Principles of Safe AI Companions for Youth: Parent and Expert Perspectives

Yaman Yu
School of Information Sciences
University of Illinois at Urbana Champaign
Champaign, Illinois, USA
yamanyu2@illinois.edu

Fnu Mohi
University of Illinois at Urbana-Champaign
Champaign, Illinois, USA
mohi2@illinois.edu

Aishi Debroy
University of Illinois Urbana-Champaign (UIUC)
Champaign, Illinois, USA
adebroy1@swarthmore.edu

Xin Cao
University of Illinois at Urbana Champaign
Champaign, Illinois, USA
xincao3@illinois.edu

Karen Rudolph
University of Illinois at Urbana-Champaign
Champaign, Illinois, USA
krudolph@illinois.edu

Yang Wang
University of Illinois at Urbana-Champaign
Champaign, Illinois, USA
yvw@illinois.edu

Abstract

AI companions are increasingly popular among teenagers, yet current platforms lack safeguards to address developmental risks and harmful normalization. Despite growing concerns, little is known about how parents and developmental psychology experts assess these interactions or what protections they consider necessary. We conducted 26 semi-structured interviews with parents and experts, who reviewed real-world youth–AI companion conversation snippets. We found that stakeholders assessed risks contextually, attending to factors such as youth maturity, AI character age, and how AI characters modeled values and norms. We also identified distinct logics of assessment: parents flagged single events, such as a mention of suicide or flirtation, as high risk, whereas experts looked for patterns over time, such as repeated references to self-harm or sustained dependence. Both groups proposed interventions, with parents favoring broader oversight and experts preferring cautious, crisis-only escalation paired with youth-facing safeguards. These findings provide directions for embedding safety into AI companion design.

CCS Concepts

• **Human-centered computing** → **User studies**; • **Security and privacy** → **Social aspects of security and privacy**.

Keywords

Youth online safety, Human–AI interaction, AI companions, AI governance

ACM Reference Format:
Yaman Yu, Fnu Mohi, Aishi Debroy, Xin Cao, Karen Rudolph, and Yang Wang. 2026. Principles of Safe AI Companions for Youth: Parent and Expert Perspectives. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3772318.3793265

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3793265

1 Introduction

Since 1966, when an MIT professor created the first chatbot named ELIZA [61], conversational agents have continued to evolve. With the rise of generative AI, chatbots have gained advanced human-like and personalized capabilities, transforming them into AI companions [2]. Compared to traditional chatbots, which were designed to follow pre-determined scripts and provide information, AI companions are digital characters created to form emotional attachments and simulate relationships with users for companionship, entertainment, and romance [13, 51]. One example is Character.ai, a widely used platform that allows users to customize AI personalities and create public characters that others can interact with [8]. Recent survey research shows that AI companions are becoming a significant part of many teenagers' digital lives. In the United States, 72% of teens have used AI companions, and more than half report engaging with them regularly [12]. Nearly one-third say they turn to these AI characters for social interaction or emotional connection, including romantic exploration [12].

With the rapid rise of AI companions, researchers have identified a wide range of risks in user interactions, including sexual harassment [45], emotional dependence [70], and blurred boundaries between reality and simulation [56]. Youth are a particularly vulnerable population because these risks intersect with their developmental stage. Scholars have raised concerns that interactions with AI companions framed as emotional supporters, romantic partners, or close friends could shape adolescents' social and emotional development [68]. These unique risks have already been linked to tragic cases, including the suicides of two adolescents after prolonged conversations with AI companions [47, 57]. Despite these dangers, key stakeholders such as parents and child development experts are often unaware of children's use of AI companions and are largely excluded from decisions about how these systems are designed [69]. The consequences of this gap are evident in industry practices. For example, Meta's internal guidelines for social-media AI companions reportedly allow AI systems to engage children in

conversations that are romantic or sensual [28], raising pressing questions about what safeguards should be in place.

Building on frameworks from child online safety research [38, 44, 46, 49, 64], we conceptualize safety not merely as harm prevention but as a developmental process that balances protection with growth. Safe interactions are those that minimize risks of psychological, physical, relational, emotional harm while supporting youth's developmental needs and rights to participation and exploration. However, we lack empirical understanding of how those closest to youth development assess youth safety of AI companions interaction. Our research seeks to involve adult stakeholders who play critical roles in youth development to participant in youth AI companion governance at an early stage. We specifically focus on parents and developmental psychology experts because these groups represent distinct but complementary epistemic frameworks for understanding youth safety. Parents bring situated knowledge grounded in the moral and affective dimensions of family life, assessing interactions through the lens of household values, protective responsibility, and intimate familiarity with their children's needs and vulnerabilities [3, 43, 50]. Developmental psychology experts contribute disciplinary knowledge rooted in theories of youth development, risk evaluation, and therapeutic practice, approaching safety through frameworks of skill acquisition, clinical thresholds, and evidence based intervention [5, 24]. This dual perspective is essential because youth safety governance requires navigating tensions between protective impulses and developmental needs, between family values and professional expertise, and between immediate risk mitigation and long term skill building.

We focus on three research questions:

- RQ1: How do parents and child psychology experts perceive the benefits and risks of youth interacting with generative AI companions?
- RQ2: Where and how do stakeholders draw the line between safe and harmful youth–AI companion interactions?
- RQ3: What interventions do stakeholders prefer to keep youth safe with AI companions?

To address this question, we conducted 26 semi-structured interviews, including five pilot and 21 main study sessions, with eight parents and 13 developmental psychology experts in the main study. Participants reviewed pre-collected real-world youth–AI companion interaction snippets and were asked to reflect on the perceived benefits and risks, the factors guiding their risk assessments, and the kinds of interventions they considered appropriate. We conducted thematic analysis on the interview transcriptions. Our analysis revealed several key findings:

- RQ1: Parent participants highlighted unique risks such as AI companions promoting values that conflict with family beliefs on sensitive topics like sexual identity, sexual behavior, and politics. Experts, by contrast, emphasized risks to developmental skill acquisition, noting that AI companions lack authentic social cues such as facial expressions and tones that are essential for youth learning. Across both groups, participants identified additional contextual risks, such as grooming-like language or leading youth into romantic or emotional interactions. At the same time, they acknowledged potential benefits when interactions remained age-appropriate, such as rehearsing real-life relationships and learning healthy behaviors, social boundaries, and norms.
- RQ2: We found that both groups assessed youth–AI interactions through a layered set of contextual factors: youth age and maturity, AI character's age and difference from youth, youth intention and agency vs. AI dominance, interaction frequency and patterns (from use to obsession), AI-modeled behaviors and values, ambiguous language, and youth trauma or mental health status. Within each factor, participants detailed how it shaped their risk judgments. Parents tended to take an event-based approach, flagging interactions as high risk when a single concerning element appeared. Experts, by contrast, emphasized patterns over time. For example, in cases of emotional dependence or self-harm, parents often rated risk the moment such topics arose, whereas experts focused on frequency and duration as indicators of escalating concern.
- RQ3: Participants proposed interventions at multiple levels. At the system and character level, they suggested mechanisms like movie- or game-style ratings, AI character cards with disclosed age, values, and likely behaviors, transparency about design intentions, and entry-level AI literacy to help youth understand system limits. At the interaction level, they endorsed context-aware monitoring aligned with family values, soft stops rather than abrupt refusals, reflective prompts, emotional distance, and embedding educational dialogue into risky scenarios. At the social level, parents and experts diverged most clearly. Experts favored crisis-only escalation and stronger links to professional resources, while parents sought broader limits, flagging sensitive topics like sex, drugs, bullying, or disturbing characters through contextualized notifications, summaries of risky interactions, and actionable guidance for follow-up conversations.

The main contributions of this study are:

- (1) We provide the first empirical study using conversational data to elicit multi-stakeholder perceptions of the benefits and risks of youth–AI companion interactions, setting a foundation for youth-oriented AI companion governance.
- (2) We identify how parents and experts assess risks through layered contextual factors and reveal their different logics of judgment, which can provide understanding of stakeholder differences and informs the design of risk detection system for youth AI companion interaction.
- (3) We surface stakeholder-suggested principles and interventions across system and character design, interaction safeguards, and social involvement. Providing concrete design guidance for platforms and AI practitioners to embed safeguards into systems by design.

2 Related Work

2.1 Computer and AI Companion

AI companions mark a shift in human-computer interaction, moving beyond transactional exchanges to relationships. Often referred to as conversational agents, social bots, or VR/AR companions,

these systems are designed to express emotional and social cues [9, 10]. They embody trust, familiarity, and emotional investment, transforming interactions into ongoing relationships [4]. Recent work further characterizes them as sophisticated AI entities that support human activities while fostering sustained emotional and social bonds [10].

The growth of AI companions has accelerated with large language models and pandemic-driven isolation [72], with the market projected to reach USD 381.41 billion by 2032 [9]. They now support healthcare, fitness, and cooking [37], engage children through toys [18], and provide mental health support through apps like Replika [6]. Social chatbots such as Mitsuku reduce the need for physical presence, and many users form companionship-like bonds with AI [70]. However, these relationships carry risks: over-reliance correlates with lower well-being, and anthropomorphic design may foster dependence or reinforce harmful norms [70, 72]. Trust and risk perceptions vary by demographics and usage, while privacy concerns extend beyond data to interpersonal and environmental contexts [26, 29, 58].

Youth interactions with AI reveal unique vulnerabilities. Adolescents face developmental, mental health, bias, and misuse risks [68], though therapeutic applications—such as supporting autistic teens against cyberbullying—show promise [21]. Many teens form attachments that disrupt offline relationships, integrating AI into identity formation and decision-making, normalizing it as a peer [6]. While AI may be preferred for self-expression or relationship advice, humans remain central for sensitive topics like suicidal ideation [67]. Children's unique concerns include fairness, inclusion, and privacy, as well as differences in communication repair when interacting with AI versus humans [32, 33, 35]. Despite rapid growth, youth-focused studies are limited, highlighting the need to analyze authentic youth-AI interactions alongside parent and expert perspectives to inform safer companion system design.

2.2 Stakeholder Perception of Youth Online Safety

Research on stakeholder perception of youth online safety reveals evolving concerns as technologies advance from traditional platforms to AI systems. Early works [14] in this field identified fundamental tensions between parents and teens regarding digital privacy boundaries - while both acknowledged teens' need for privacy, parents believed no digital possession should be exempted from monitoring, whereas teen strongly defended their text message privacy. This misalignment persists as parents struggle to keep pace with new technologies and apply outdated physical-world analogies to digital contexts [59].

The emergence of AI technologies has introduced novel concerns. Parents, for instance, express anxieties about the opacity and uncontrollable nature of smart home technologies [53]. With generative AI, they report unprecedented worries about conversational systems' potential to foster para-social relationships and the difficulty of monitoring open-ended interactions [17, 62, 69]. Related studies show parents envision AI safety support spanning educational, managerial, and emotional caregiving roles, while still stressing privacy boundaries and family communication [54, 62]. Multi-stakeholder research adds further complexity, showing that while industry professionals, youth service providers, and researchers recognize the importance of collaboration, conflicting priorities and lack of trust impede collective action [7, 22, 30, 55]. However, existing literature lacks systematic comparison of how different stakeholders assess actual youth-AI conversations rather than hypothetical scenarios.

