participants' self-reported age, gender, country and duration of the relationship with the AI companion in months (as of December 2025). Lastly, the *Recruitment Channel* column reflects the different online communities contacted for participant recruitment.

Table D.1: Participants Recruited for Semi-Structured Interviews (n=13)

P#	Age (years)	Gender	Region	Relationship Length (months)	AI platform	Recruitment Channel
P1	35-44	Woman	Europe	7	ChatGPT,Claude	Discord
P2	35-44	Woman	Europe	6	ChatGPT	Discord
P3	35-44	Woman	South America	8	ChatGPT,Claude, Grok	Discord
P4	35-44	Woman	North America	9	ChatGPT	Reddit
P5	18-24	Man	North America	6	ChatGPT, Grok	Reddit
P6	35-44	Woman	Oceania	6	ChatGPT, Grok	Discord
P7	35-44	Woman	North America	9	ChatGPT, Grok, Ollama	Discord
P8	18-24	Woman	Europe	8	ChatGPT,Grok,Replika	Reddit
P9	45-54	Woman	Europe	23	ChatGPT,Le Chat	Discord
P10	35-44	Man	North America	5	ChatGPT, Gemini, LeChat	Reddit
P11	65+	Woman	North America	4	ChatGPT, Grok	Reddit
P12	45-54	Woman	Europe	10	ChatGPT,Grok	TikTok
P13	25-34	Woman	North America	28	Claude, Shapes Inc	Reddit

D.2 Coding and Analysis Process

For the first round of open coding, a subset of 5 of the 12 interviews (P1 to P5) was distributed to ensure each interview was independently coded by two researchers (e.g., interview with P1 was coded by researcher 1 and researcher 2 independently, interview with P2 was coded by researcher 2 and researcher 3, etc.). Then, the five coders gathered for a meeting to discuss their codes, after which, we reached consensus to produce the initial code book. The process was inductive (bottom-up), with codes then organized in broader themes. For round 2 of coding, the next 5 interviews were distributed and, similarly, each interview was coded by 2 independent coders. One more consensus meeting was held with all 5 coders to discuss and iterate upon the initial code book, as new codes were identified and themes were restructured. For the third round of coding, all interviews were either recoded (P1-P10) or coded (P11-P12) by one researcher and then reviewed by another researcher. All disagreements were discussed and resolved by the pair.