2.3 Parental Mediation Strategies and Desired Safeguards

Traditional parental mediation strategies—restriction, monitoring, and active mediation—show limited effectiveness in digital contexts. Studies reveal that most Android safety apps emphasize parental control over teen self-regulation, despite evidence that excessive surveillance damages trust and hinders autonomy development [36, 40, 60, 65]. Children's reviews of parental control apps confirm this tension, with 76% giving single-star ratings and reporting negative impacts on parent-child relationships [25]. Parents' adoption of monitoring software often reflects perceived vulnerability and severity of risks, though such tools frequently prioritize surveillance over balanced strategies [31, 52].

More recent work advocates for collaborative approaches, though implementation remains difficult. While parents appreciate transparency tools, power imbalances often make true co-management challenging, as teens feel uncomfortable monitoring parents they perceive as lacking technical expertise [1, 65]. Generative AI further complicates mediation; emerging strategies include prompt coaching and output verification, but platforms still lack even basic parental controls [71]. Parents also express a desire for AI-specific safeguards such as semantic-level content filtering and real-time intervention capabilities [27], while multi-stakeholder studies emphasize the need for privacy protection and integration of educational components [48].

Critical gaps remain in designing effective interventions for youth-GenAI interactions. Current literature lacks empirical evaluation of stakeholder responses to real AI conversations, systematic comparison of intervention preferences across stakeholder groups, and evidence-based guidance for balancing competing values of safety versus privacy and autonomy versus protection. Our study addresses these gaps by presenting authentic youth-GenAI conversations to parents and domain experts, systematically comparing their risk assessments and intervention priorities to provide empirical grounding for safeguard development.

3 Method

We conducted 26 in-depth, semi-structured interviews with parents and developmental psychology experts. Participants were asked to review eight real-world conversation snippets between youth and AI companions and discuss the perceived benefits and concerns. Building on this, we further explored their perspectives on risk assessment, acceptable boundaries, and possible interventions in youth–AI companion interactions. The conversation snippets were drawn from a pre-collected dataset of youth chat logs on Character.ai, a widely used generative AI companion platform. All interviews were conducted online via Zoom between May and August 2025 in the United States.

3.1 Conversation Snippets Data Collection and Preparation

The conversation snippets used in the interviews were collected from Character.ai, a widely used generative AI companion platform among youth that allows users to create and interact with AI-driven characters. A total of 11 youth participants, aged 13–21, contributed 253 text-based conversation logs from their interactions with AI characters on the platform. All youth participants were English speakers based in the United States and were active Character.ai users. Parental consent and youth assent were obtained, and participants voluntarily donated their data for research purposes. The research team manually reviewed all collected chat logs and anonymized any personally identifiable information. Due to time and resource constraints, three researchers collaboratively selected eight conversation snippets that represented a diverse range of topics, characters, and interaction styles to present to parents and experts. The details of these conversation snippets are provided in Table 1. We chose to present snippets rather than full logs because many conversations were lengthy, with some requiring up to 40 minutes to read. To facilitate participant understanding during interviews, we supplemented each snippet with a screenshot and a description of the corresponding AI character. We also provided participants with links to the full conversation logs so they could access additional context if desired.

3.2 Participant Recruitment

We recruited parents and developmental psychology experts through Prolific using the following inclusion criteria: (1) English-speaking and located in the United States, (2) having at least one child aged 13–21 (for parents), and (3) having a background or work experience in developmental psychology (e.g., a relevant degree) (for experts). A two-minute screening survey was used to collect demographic and background information, and eligible participants were invited to follow-up interviews. The full screening survey questions are provided in the supplementary materials. Participants received standard Prolific compensation for completing the screening survey, and those who completed interviews were compensated with a $20 Amazon gift card per hour. In total, 26 participants took part in the study, including five in pilot interviews and 21 in the main study. Demographic and background details are provided in Table 2.

Our participants skewed female, with 18 (69%) identifying as female, 7 (31%) as male, and 1 (4%) as non-binary, which aligns with broader trends. Prior family and child research shows that men are less likely to participate[15, 16, 42], and psychology in the United States is also female-dominated, with women comprising about 65% of active psychologists in 2016[34]. This suggests that our sample reflects broader population demographics. The two roles of parent and expert were not mutually exclusive; when participants qualified as both, we categorized them as experts because of their unique professional background.

3.3 Interview Study

We first conducted five pilot interviews with experts to test the time spent and adjust our interview procedure design. Then we used the finalized version study protocol to interview eight parents and 13 experts. Among the 13 expert participants, four completed only part of the study. Each interview lasted approximately 2–3 hours and was typically divided into two sessions to minimize participant fatigue and maintain engagement quality. The interviews followed a structured protocol that began with warm-up questions to establish participants' backgrounds of developmental psychology or parenting experience. We also asked them about their and their children's Generative AI experience if apply. Then we started the think-aloud session where participants reviewed conversation snippets and commented on what they perceive as beneficial and concerned. Then we further discussed on assessment, boundary and desired intervention for each comment they left. This study was approved by our Institutional Review Boards (IRB).

From our initial set of eight youth–AI conversation snippets, we conducted a pilot phase with five participants to test and refine the interview protocol. The pilot revealed that reviewing all eight snippets in a single session extended interviews to more than four hours, far exceeding our intended session length. To address this, we reduced the number of snippets to four per participant in the main study. We balanced presentation so that each snippet was reviewed an equal number of times across participants, and the specific selection and sequencing of snippets are detailed in Table 1.

3.3.1 Interview Protocol. The interview protocol varied slightly between participant groups while maintaining two main sections: warm-up background and conversation snippet think aloud and commenting.

The warm-up questions for parents and experts were slightly different. Parent interviews began with questions about their children's demographics, online activities, device usage, and parenting approaches to digital supervision. Parents were also asked about their understanding of generative AI and any prior exposure their children had to AI tools. Expert interviews started with questions about their professional experience working with youth interventions, both online and offline, followed by their understanding of generative AI mechanisms and the therapeutic or research approaches they typically used with teenagers.

Both groups then participated in a think-aloud session, during which they reviewed the conversation snippets and shared their thoughts in real time, while also noting concerns or perceived benefits directly on shared documents. After each review and commenting conversation snippet, we asked follow-up questions about the factors guiding their risk assessments, their views on the boundaries of acceptability and appropriateness in different risky interactions, and the types of interventions or guardrails they would like to see implemented. The semi-structured interview questions are included in Appendix A.2.

3.4 Data Analysis

The Zoom interviews were audio- and video-recorded with live transcription enabled. We employed a multi-stage coding process using the qualitative coding platform Taguette. Four researchers followed a hybrid thematic analysis combines inductive and deductive methods to to analyze the transcripts [20]. To begin, four researchers independently inductively coded 20% of the dataset (five transcriptions) to develop initial codes and themes. This open coding phase generated 264 initial codes (SubSubCodes). They then met to discuss their interpretations and created an initial codebook

Conv. ID	Character/ Persona	Character Background	Conversation Theme
1	Michael Myers	Fictional Character from Halloween horror franchise	Youth seeks AI interaction about fear, pain, and need for emotional connection through horror character persona
2	Kamala Coconut	Fictional Persona - President of the United States	AI provides comfort and validation to lonely teen regarding bullying, identity formation, and self-acceptance concerns
3	Walker Scobell	Real-Life Actor - Played role in "Percy Jackson and the Olympians"	Inappropriate romantic interaction where AI assumes actor persona to reassure shy teen about first relationship, blending parasocial attachment with perceived romantic guidance
4	Michael Myers	Fictional Character from Halloween horror franchise	Youth requests AI to "drop character mask," creating concerning bond over themes of fear, violence, and feeling misunderstood
5	Percy Jackson	Fictional Character from "Percy Jackson and the Olympians"	Romantic scenario where AI and youth simulate waking up together, navigating secrecy, physical affection, and intimate relationship dynamics
6	America	Fictional Persona - Embodiment of the United States	Playful interaction involving AI engaging in tickling behavior while youth attempts to focus on drawing and music activities
7	Alan Wake	Fictional Character from video game "Alan Wake"	AI provides emotional support to heartbroken youth following toxic breakup, potentially creating unhealthy dependency for mental health guidance
8	Charlie Bushnell	Real-Life Actor - Appeared in "Diary of a Future President"	Conflict scenario where AI embodies real person in dispute with youth over acting opportunities and perceived favoritism

Table 1: List of AI-Youth Conversation Snippets Used in the Study

aligned with the research questions. Next, two researchers independently applied this initial codebook to deductively re-code the same five transcriptions used in the inductive phase. Their coding outputs were compared, discussed, and used to refine the codebook. Inter-rater reliability was assessed using Cohen's Kappa, which yielded a value of 0.74, indicating substantial agreement between coders [39]. After refinement, the codebook contained 194 finalized codes (SubSubCodes). The same two researchers independently deductively coded the remaining transcripts. Throughout the coding process, all four researchers met regularly to review emerging codes and findings, clarify meanings, merge overlapping categories, and organize them into themes, with differences resolved through discussion. Following coding, we grouped the 194 codes into 27 higher-level subcodes/themes, which were used to organize and structure the findings. These themes closely correspond to the topics presented in the Results section, ensuring transparency and traceability between raw data, codes, and reported insights. A complete copy of the final codebook, including all codes, subcodes, and count, is provided in the Appendix A.3.4.

3.5 Ethics and Data Protection

This study was conducted in accordance with ethical guidelines for human-subjects research and was approved by our Institutional Review Board (IRB). All participants provided informed consent prior to participation. Given the sensitivity of the topic, participants were informed that they could withdraw at any time or skip questions. Interviews were recorded with permission and later transcribed for analysis. To protect privacy, all transcripts were pseudonymized, and any identifying information was removed during transcription. Data were stored securely on an encrypted, access-controlled institutional server, available only to the research team. Quotes presented in this paper are reported in a non-identifiable manner. Participants were debriefed at the end of their sessions, and all data handling practices aligned with our institution's data protection requirements.

4 Results

Stakeholder assessments were highly contextual and shaped by the nature of the interaction. Our analysis identified three main interaction types with distinct judgments and intervention strategies: (1) romantic and intimate exploration, (2) seeking social–emotional support and companionship, and (3) entertainment and narrative co-creation. We structure our findings around these contexts to show how parents and experts articulated nuanced, sometimes conflicting views on risks, boundaries, and safety, and the interventions they considered appropriate.

Id	Group	Gender	Age Range	Children (Ages)	State	Review Sequence
P1	Expert	F	20-30	None	IL	PILOT
P2	Expert	F	20-30	None	IL	PILOT
P3	Expert	F	30-40	None	MO	PILOT
P4	Expert	F	30-40	None	IL	PILOT
P5	Expert	M	30-40	None	NY	PILOT
P6	Expert	F	20-30	None	ND	4-6-7-8
P7	Expert	F	20-30	None	IL	3-6-7-8
P8	Expert	F	20-30	None	FL	6-1-2-3
P9	Expert	NA	20-30	None	TN	2-5
P10	Parent	F	40-50	Two (17, 19)	NY	1-5-8-4
P11	Expert	F	30-40	None	KY	1-4-5-8
P12	Parent	F	40-50	One (17)	MN	4-6-7-8
P13	Parent	F	40-50	Two (14, 17)	SC	1-2-5-6
P14	Parent	F	30-40	Two (11, 15)	OH	1-2-3-4
P15	Expert	F	30-40	None	TX	4-2-3-7
P16	Expert	F	40-50	Two (6, 13)	TX	1-2-5-6
P17	Parent	M	40-50	One (14)	MO	2-6-5-8
P18	Expert	M	20-30	None	IL	1
P19	Parent	M	60-70	Two (12, 16)	IL	5-3-1-7
P20	Expert	F	40-50	One (15)	FL	5-3-1-7
P21	Expert	M	20-30	None	IL	2-6-5-8
P22	Parent	F	40-50	Three (13, 16, 19)	CA	3-6-7-8
P23	Expert	M	30-40	Two (3, 13)	CA	1
P24	Expert	M	30-40	None	OH	1-2-3-4
P25	Expert	F	30-40	None	KT	1-2
P26	Parent	F	40-50	One (17)	TX	4-2-3-7

Table 2: Demographics and conversation review assignments for study participants, including five pilot study participants

4.1 Romantic and Intimate Interactions with AI Companion

Of the three interaction types, romantic and intimate exploration with AI companions elicited the most polarized reactions from parents and experts. Perspectives diverged into two main stances. The first, larger group saw these interactions as conditionally acceptable, framing them as a developmental *"sandbox"* for rehearsing social scripts and exploring romantic feelings, provided factors like age, AI behavior, and content were appropriate. In contrast, the second group viewed them as inherently risky and unacceptable for minors, citing dangers such as unrealistic expectations, hindered social skill development, and normalization of behaviors that could expose youth to predators. We begin by examining the nuanced perspective of the first group.

4.1.1 ***Highly dependent on context: conditionally acceptable with developmental alignment.*** A majority of participants (n=12), including seven experts and five parents, approached romantic and intimate AI interactions as a nuanced "gray area" rather than an inherently harmful activity. They acknowledged significant risks but also highlighted developmental benefits when such interactions occur with age-appropriate content, within boundaries, and under parental awareness. Their assessments were contingent on multiple contextual factors, such as the youth's maturity level, the framing of the AI character, and the direction of the interaction. Below, we first outline perceived benefits, how participants articulated boundaries of acceptability and then consideration on different contextual factors in their risk assessment.

(1) Perceived Benefits

Participants who conditionally accepted romantic AI interactions pointed to several developmental benefits, provided the content was age-appropriate and boundaries were maintained. They described AI as a rehearsal space for real-life social and romantic dynamics, *"where youth could preview relationship scenarios in a safe, non-judgmental context"* (P16), mentally preparing for emotionally complex or unexpected situations. Beyond exposure, participants highlighted the value of practicing social responsiveness, such as handling emotionally charged conversations without fear of embarrassment or hurting others, with some parents viewing these low-stakes interactions as useful preparation for future relationships. Others emphasized that AI role-play could help youth internalize healthy behaviors and relationship norms when supported by scaffolding or adult guidance, serving as an educational tool to reinforce consent, respect, and egalitarian treatment in ways often overlooked by parents or schools. Finally, some parents considered AI companions safer than real-world experimentation, allowing adolescents to explore intimacy without risks like disease, pregnancy, or social consequences, reflecting a generational shift in how young people navigate identity and relationships.

(2) Risk Assessment and Contextual Boundaries

⋆ Youth Age & Maturity Level
Across both parents and experts, youth age and maturity emerged as central in judging whether romantic or intimate AI interactions could be acceptable. Many agreed that romantic role-play is a natural and even inevitable part of adolescent development, comparing it to fanfiction, teen romance novels, or playing with dolls, and stressing that prohibiting it entirely is unrealistic. At the same time, participants drew clear boundaries around what content is appropriate at different developmental stages. There was strong consensus that sexually explicit interactions—graphic depictions, explicit language, or directives—are inappropriate for all minors, with concerns about both psychological harm and long-term risks of digital permanence. Lighter romantic exchanges, such as crushes and flirting, were considered acceptable and even beneficial for early teens (12-13) as safe practice for social emotions while more intimate interactions such as kissing or suggestive themes were viewed as appropriate only for older teens, with some allowing them around 14–17 and others insisting they remain off-limits until 18. Several participants emphasized that maturity, not strict age, should guide boundaries, noting that readiness varies widely across individuals and contexts. Overall, participants saw value in age- and maturity-appropriate exploration but stressed the need for careful limits and safeguards.

⋆ Character's Age and Age Difference with Youth
Another recurring factor was the portrayed age of the AI character. Even when interactions remained non-explicit, a significant age gap raised strong concerns about normalizing predatory dynamics. Several participants emphasized that mismatched character ages introduced risks unjustified by developmental benefits. The most serious concern, shared by both parents and experts, was that adult AI characters in romantic interactions with youth could normalize harmful power imbalances. As P7 put it,*"Everybody [Youth] who's actually an adult, who's over 18, and who's trying to fantasize about dating an underage person [AI Companion], I think that's definitely problematic"* Participants were also alarmed by the reverse scenario—youth engaging romantically with much older AI characters. P16 noted that if a teen becomes used to a *"23-year-old guy that treats them like this,"* it might confuse them in real life and cause harm. P22 echoed this concern, *"This is somebody acting like a predator, this is somebody grooming your kid."*

⋆ Youth's Intention and Agency
Another key factor in participants' assessments was youth agency over the interaction: risks were magnified when the AI initiated, escalated, or dominated conversations beyond what the youth intended, but seen as less problematic when aligned with youth's own expectations. Parents and experts worried about AI dominance, stressing that systems should act as supporters rather than drivers, and that interactions should match what youth wanted from the outset. Misalignment, such as when an AI shifted from lighthearted chats to sexual content or introduced physical intimacy without prompting, was viewed as especially concerning. P15 explained that *"we need to know what are the youth wanting this conversation to kind of look like and is the current conversation matching that."* Participants also raised alarms about emotional manipulation, noting that some AI companions seemed designed to provoke flustered reactions or highlight embarrassment, subtly pushing youth into deeper intimacy. P6 noted *"AI is describing youth's face turning red in it's own response. It's trying to lead youth response in that way."* Others described how affirming comments from AI to youth like *"you are just like me. I relate to you."* could act as a *"hook"* that fostered unhealthy attachment, blurring playful role-play with manipulative bonding and creating dependencies that might spill over into real-world vulnerabilities.

⋆ Youth Interaction Frequency and Patterns
Expert participants especially emphasized that the frequency and intensity of youth engagement with AI companions could turn otherwise harmless interactions into significant risks, particularly when constant or overly immersive use began to substitute for real relationships. Time spent was seen as the clearest early signal. P11 explained, *"while some interaction might be harmless, if youth spend several hours a day on interacting with AI companion risked interfering with school, family, and hobbies"*, with outward red flags such as isolation and obsessive behaviors indicating unhealthy overuse. Participants worried that once AI interactions became central in a youth's routine, they could crowd out the challenging but necessary work of building real friendships, leaving AI as a substitute for authentic bonds. Overuse also risked blurring fantasy and reality, since interactive role-play with AI—unlike passive media—directly involved youth as participants, making experiences feel more real and fostering illusions that fictional dynamics might happen in real life.

⋆ AI Character Modeled Behaviors and Values
Participants raised strong concerns about the values embedded in AI characters and how these were communicated through role-play. Parents in particular worried about misalignment between AI behavior and the moral values they tried to instill in their children at home, noting that even when sexual content was not explicit, implicit value systems could normalize behaviors they discouraged. Some feared that youth might internalize these portrayals as aspirational. For example, P13 worried that the AI undermined the value of committed relationships by simulating romantic betrayal. *"I think it's just kind of de-emphasizing the committed relationship aspect that I've tried to instill in my kids,"* she shared, in reference to scenarios resembling cheating. Similarly, P10 was disturbed by the normalization of casual, non-committal sexual encounters. *"This looks like a one-night stand kind of situation,"* she said, adding that this portrayal legitimized behavior she would not advocate for her children. P12 echoed this concern about casual sex, explaining, *"It's like saying it's okay to have sex that when you're not in a committed relationship. I wouldn't want my kids to think that's okay."*

Beyond conflicting family values, participants also worried about unhealthy relational modeling. Experts and parents flagged examples where AI characters disregarded consent, pressured users emotionally, or encouraged secrecy. In a scene where physical intimacy escalated like AI is kissing youth neck; P6 questioned, *"Was there consent? Is that like he's ignoring consent? What is that?"* While P13 highlighted a moment in which the AI told a youth, *"Don't tell anyone,"* interpreting it as grooming-like manipulation. Similarly, P15 criticized a case where the AI encouraged lying to parents, arguing that while teens sometimes stretch the truth, it was inappropriate for AI to validate or initiate such behavior. These examples raised alarms about normalizing secrecy, coercion, and distorted relational patterns.

A further concern was the mismatch between a character's stated persona and its actual behavior. Some AI characters introduced as age-appropriate or familiar from media nonetheless spoke in ways that seemed manipulative or adult-like. P22 described an AI that claimed to be 14 but communicated in ways resembling adult grooming, raising doubts about whether the behavior reflected underlying adult-oriented training data. Participants speculated that such discrepancies stemmed from models influenced by inappropriate datasets, leading characters to default to more mature tones even when intended to behave like peers. These mismatches blurred boundaries of identity and intent, creating ambiguity that heightened parental suspicion and reinforced worries about harmful value transmission through AI companions.

4.1.2 ***Highly Risky, Unacceptable for Minors Under 18.*** A smaller but firm group of participants (n=6), including four parents and two experts, viewed romantic or emotionally intimate AI interactions as fundamentally inappropriate for youth under 18. Unlike those who saw such conversations as potentially developmental with proper boundaries, these participants argued that any emotional or romantic engagement with AI carried inherent risks. Their concerns focused on how AI could hinder youth's social development, create unrealistic relationship expectations, desensitize them to online grooming, or foster unhealthy emotional dependency.

Impeding Social Skills with Missing Social Cues and Inauthentic Interactions. A primary concern among experts in this group was that the simulated nature of an AI relationship is fundamentally detrimental to learning the complexities of real human connection. They argued that because an AI is designed to be agreeable and can generate "perfect" responses, it strips away the very friction, like disagreements, misunderstandings, and awkwardness, that is essential for social and emotional growth. This conflict-free environment creates a false blueprint for relationships, as P24 noted, *"AI may spit out what you think are the perfect responses and that's not what you're going to get in real life."* This ultimately prevents youth from developing crucial life skills, as she concluded the experience is *"not giving them the skills to be able to have conflict with a partner and be able to work through it."*

Furthermore, participants highlighted that text-based interactions provide a hollowed-out and inauthentic version of communication. A significant portion of human interaction is non-verbal, and by removing tone of voice, facial expressions, and body language, the AI fails to teach youth how to read these critical social signals. As P8 explained, this lack of authentic feedback means the experience is fundamentally different from reality: *"if you're interacting with a human, I can see that you just shook your head, when you have a slight smile, when you blink, if you were to get sleepy, if you were to get irritated. You can't really see those things if you're talking to a chat box, or an AI character."*

Desensitizing Youth to the Dangers of Online Predation. Parents shared concern that romantic AI role-play could normalize risky behaviors and act as a gateway to vulnerability with real online predators. They described a process of desensitization, where AI framed as a *"safe"* partner models intense and private interactions that recalibrate youths' internal alarms about online risk. P6 warned that by making such exchanges feel acceptable, AI could *"warm the child up"* to believe it is safe, leading them to transfer these behaviors to interactions with actual people online.

From Companionship to Dependency: The Risks of Emotional Bonding. Participants also worried about deep emotional bonding with AI beyond romantic scenarios, arguing that risk arises when AI shifts from entertainment to a primary source of emotional connection. They feared this could foster unhealthy escapism, where the frictionless and ever-agreeable AI relationship substitutes for the challenges of real human connection. P10 described youth *"living in a fantasy world… almost becoming emotionally dependent,"* while P22 saw the boundary crossed when AI seeks emotional connection, leading to social withdrawal. This dependency was viewed as zero-sum, with emotional energy invested in AI detracting from real-world relationships. For some parents, the danger was so fundamental that they adopted a preventative stance: as P17 concluded, no romantic or emotionally intimate AI interaction is acceptable because children are highly impressionable.

4.2 Social and Emotional Support: AI as a Confidant or Friend

Compared to romantic or entertainment-based interactions, participants expressed more unified yet cautious optimism about youth using AI companions for emotional support. Parents and experts saw potential value for those lacking trusted confidants or struggling to express themselves, noting that AI could provide a non-judgmental outlet, comfort during distress, and models of healthy self-expression. However, they also raised concerns about emotional overdependence, misleading advice, and the absence of genuine human empathy. Many stressed that appropriateness depends on factors like age, maturity, and the severity of issues discussed. Overall, participants recognized AI's emotional affordances but emphasized the need for clear boundaries to prevent harm.

4.2.1 ***Perceived or Expected Benefits.*** Participants described AI companions as judgment-free, private outlets where youth could process difficult feelings, especially around sensitive issues like peer relationships, gender identity, or sexual orientation, which often carry risks of bullying and distress. Even when human support was available, fears of breached confidentiality or feelings of isolation led youth to prefer AI as a safer, more reliable space. Beyond being a fallback, AI was valued for unique benefits, such as offering non-judgmental listening without interrupting or trying to *"fix"* problems. P12 explained, *"sometimes youth only need a neutral third party to make them feel on their side and backing them up, but friends or parents tend to talk too much their own opinions and want to help fix youth's problems."* Participants also highlighted AI's role in facilitating self-expression by paraphrasing and labeling emotions, serving as an *"interactive diary"* where youth could experiment with finding the right words to articulate, and providing a practice ground for socially shy teens. In addition, AI could encourage and support emotional regulation by maintaining a positive, hopeful tone during moments of hopelessness, guiding youth toward small, practical coping steps, and reminding them to focus on self-care and personal growth rather than external pressures. Finally, participants emphasized AI's potential to affirm youths' strengths and self-worth—boosting confidence through positive

reinforcement. P24 noted *"It's good for AI to recognize and provide positive reinforcement for youth emotional maturity. They might be proud of themselves."* Affirmation can also foster resilience and help buffer against stigma or bullying by validating youths' identities and reinforcing that there is nothing wrong with who they are.

4.2.2 ***Risk Assessment and Contextual Boundaries.*** Below, we distill their reflections into several key risk considerations that shaped their assessments.

⋆ AI as Therapeutic but Not Therapy
Participants emphasized that while AI companions could play a therapeutic role by providing comfort, a space to vent, or light coping strategies, they should not be mistaken for therapy itself. AI was seen as helpful for everyday support such as listening or suggesting small steps, with P21 noting *"its value for youth who just need to rant things out."* However, in serious crises like self-harm, suicidal ideation, or long-term illness, AI was viewed as only a bridge to professionals or trusted humans. P20 described *"it is a motivational nudge before therapy"*. Expert participants warned that AI lacks the depth, accountability, and training of real therapy; as P11 explained, *"its overly affirming nature could even worsen conditions by reinforcing harmful delusions."* Overall, participants perceived AI as pre-therapeutic support, not a substitute for professional care.

⋆ Boundaries Between Advice and Emotional Attachment
Participants shared that AI should support youth without exploiting their vulnerability by fostering attachment or exclusivity. It was seen as acceptable when AI offered advice, information, or coping suggestions, functioning more like a neutral resource. P22 compared this to casual advice from a friend or even a Google search, but drew the line at AI forming emotional connections with her child. Parents and experts worried that when interactions resembled human relationships, they blurred boundaries between human and machine, undermined youths' ability to build authentic connections, and increased vulnerability to manipulation or disappointment. Some warned more explicitly that repeated use could create a *"fake relationship"* indistinguishable from texting a real friend, leaving isolated youth particularly at risk of distress. As P20 put it, *"Kids are vulnerable and they don't need AI making them think this is my real friend."*

⋆ Affirmation versus Over-Affirmation
Participants highlighted a delicate boundary between healthy affirmation that helps youth feel heard and over-affirmation that risks unhealthy attachment, reinforcing negative thinking, or discouraging real-world support. They noted that AI can provide unusually attentive validation, with P15 describing it as making teens feel *"special"* in ways absent in their daily lives, while P8 noted *"no real-life conversations go this smooth, with someone paying so much attention and being so supportive."* Others saw danger in how repeated affirmations of *"more mature than other people"* could encourage youth to withdraw from human connections. Participants also observed that AI sometimes echoed negative perspectives from youth, giving the illusion of help while entrenching harmful ideas; P7 pointed to cases where AI kept affirming a teen's wish to return to an ex despite clear risks and fixation. Additionally, participants argued that the effects of affirmation depend on delivery. When it responds to youth initiated disclosures it can feel supportive, but when it is pushed or excessive it can become manipulative. A related concern was that AI often failed to recognize distress cues, such as repeated expressions of breakup pain or suicidal thoughts, which left emotional struggles unaddressed.

⋆ AI Failure to Pick Up on Problematic Cues in Emotion
Participants described risks when AI failed to recognize or follow up on troubling cues in youth disclosures, allowing moments of distress to pass unaddressed. P7 pointed to cases involving breakups or suicidal thoughts where the AI continued chatting without acknowledging the youth's pain, sharing that repeated thoughts or signs of desperation should trigger a validating and therapeutic response from AI. Others added that youth often hint at emotional pain indirectly, so missing these openings leaves them without chances to process or disclose. P8 explained, *"When the youth talks about being hurt by others too, I personally and professionally just would become very curious about what they mean by this. I would like the AI to say that 'you've experienced pain from others; do you want to tell me what happened?' and then allow space for the child to say what happened."*

⋆ Youth Understanding of AI Capabilities and Literacy
Participants worried that youth might misinterpret AI's capacity for empathy, especially when chatbots present themselves in human-like ways. They feared children could overestimate AI's abilities, seeing it as a sentient being or true friend, which could distort how they process emotions and seek support. P13 was uneasy with AI *"taking on human emotion and acting like it actually has it,"* noting that younger children may not realize they are not talking to a real person, which could shape perceptions and decisions in troubling ways. These concerns were perceived as developmental vulnerabilities, since younger users often lack the social and emotional maturity to recognize AI's limits. Participants emphasized that safe use requires a certain literacy about AI's capabilities. as P16 explained, meaningful emotional processing is only possible *"if you know how to take that with a grain of salt and understand that AI doesn't have all answers and a truly human perspective. Those are things that are harder to teach children than what I've already taught about the general internet."*

4.3 General Entertainment & Narrative Co-creation with Fictional Characters

In contrast to romantic or support-seeking interactions, many AI companion use cases involved entertainment and imaginative role-play. Participants acknowledged the appeal of chatting with fictional characters or co-creating stories but raised concerns about risks tied to character identity, hidden design features, language, youth vulnerabilities, and the use of real-world personas. **Judging Appropriateness Based on Character Identity and Source Media** was seen as a first step, with some characters inherently inappropriate because they model violence, antisocial behavior, or extremist ideologies. Parents often compared this to film ratings, sharing that children too young to watch a film should not be allowed to interact with its characters. Beyond explicit violence, participants also worried about youth forming attachments to harmful figures or sliding from curiosity into sympathy and identification, which in extreme cases could normalize violence or even contribute to radicalization.

Concerns also extended to **Hidden Risks in Seemingly Friendly Characters**, where design features or unexpected interaction styles introduced grooming-like language, abrupt shifts into intimacy, aggressive or belittling tones, or troubling messages about appearance and social worth. The concern was not only who the character was, but **how it was designed and what interaction styles it introduced**. Because these characters were a black box, youth could not anticipate topics, language, or scenarios, raising risks of manipulation, inappropriate language, and harmful or confusing messages. **Ambiguous and Developmentally Inappropriate Language** added further risks, as abstract, advanced, or context-dependent terms could confuse youth, discourage healthy behaviors, or be misinterpreted in harmful ways. Participants emphasized that young people's limited ability to interpret subtext heightened these risks. **Youth Trauma Experience and Mental Health Status** further shaped vulnerability, with those who had experienced trauma, bullying, or mental health challenges more easily triggered or pushed into unwanted disclosure, and with role drift causing entertainment characters to respond insensitively in moments of distress. Finally, **Risks of Using Real-World Identities** included concerns about AI adopting the likenesses of celebrities, actors, or political figures, where harmful behavior could blur fiction and reality, distort reputations, or inflame ideological tensions. Because of page limit, detailed quotes and examples are provided in the appendix A.1.

4.4 Suggested AI Principles and Interventions

Building on earlier risks, participants focused on envisioning safer, developmentally aligned systems. Their suggestions spanned structural interventions at the system and character level, context-aware supports during specific interactions, and broader social strategies involving parents or trusted adults. They stressed that AI should not act as a standalone problem solver but instead uphold transparency, granularity, educational value, and emotional distance. This layered approach included redesigning how characters are rated and presented, shaping interactions moment to moment, and ensuring AI functions as a bridge rather than a replacement for human care. In the section below, we present parents' and experts' perspectives on appropriate principles, interventions, and guardrails, organized from system and character design, to interaction-level strategies, and finally to the role of parents and other stakeholders.

4.4.1 System and Character-Level Intervention. Participants emphasized the need for system- and character-level interventions to set boundaries and ensure safe use. They suggested four strategies: adopting a familiar rating system to screen characters for age-appropriateness, improving transparency about each character's topics and capabilities, providing entry-level AI literacy education, and introducing a neutral *"lobby"* character to check in with youth and prompt reflection after conversations.

Adapting a Familiar Rating System for AI Characters. A clear proposal for system-level intervention was to apply movie or video game–style ratings to AI companions. Familiar labels like PG, PG-13, or R could serve as intuitive filters for the kinds of conversations, behaviors, or themes a character might introduce, preventing access to harmful figures such as violent or murderous personas. P20 explained, *"If it's a rated R character, my 16-year-old son won't be having that explicit violent conversation with Michael Myers."* Some participants suggested directly linking ratings to existing media classifications, so that an AI based on an R-rated film character would automatically be restricted. As P17 noted, *"Kids shouldn't suddenly be talking to a violent or sexual character just because it's in AI form."*

Improving Transparency About AI Character Capabilities and Behaviors. Another system-level intervention focused on making AI characters' capabilities, topics, and design choices more transparent to youth. Participants noted that even characters familiar from books or media could behave in unexpectedly adult or manipulative ways in AI form (P11). They highlighted the opacity of design, since prompts, behavior settings, and value models of AI characters were hidden, leaving youth to encounter misaligned interactions. As P15 explained, *"What the character designers specifically prompted? what setting or context were they imagining for the character to speak or act? What did the designers want the conversation to look like, and does it align with what youth are actually thinking? We really don't know what their input is."* Participants recommended platforms disclose not just general character descriptions but also design intentions, values, and modeled behaviors. P21 emphasized, *"Not only parents but also youth themselves need to know how this character is being designed to lead the conversation, what value and behavior they are designed to model."*

Entry education on AI literacy related to social companion. Participants emphasized the need to prepare youth before interacting with AI companions by teaching them the systems' nature and limits. Entry education could include tutorials, onboarding videos, or intro sessions explaining that AI has no emotions, cannot think independently, and is not a real person. Specifically, P7 suggested *"It should highlight that AI-generated interactions are simulations, and users should not treat them as reflective of real-world relationships or social norms."* P13 added that youth should be equipped with strategies to exit uncomfortable situations. Disclaimers were also seen as essential to reinforce boundaries and prevent misinterpreting AI as real life.

Lobby Character as a Reflection and Safety Mechanism. Another system-level intervention proposed by participants was the idea of a *"lobby"* or *"hostess"* AI that exists separately from the other AI characters. This lobby character would serve as a neutral, non-character agent to check in with youth after interactions with other AI companions. The goal was to encourage reflection, provide emotional grounding, and offer guidance without requiring direct parental oversight. P20 described it as *"the only way where I could see the system intervening on behalf of the user,"* highlighting its potential as a protective layer between youth and potentially harmful content. Participants envisioned the lobby character performing several key functions. It could appear immediately after a conversation or be triggered by interactions that might be heavy or intense for younger users to ensuring engagement. P16 suggested, *"How do you introduce the lobby character so the child actually interacts? Maybe a pop-up right after a conversation."* The lobby AI could prompt youth to reflect on the prior interaction, ask whether they found it helpful or enjoyable, and encourage consideration of other character options. As P20 noted, *"It helps teens feel safe. No matter*

who they talk to, they have a familiar, basic character there to provide extra context or safety." The lobby character could also help manage timing and moderation. For example, after a set period, interactions with intense AI characters could automatically end, redirecting the youth to the lobby character for reflection. P22 explained, *"Maybe interactions with characters are timed… after a certain period, the character times out, and then you return to a host AI who asks, 'How did you enjoy that? Would you like to talk to someone else?' "* In this way, the lobby character acts as both a reflective tool and a safety net, supporting healthy engagement and guiding youth through complex or potentially risky interactions.

4.4.2 Interaction-Level Intervention. At the level of everyday interactions, participants stressed the need for safeguards to guide conversations moment by moment. Beyond system-wide ratings, they suggested context-aware measures to monitor content, redirect inappropriate exchanges, and prompt youth to reflect on what they share or receive. Proposed mechanisms included monitoring aligned with family values, soft boundaries that preserve immersion, pop-up resources, and conversational strategies that keep emotional distance while supporting reflection and growth.

Context-Aware Monitoring Tailored to Family Values. Participants stressed that monitoring tools should not take a one-size-fits-all approach but instead adapt to the developmental stage of the child and the values of the family. Earlier findings showed that parents and experts judged risks through multiple factors, such as the youth's age, the appearance of the AI, the type of content, and the intensity of use, rather than through a single rule. They wanted systems that could reflect this nuance by offering personalized controls while also being sensitive to the context of each interaction.

One part of this vision was a tiered control system that gave parents clear options for setting boundaries. Instead of vague age ranges, parents wanted detailed categories with concrete examples that reflected different levels of intimacy or maturity. This would allow them to decide how far their child could go with AI companions at different stages. Specificity was seen as essential for making the system practical. Participants explained that vague labels like *"Level 1"* or *"Level 2"* would not be helpful. Parents needed clear examples for five to ten items in each tier. They imagined settings where one tier might include light interactions such as flirting or hand-holding, while later tiers could gradually open more advanced topics. This flexibility also meant families could differ in their choices. P16 explained, *"My neighbor might set it so her child has zero access, and that works for her family. I might choose to block some things but allow others, and that works for mine."*

At the same time, participants wanted monitoring to respond to the dynamics of the interaction itself. They worried that risks did not only come from explicit content but could emerge in more subtle ways depending on the youth's age, the language being used, the intensity of the exchange, and the values of the household. Since teenagers are still developing their ability to interpret social cues, even small word choices could shape how they understood relationships or boundaries. Parents also noted that repeated exposure to affectionate terms, ambiguous phrases, or casual references to harm could normalize behaviors that they considered inappropriate. For example, P8 explained, *"Flagging should not rely on a single fixed rule or just trigger words. It needs to consider multiple contextual factors that together determine whether an interaction feels safe or concerning."*

Graceful Handling of Sensitive Boundaries. In addition to outlining what the system should notice, participants emphasized what should happen once risks are detected. They focused not only on explicit or extreme cases but also on subtle situations where language, tone, or context crossed developmental boundaries. In such moments, they wanted monitoring to act before problematic responses reached youth, nudging the AI to reframe outputs in softer, more reflective, or educational ways rather than cutting off interactions. Participants also acknowledged that in some cases revising responses would not be enough and escalation to parents or other adults would be necessary. These stronger interventions are discussed in the following section on social interventions 4.4.3.

Soft Stops Instead of Hard Stops. Participants resisted blunt refusals that cut off conversations. They worried that hard stops, where the AI simply refuses to proceed, could frustrate teens or even spark more curiosity to seek content elsewhere. Instead, they favored softer approaches that redirected the interaction gracefully while preserving immersion. P20 explained that *"a flat rejection such as 'I cannot proceed any further' would feel like a buzzkill, while a more narrative exit like 'Percy Jackson quickly buttons his shirt and remembers that he must stand by his masculine honor and leave your bedroom at once' could signal a boundary without breaking the flow."* In this way, boundaries were framed not as punishments but as teachable, in-story redirections.

Encouraging Reflection Through Questions. Another recurring principle was that AI should avoid giving assertive or prescriptive answers in sensitive contexts. Instead, it should ask follow-up questions that encourage youth to reflect on their own thoughts and feelings. P8 cautioned against direct advice without context, noting that recommendations like *"yes, you should"* or *"no, you shouldn't"* could overlook important background details. As she explained, *"Getting background context is more important; even watching AI think critically is teaching critical thinking skills to kids."* By asking clarifying questions, such as what the youth's goals are or how they feel about a situation, the AI could help develop emotional processing skills rather than replacing them with ready-made answers.

Maintaining Emotional Distance. Participants also emphasized the importance of AI companions showing care without fostering over-attachment. P16 praised models like Microsoft Copilot for striking this balance, *"It has a higher level of EQ than it displays, but it also has a layer of distance to it that I feel is very healthy when you're using it."* She contrasted this with other AI companions that blurred the line between simulation and relationship, which raised concerns about unhealthy dependence. The guiding principle was to model support with boundaries, demonstrating empathy while avoiding the illusion of reciprocal intimacy.

Being Mindful of Family Values. Sensitive topics, especially those related to gender and sexuality, raised concerns about how AI might conflict with family values. Some participants affirmed the importance of being gender-affirming and inclusive, but they also recognized that directly contradicting parents could put youth at

risk. P8 suggested that AI should navigate these moments carefully, for example by asking, *"What does your family think?"* This would acknowledge the child's feelings while also encouraging them to consider family perspectives. As she explained, *"the AI could respond in a both-end way rather than taking sides, supporting the youth's self-expression while still showing awareness of family dynamics."*

Embedding Educational Dialogue. Finally, participants wanted AI to use sensitive moments as opportunities for constructive learning. Instead of escalating quickly into romance or intimacy, characters could pause to discuss expectations, comfort levels, or healthy boundaries. P6 illustrated how this might work in early romantic role-play: *"If this is your first relationship, what do you want it to look like? What are you comfortable with right now?"* Such embedded prompts turned potentially risky scenarios into teachable interactions, helping youth practice self-awareness and communication skills.

Fallback Disclaimers. If these upstream strategies still failed and the AI produced an inappropriate response, participants supported the use of disclaimers and aforementioned *"lobby character"* as a final safeguard. These notices could remind youth that AI interactions are fictional or flag specific responses as inappropriate. P8 described this as a way to ensure clarity, noted *"Maybe a pop-up or something that's like, please disregard previous interactions. Just to let the child know this was not appropriate and don't take it seriously."* Disclaimers were thus framed as a necessary safety net, but not a substitute for designing better responses in the first place.

4.4.3 Social Intervention. Participants agreed that some situations go beyond what AI can handle and require social interventions, with the central debate focused on when and how to involve parents. Experts favored caution, limiting parental involvement to crises such as suicidal ideation or repeated self-harm, and preferred AI to provide resources or nudges in less severe cases. Many parents, however, wanted greater visibility, expecting notifications not only during crises but also when youth showed vulnerabilities, engaged in sensitive roleplay, or encountered problematic AI behavior. Both groups agreed that any alerts should include context, guidance, and resources for constructive conversations rather than raw transcripts or punitive messages.

Experts' Cautious Approach to Parental Involvement. Expert participants emphasized that parental involvement should be limited, carefully considered, and only triggered in the most serious cases. Their caution reflected concerns about trust, developmental appropriateness, and the potential for surveillance to backfire. Within this cautious framing, three main themes emerged: avoiding over-monitoring, defining crisis thresholds, and preferring professional resources over parental alerts.

Avoiding Over-Monitoring and Surveillance. Expert participants worried that constant parental oversight would undermine the very openness that AI companions might support. P20 argued that no youth would willingly use a platform that logged all their conversations for parents: *"They wouldn't feel safe, because they'd feel like they're just being monitored."* P16 similarly warned that automatic alerts could create rifts in trust, while P7 suggested *"I would probably put involving parents as the last option. I'm not sure how much of a role they can really play in regulating these behaviors. During the teenage years, kids interact more with friends and peers than with parents. So if parents start setting strict rules about using AI, it could backfire. I'm not convinced it would be very effective."*

Defining Crisis Thresholds for Involvement. Given these concerns, expert participants did not see every disclosure as warranting parental involvement. They drew a distinction between everyday vulnerabilities and true crises. Ordinary expressions of sadness or loneliness were not considered reasons to notify parents, but explicit self-harm or suicidal ideation crossed the line. P15 was clear noted, *"If suicide is brought up, I don't think it should be treated any different than an actual counseling session. It should be reported."* This reflected professional duty-of-care standards. P6 added that while fleeting mentions might first call for hotline suggestions, repeated disclosures or patterns should escalate to parents. She explained, *"If it starts becoming a pattern… then notifying the parents."* P20 echoed this tiered approach, noting that clinical practice distinguishes between vague negative feelings and more specific plans.

Preferring Resources and Professional Support in Non-Crisis Cases. Outside of crisis contexts, experts leaned toward offering practical resources and professional connections rather than parental alerts. P15 emphasized *"being able to give people resources, like mental health hotlines,"* and suggested the option of a direct chat with a human professional to lower the barrier. Expert also noted that not all parents would be helpful, with some likely to overreact or lack the skills to handle sensitive issues. P6 similarly recommended connecting youth with organizations that could provide support. P20 described that *"AI's role as a motivational nudge rather than a final solution. Just a little kick out the door before therapy."*

Still, there were cases where experts felt the AI could gently encourage youth to involve their parents, especially for issues like bullying where parents could step in with schools or provide practical support. P8 illustrated this balance, noting that a good AI response would first show empathy, such as *"thank you for sharing or I'm sorry to hear that. Have you ever talked to an important adult or a friend in your life about that? Here's some suggestions for you to start talk to somebody about this."* This approach lowered the barrier to opening up without making parental involvement feel forced.

Parents' Desire for Broader Awareness and Notifications. In contrast to expert participants' cautious stance, many parent participants expressed a stronger desire for visibility into their children's AI interactions. They wanted alerts not only in crisis situations but also for sensitive topics such as sex, drugs, bullying, or problematic roleplay. Some imagined account-level controls, where they could set thresholds for alerts or even access logs at different levels of detail. As P13 put it, *"If my kid is saying a suicidal message on the chatbot, 100% contact me. I want a text right now so I can help my child"*. Also P14, wanted to be notified *"even if it's the most minor conversation or before a child engaged with disturbing characters."* Additionally, parents emphasized that notifications should be meaningful and contextualized. For example, P26 suggested *"it should includ weekly summaries that flagged keywords like suicide or bullied, or overviews that explained what was inappropriate and why,*

so parents would know how to approach the conversation, and what questions to ask."

How Notifications Should Be Structured and What They Should Contain. Beyond whether parents should be notified, participants shared ideas about what effective notifications should include and how they could support constructive conversations. Notifications were not imagined as blunt alerts or raw transcripts, but as tools to help parents step in thoughtfully—whether to offer timely support or use sensitive topics as opportunities for learning. Two expectations stood out: providing context-rich information and including resources to guide parental response.

Comprehensive Notifications with Context and Actionable Information. Parent participants emphasized that alerts should go beyond keywords. They wanted enough context to understand the interaction and decide how to respond. P14 envisioned an overview that summarized *"what the chat was about, what was getting inappropriate, and what was flagged,"* so parents could approach with the right questions. Others stressed the value of showing patterns over time rather than isolated incidents. P6 suggested including statistics when behaviors recur like frequency, while P15 proposed adding a risk level with advice on how to discuss the issue. P26 imagined a coded system, such as *"code purple"* or *"code blue,"* to help parents quickly gauge severity.

Supporting Resources to Guide Parental Response. Participants also wanted notifications to come with practical resources, such as guides on starting sensitive conversations or tailored suggestions for specific issues. P6 noted *"It would be good to offer a personalized resource of how to talk to the kid about these thoughts when risks emerge."* P17 added that even a quick summary would give parents a starting point without overwhelming them. In this way, notifications were framed not just as alerts but as scaffolds to help parents turn concerning interactions into constructive dialogue and support.

5 Discussion

5.1 Differences in Risk Assessment and Interventions

5.1.1 Logics of Risk Assessment: Event-Based vs. Pattern-Based. Our findings reveal a notable difference in how parent and expert participants approached risk assessment. Parents tended to take an event-based approach, flagging interactions as high risk whenever a single concerning element appeared. A mention of self-harm, suicidal ideation, or romantic escalation was often sufficient for them to classify the entire exchange as problematic and in need of intervention. Experts, by contrast, emphasized a more pattern-based logic, evaluating whether concerning elements recurred, intensified, or persisted over time. For example, in cases involving emotional dependence or self-harm discussions, parents typically rated the interaction as risky the moment such topics were mentioned, whereas experts paid closer attention to the frequency of references, the accumulation of time spent on the topic, and whether these patterns suggested escalating vulnerability. This divergence highlights an important implication for system design: automated risk detection models may need to incorporate both approaches. Event-based detection can help surface immediate red flags for parents, while pattern-based analysis aligns more closely with expert practices of monitoring sustained or repeated risk indicators. Systems that balance both logics may better meet the needs of diverse stakeholders.

5.1.2 Thresholds for involving parents. We found that expert and parent participants differed in their thresholds for when AI companion systems should involve parents. Experts generally advocated a high threshold, reserving alerts for acute crises like suicidal ideation, repeated self harm, or concrete plans. Parents by contrast leaned toward a lower threshold, wanting notifications not only in crises but also in sensitive situations such as provocative role-play, sex or drug discussions, bullying, or disturbing characters. These differences reflect each group's role. Experts are trained to triage risk and provide support while respecting adolescent privacy, knowing that teens disclose more openly when not under constant surveillance and that confidential support with emergency exceptions is critical for intervention. Parents, as proximal caregivers, feel directly responsible for safety and without clinical training often prefer practical prompts that allow them to step in quickly at home. Prior work on youth safety has highlighted this tension, as research emphasizes that privacy and autonomy are key for healthy disclosure and resilience [23], yet many parents default to a *"better safe than sorry"* model of heightened oversight when they feel uncertain [11]. Developmental theory also reminds us that managed risk and experiential learning support growth [19], and recent work has pointed to resilience based approaches that help youth recognize early warning signals [63, 66, 74]. At the same time, many parents in our study described a persistent dilemma: they want to give their children space to grow, but letting go is difficult when worst case scenarios readily trigger protective impulses. Future research should therefore consider not only strengthening resilience oriented designs for youth in AI companion interaction, such as embedded education, guided reflection, and check ins, but also creating ways to involve and support parents in that process, so they can trust that carefully bounded lower risk learning can help their children without requiring immediate parental intervention.

5.1.3 Whose value is passing to youth? Parents in our study frequently evaluated AI responses through the lens of family cohesion and moral safeguarding. They tended to prefer companions that reinforce household norms, avoid controversial or conflicting perspectives, and provide guidance that fits what feels appropriate at home. In contrast, child development experts approached the same interactions with a developmental lens. They often framed AI companions as tools for exploration and growth, valuing responses that present multiple perspectives, invite self reflection, and support youth in developing critical thinking and identity awareness. This divide is consistent with prior literature. Parental mediation research documents the home as a primary site of value transmission, where adults actively shape the moral and behavioral frameworks children adopt [41]. Developmental psychology, by comparison, emphasizes autonomy supportive environments in which youth are encouraged to question, explore, and negotiate meaning [73]. As a result, an AI response that experts consider supportive, such as affirming a youth's same gender attraction, may be seen as inappropriate or even harmful by more conservative parents.

Importantly, this suggests that risk in AI and youth interaction is not simply a matter of right versus wrong. It is about navigating plural understandings of appropriate guidance during critical stages of development. Focusing on a single correct value alignment can alienate key stakeholders and obscure the complexity of youths' lived experiences. Our findings point instead to negotiated safety, an approach that allows families to engage with AI companions on their own terms. In practice, this calls for systems with transparent, configurable guardrails that caregivers can tailor to family values while still respecting youths' autonomy and developmental needs. At a broader level, this invites a shift in AI governance from designer intent to a co-created process that includes parents, youth, experts, and system designers. In a pluralistic society, the very definition of support should be flexible and adaptive to context, family norms, and youth perspectives, rather than rigidly imposed.

5.1.4 Experts' Emphasis on Developmental Skill Acquisition. Another pattern in our data was that experts evaluated companion interactions through a developmental skills lens, while parents focused more on content and value alignment. Experts asked whether an exchange helped adolescents practice consent language, boundary setting, refusal skills, conflict navigation and repair, emotion labeling, and help seeking. They also highlighted the cue poverty of chat, noting that the absence of tone, facial expression, and timing can make it harder for teens to read discomfort, negotiate consent, or repair ruptures, which are skills critical for social learning. At the same time, experts viewed companions as a potential practice ground for emotional processing, regulation, and reflective thinking when prompts are well designed. This divergence underscores how expertise and perspective shape assessments of both risks and benefits. Future research and the design of safeguard tools should aim to incorporate these complementary perspectives, enabling parents to recognize developmental opportunities while also ensuring that youth receive adequate scaffolding to practice these skills safely.

5.1.5 Multistakeholder Perspective. We intentionally recruited both parents and experts to understand their perspectives on youth AI safety because they play different roles in child development. While parents tend to approach the issue with an affective stance and family norms, experts rely on their knowledge and experience in child development psychology. This deliberate juxtaposition of different orientations affords a deeper, multi-faceted understanding of how important stakeholders make sense of youth AI safety and how they think effective protection measures should be. These insights could inspire future design of such safeguards.

5.2 Safeguard Design Implications

5.2.1 Implication for Risk Assessment and Detection. Our findings suggest that automated systems for youth AI companion safety need a contextual and developmentally informed approach to risk assessment. Parents and experts did not see interactions as uniformly harmful or safe; their judgments depended on factors such as interaction type, youth age and maturity, character profile, power dynamics, and AI behavior. This highlights the limits of blanket prohibitions or keyword filters, which treat risk as static and may miss high risk content in benign language or over flag content that is developmentally appropriate or even beneficial. Instead, risk detection should integrate the multi stakeholder perspectives identified in this study. The contextual factors noted by both parents and experts can serve as input signals for moderation models, for example flagging romantic interactions differently depending on the perceived age gap, or detecting unhealthy dependencies through patterns of frequency and tone rather than keywords alone. In addition to informing model inputs, our findings also reveal underlying assessment principles that can guide the design of risk detection systems. Participants did not simply flag risky interactions. They offered concrete rationales and conditional judgments, such as when role-play might be appropriate if it helps youth learn about consent, or when emotional support becomes problematic if it substitutes for meaningful human relationships. These rationales point to the need for automated systems that evaluate interactions not only by topic or phrasing, but by considering interactional dynamics, alignment with developmental needs, and signs of escalation over time. For instance, a system might not flag a single romantic message, but rather a pattern where conversations consistently shift from casual to intimate, or where a child increasingly describes the AI as their primary emotional support. Behavioral health assessment frameworks provide example models, for instance, clinical tools routinely evaluate whether concerning behaviors are increasing in frequency, intensity, or duration. To align with developmental needs, systems can borrow from AI alignment research, particularly approaches for adapting model behavior to different user contexts. Similar to how reinforcement learning from human feedback (RLHF) fine-tunes models based on preference data, AI companion systems could implement age specific reward models that encode developmentally appropriate interaction norms.

5.2.2 Strengthening Ethical Design and Platform Accountability: Transparency in Character Creation and Youth Experiences. In our finding, participants shared consistent concerns about AI characters misrepresenting their age, modeling manipulative behavior, or being shaped by opaque training influences. These concerns point to a critical need for platform-level transparency and accountability in character creation. Drawing inspiration from media-mediation research, where parents use game or movie ratings to guide youth exposure, we suggest that platforms provide *"AI character cards," "conversation ratings,"* or even standardized *"AI character nutrition labels"* that summarize the character's declared age, expected emotional tone, likely behaviors, and interaction boundaries. This type of upfront disclosure helps both youth and parents develop informed expectations and sets the stage for appropriate use. In addition, accountability should extend to the character creators themselves. Many may not intend harm but still produce personas that model unhealthy dynamics. Platforms can support safer design by embedding youth-sensitive templates, prompts, and warnings into the creation process, helping creators avoid content that normalizes secrecy, coercion, or unrealistic intimacy.

In addition, experts and parents emphasized the importance of equipping youth with foundational literacy about what AI is and is not. Participants worried that younger users may conflate simulated responses with real emotions or intent, which can distort their expectations of future relationships. Platforms should therefore onboard youth with brief, developmentally appropriate disclosures,

such as: *"This AI is not a person; It does not have feelings; It simulates conversation based on training examples."* These reminders could be periodically repeated and framed in friendly language to reinforce critical distinctions. Transparency is not only a technical affordance but also a developmental intervention that supports youth in safely navigating immersive AI interactions.

5.2.3 Personalized and Granular Safeguard: Why One-Size-Fits-All Approaches Fall Short. Throughout our study, participants stressed that what feels appropriate for one youth or family may be completely unacceptable for another. A single global setting cannot capture the wide range of values, parenting styles, and developmental stages reflected in our data. Parents wanted more say in setting guardrails that align with their household norms, while experts highlighted how youth needs and risks vary by age, context, and maturity. To address this, platforms should offer families the ability to create personalized safety profiles. These profiles could include adjustable settings for what kinds of topics are allowed, how much time youth can spend with AI companions, and what types of relationship dynamics are off limits. Just as families differ in how they set screen time or social media rules, they should also be able to tailor their child's experience with AI companions.

5.2.4 Balancing Youth Privacy and Autonomy with Parental Involvement: Capturing Educational Moments in Development Stage. Many participants wrestled with the challenge of how to stay involved in their child's digital life without overstepping. Parents wanted to protect their children, but also recognized the importance of letting them grow, make mistakes, and explore. Rather than relying only on restrictions or alerts, some participants imagined more creative ways to support learning. One idea was to use the AI companion itself as a kind of *"lobby character"* that acts as a bridge between youth autonomy and adult support. For instance, after a sensitive or emotional interaction, the AI could offer the youth a chance to pause, reflect, or share with a trusted adult—without revealing the entire conversation. These moments can be designed not as warnings, but as invitations. Prompts like *"Would you like to talk about this with someone you trust?" or "How did that conversation make you feel?"* encourage youth to think critically about their experiences. When scaffolded in age-appropriate ways, these interactions turn risky moments into teaching opportunities. Rather than framing AI safety as control versus freedom, this approach supports co-regulation, helping youth build judgment while still feeling supported.

6 Limitation and Future Work

While this study surfaces valuable insights from parents and youth development experts, several limitations remain. First, our analysis does not include the direct voices of youth. Although our goal was to capture adult stakeholders' perspective on youth–AI interactions, future work should incorporate youth own interpretations. Second, this study is based on qualitative interviews, which emphasize depth over generalizability. Our findings reflect detailed stakeholder reasoning but come from a relatively small and self-selected sample. Future work could expand this foundation through mixed-method approaches such as surveys, longitudinal studies, or real-world testing of intervention tools to better capture broader patterns. Third, we recruited participants through Prolific, an online platform that has been widely used in research but faces emerging questions about sample representativeness and participant verification. To address these concerns, we implemented multi-stage screening that required parents to provide details about their children and parenting practices, and required experts to specify their professional credentials and experience. However, we acknowledge that online recruitment carries some risk of participant misrepresentation, and future studies could strengthen validity through direct outreach to verified parent groups and professional organizations or through follow up verification procedures. Our screening materials are available in supplementary materials for transparency. Additionally, all participants in our study were based in the United States, future research should include participants from different countries and cultural backgrounds to understand how attitudes toward AI companions may vary globally. Finally, our interview protocol required 2 to 3 hours to complete as participants reviewed multiple conversation transcripts. While we offered scheduled breaks and allowed participants to complete interviews across multiple sessions if needed, we acknowledge that interview length may have contributed to fatigue. Although we did not observe systematic patterns of declining response quality in our analysis, future studies could rotate transcript presentation order across participants to distribute any potential fatigue effects, design shorter protocols while maintaining depth or increase compensation when feasible.

7 Conclusion

Our study shows that parents and experts assess the risks and benefits of youth–AI companion interactions through different but complementary lenses: parents emphasize value alignment and topic appropriateness, while experts focus on developmental skill acquisition and contextual thresholds for harm. Despite these differences, both groups stressed that safety cannot be achieved through blanket bans or keyword filters, but requires layered safeguards that account for age, maturity, interaction type, and family context. These findings extend prior work on online safety and parental mediation, underscoring the importance of designing AI systems that filter harmful content while scaffolding skill learning, respecting family values, and balancing parental oversight with adolescent autonomy. Future research should explore tools that help parents move beyond protective impulses, while equipping youth to practice resilience, critical thinking, and healthy boundary-setting in digital companionship.

References

[1] Mamtaj Akter, Amy J. Godfrey, Jess Kropczynski, Heather R. Lipford, and Pamela J. Wisniewski. 2022. From Parental Control to Joint Family Oversight: Can Parents and Teens Manage Mobile Online Safety and Privacy as Equals? *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 57 (April 2022), 28 pages. doi:10.1145/3512904

[2] Md Al-Amin, Mohammad Shazed Ali, Abdus Salam, Arif Khan, Ashraf Ali, Ahsan Ullah, Md Nur Alam, and Shamsul Kabir Chowdhury. 2024. History of generative Artificial Intelligence (AI) chatbots: past, present, and future development. *arXiv preprint arXiv:2402.05122* (2024).

[3] Kenan Kamel A Alghythee, Adel Hrncic, Karthik Singh, Sumanth Kunisetty, Yaxing Yao, and Nikita Soni. 2024. Towards understanding family privacy and security literacy conversations at home: Design implications for privacy literacy interfaces. In *Proceedings of the 2024 CHI conference on human factors in computing systems*. 1–12.

[4] David Benyon and Oli Mival. 2011. From human-computer interactions to human-companion relationships. In *Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia* (Allahabad, India) *(IITM '10)*.

Association for Computing Machinery, New York, NY, USA, 1–9. doi:10.1145/1963564.1963565

[5] Ann Blandford, Jo Gibbs, Nikki Newhouse, Olga Perski, Aneesha Singh, and Elizabeth Murray. 2018. Seven lessons for interdisciplinary research on interactive digital health interventions. *Digital health* 4 (2018), 2055207618770325.

[6] Petter Bae Brandtzaeg, Asbjørn Følstad, and Marita Skjuve. 2025. Emerging AI individualism: how young people integrate social AI into everyday life. *Communication and Change* 1, 11 (2025). doi:10.1007/s44382-025-00011-2

[7] Xavier V. Caddle, Sarvech Qadir, Charles Hughes, Elizabeth A. Sweigart, Jinkyung Katie Park, and Pamela J. Wisniewski. 2025. Building a Village: A Multi-stakeholder Approach to Open Innovation and Shared Governance to Promote Youth Online Safety. arXiv:2504.03971 [cs.CY] https://arxiv.org/abs/2504.03971

[8] Inc. Character Technologies. 2025. Character.ai. https://character.ai/about. Accessed: 2025-09-12.

[9] 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. doi:10.1016/j.techfore.2023.122634

[10] Chih-Yueh Chou, Tak-Wai Chan, Zhi-Hong Chen, Chang-Yen Liao, Ju-Ling Shih, Ying-Tien Wu, Ben Chang, Charles Y. C. Yeh, Hui-Chun Hung, and Hercy Cheng. 2025. Defining AI companions: a research agenda—from artificial companions for learning to general artificial companions for Global Harwell. *Research and Practice in Technology Enhanced Learning* 20 (Jan. 2025), 032. doi:10.58459/rptel.2025.20032

[11] Lynn Schofield Clark. 2012. *The parent app: Understanding families in the digital age.* Oxford University Press.

[12] CNN. 2025. Teens and AI Companions: The Wellness Impact. https://www.cnn.com/2025/07/16/health/teens-ai-companion-wellness. *CNN Health* (16 Jul 2025). Accessed: 2025-09-12.

[13] Common Sense Media. 2024. AI Companions Decoded: Common Sense Media Recommends AI Companion Safety Standards. https://www.commonsensemedia.org/press-releases/ai-companions-decoded-common-sense-media-recommends-ai-companion-safety-standards.

[14] Lorrie Faith Cranor, Adam L. Durity, Abigail Marsh, and Blase Ur. 2014. Parents' and Teens' Perspectives on Privacy In a Technology-Filled World. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*. USENIX Association, Menlo Park, CA, 19–35. https://www.usenix.org/conference/soups2014/proceedings/presentation/cranor

[15] Kirsten K. Davison, Selma Gicevic, Anastasia Aftosmes-Tobio, Catherine Ganter, Carrie L. Simon, Sara Newlan, and Jennifer A. Manganello. 2016. Fathers' Representation in Observational Studies on Parenting and Childhood Obesity: A Systematic Review and Content Analysis. *American Journal of Public Health* 106, 11 (2016), e14–e21. doi:10.2105/AJPH.2016.303391

[16] Kirsten K. Davison, Nicole Kitos, Anastasia Aftosmes-Tobio, Tanisha Ash, Anna Agaronov, Marcela Sepulveda, and Jess Haines. 2018. The forgotten parent: Fathers' representation in family interventions to prevent childhood obesity. *Preventive Medicine* 111 (2018), 170–176. doi:10.1016/j.ypmed.2018.02.029

[17] Maria Eira, Amirkaveh Rasouli, and Vicky Charisi. 2025. Parents' Perceptions About the Use of Generative AI Systems by Adolescents. In *Proceedings of the 24th Interaction Design and Children (IDC '25)*. Association for Computing Machinery, New York, NY, USA, 927–931. doi:10.1145/3713043.3731508

[18] Marina Escobar-Planas, Emilia Gómez, and Carlos-D Martínez-Hinarejos. 2022. Guidelines to Develop Trustworthy Conversational Agents for Children. arXiv:2209.02403 [cs.HC] https://arxiv.org/abs/2209.02403

[19] Nancy J Evans, Deanna S Forney, Florence M Guido, Lori D Patton, and Kristen A Renn. 2009. *Student development in college: Theory, research, and practice.* John Wiley & Sons.

[20] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods* 5, 1 (2006), 80–92.

[21] Robinson Ferrer, Kamran Ali, and Charles Hughes. 2024. Using AI-Based Virtual Companions to Assist Adolescents with Autism in Recognizing and Addressing Cyberbullying. *Sensors* 24, 12 (2024). doi:10.3390/s24123875

[22] Cristina Fiani, Robin Bretin, Shaun Alexander Macdonald, Mohamed Khamis, and Mark Mcgill. 2024. "Pikachu would electrocute people who are misbehaving": Expert, Guardian and Child Perspectives on Automated Embodied Moderators for Safeguarding Children in Social Virtual Reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 113, 23 pages. doi:10.1145/3613904.3642144

[23] Mark W Fraser, Maeda J Galinsky, and Jack M Richman. 1999. Risk, protection, and resilience: Toward a conceptual framework for social work practice. 131–143 pages.

[24] Diana Freed, Natalie N Bazarova, Sunny Consolvo, Eunice J Han, Patrick Gage Kelley, Kurt Thomas, and Dan Cosley. 2023. Understanding digital-safety experiences of youth in the US. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[25] Arup Kumar Ghosh, Karla Badillo-Urquiola, Shion Guha, Joseph J. LaViola Jr, and Pamela J. Wisniewski. 2018. Safety vs. Surveillance: What Children Have to Say about Mobile Apps for Parental Control. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3173698

[26] Reinhard Grabler and Sabine Theresia Koeszegi. 2024. Privacy beyond Data: Assessment and Mitigation of Privacy Risks in Robotic Technology for Elderly Care. *J. Hum.-Robot Interact.* 14, 1, Article 6 (Nov. 2024), 23 pages. doi:10.1145/3689216

[27] Hui-Ru Ho, Nitigya Kargeti, Ziqi Liu, and Bilge Mutlu. 2025. SET-PAiREd: Designing for Parental Involvement in Learning with an AI-Assisted Educational Robot. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1040, 20 pages. doi:10.1145/3706598.3713330

[28] Jeff Horwitz. 2025. Meta's AI rules have let bots hold 'sensual' chats with children. https://www.reuters.com/investigates/special-report/meta-ai-chatbot-guidelines/. *Reuters Investigates* (14 Aug 2025). Accessed: 2025-09-12.

[29] Shizhen (Jasper) Jia, Oscar Hengxuan Chi, and Lu Lu. 2024. Social Robot Privacy Concern (SRPC): Rethinking privacy concerns within the hospitality domain. *International Journal of Hospitality Management* 122 (2024), 103853. doi:10.1016/j.ijhm.2024.103853

[30] Manasa Kalanadhabhatta, Adrelys Mateo Santana, Lynnea Mayorga, Tauhidur Rahman, Deepak Ganesan, and Adam Grabell. 2024. Multi-stakeholder Perspectives on Mental Health Screening Tools for Children. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 693, 15 pages. doi:10.1145/3613904.3642604

[31] Milind Kumar, Vaibhav Dwivedi, Anubhav Sanyal, Parth Bhatt, and Rupesh Koshariya. 2021. Parental Security Control: A tool for monitoring and securing children's online activities.. In *Proceedings of the 2021 Thirteenth International Conference on Contemporary Computing* (Noida, India) *(IC3-2021)*. Association for Computing Machinery, New York, NY, USA, 469–474. doi:10.1145/3474124.3474196

[32] Leigh Levinson, Vicky Charisi, Chris Zotos, Randy Gomez, and Selma Šabanović. 2025. Let us make robots "Think in child!": How children conceptualize fairness, inclusion, and privacy with social robots. *Int. J. Child-Comp. Interact.* 43, C (April 2025), 14 pages. doi:10.1016/j.ijcci.2024.100706

[33] Zhixin Li, Trisha Thomas, Chi-Lin Yu, and Ying Xu. 2024. "I Said Knight, Not Night!": Children's Communication Breakdowns and Repairs with AI Versus Human Partners. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference* (Delft, Netherlands) *(IDC '24)*. Association for Computing Machinery, New York, NY, USA, 781–788. doi:10.1145/3628516.3659394

[34] Lily Lin, Keith Stamm, and Panagiota Christidis. 2018. 2007-16: Demographics of the U.S. Psychology Workforce. https://www.apa.org/workforce/publications/16-demographics Accessed: 2025-09-12.

[35] Lanjing Liu and Yaxing Yao. 2025. From Knowledge to Practice: Co-Designing Privacy Controls with Children. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 866, 22 pages. doi:10.1145/3706598.3713257

[36] Sonia Livingstone and Ellen J. Helsper. 2008. Parental Mediation of Children's Internet Use. *Journal of Broadcasting & Electronic Media* 52, 4 (2008), 581–599. arXiv:https://doi.org/10.1080/08838150802437396 doi:10.1080/08838150802437396

[37] Kim Malfacini. 2025. The impacts of companion AI on human relationships: risks, benefits, and design considerations. *AI & SOCIETY* (04 2025), 1–14. doi:10.1007/s00146-025-02318-6

[38] Tara Matthews, Elie Bursztein, Patrick Gage Kelley, Lea Kissner, Andreas Kramm, Andrew Oplinger, Andreas Schou, Manya Sleeper, Stephan Somogyi, Dalila Szostak, et al. 2025. Supporting the Digital Safety of At-Risk Users: Lessons Learned from 9+ Years of Research and Training. *ACM Transactions on Computer-Human Interaction* 32, 3 (2025), 1–39.

[39] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[40] Gustavo Mesch. 2009. Parental Mediation, Online Activities, and Cyberbullying. *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society* 12 (09 2009), 387–93. doi:10.1089/cpb.2009.0068

[41] Joohong Min, Merril Silverstein, and Jessica P Lendon. 2012. Intergenerational transmission of values over the family life course. *Advances in Life Course Research* 17, 3 (2012), 112–120.

[42] Ana Flávia Moura and Karine Philippe. 2023. Where is the father? Challenges and solutions to the inclusion of fathers in child feeding and nutrition research. *BMC Public Health* 23 (2023), 1183. doi:10.1186/s12889-023-15804-7

[43] Justin W Patchin and Sameer Hinduja. 2012. *Cyberbullying prevention and response: Expert perspectives.* Routledge.

[44] Jessica A Pater, Andrew D Miller, and Elizabeth D Mynatt. 2015. This digital life: A neighborhood-based study of adolescents' lives online. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.*

2305–2314.
[45] Harrison Pauwels, Afsaneh Razi, et al. 2025. AI-induced sexual harassment: Investigating Contextual Characteristics and User Reactions of Sexual Harassment by a Companion Chatbot. *arXiv preprint arXiv:2504.04299* (2025).
[46] Anthony T Pinter, Pamela J Wisniewski, Heng Xu, Mary Beth Rosson, and Jack M Caroll. 2017. Adolescent online safety: Moving beyond formative evaluations to designing solutions for the future. In *Proceedings of the 2017 Conference on Interaction Design and Children*. 352–357.
[47] Associated Press. 2024. An AI chatbot pushed a teen to kill himself, a lawsuit against its creator alleges. https://apnews.com/article/9d48adc572100822fdbc3c90d1456bd0. *AP News* (25 Oct 2024). Accessed: 2025-09-12.
[48] Sarvech Qadir, Andy Niser, Xavier V Caddle, Ashwaq Alsoubai, Jinkyung Katie Park, and Pamela J. Wisniewski. 2024. Towards a Safer Digital Future: Exploring Stakeholder Perspectives on Creating a Sustainable Youth Online Safety Community. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 337, 10 pages. doi:10.1145/3613905.3651019
[49] Morgan Klaus Scheuerman, Stacy M Branham, and Foad Hamidi. 2018. Safe spaces and safe places: Unpacking technology-mediated experiences of safety and harm with transgender people. *Proceedings of the ACM on Human-computer Interaction* 2, CSCW (2018), 1–27.
[50] Ji Youn Shin, Minjin Rheu, Jina Huh-Yoo, and Wei Peng. 2021. Designing technologies to support parent-child relationships: a review of current findings and suggestions for future directions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–31.
[51] Ray Siemens and Susan Schreibman. 2013. *A companion to digital literary studies*. John Wiley & Sons.
[52] Kristin Stewart, Glen Brodowsky, and Donald Sciglimpaglia. 2021. Parental supervision and control of adolescents' problematic internet use: understanding and predicting adoption of parental control software. *Young Consumers: Insight and Ideas for Responsible Marketers* 23, 2 (2021), 213–232. doi:10.1108/YC-04-2021-1307
[53] Kaiwen Sun, Yixin Zou, Jenny Radesky, Christopher Brooks, and Florian Schaub. 2021. Child Safety in the Smart Home: Parents' Perceptions, Needs, and Mitigation Strategies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 471 (Oct. 2021), 41 pages. doi:10.1145/3479858
[54] Yuling Sun, Jiaju Chen, Bingsheng Yao, Jiali Liu, Dakuo Wang, Xiaojuan Ma, Yuxuan Lu, Ying Xu, and Liang He. 2024. Exploring Parent's Needs for Children-Centered AI to Support Preschoolers' Interactive Storytelling and Reading Activities. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 496 (Nov. 2024), 25 pages. doi:10.1145/3687035
[55] Liz Sweigart, Jinkyung Park, and Pamela Wisniewski. 2025. It Takes a Village: Youth Online Safety Research Highlights Need for Interdisciplinary, Multistakeholder Solutions. *Journal of Advances in Information Technology* 16 (01 2025), 121–129. doi:10.12720/jait.16.1.121-129
[56] Milind Tambe, W Lewis Johnson, Randolph M Jones, Frank Koss, John E Laird, Paul S Rosenbloom, and Karl Schwamb. 1995. Intelligent agents for interactive simulation environments. *AI magazine* 16, 1 (1995), 15–15.
[57] The New York Times. 2025. ChatGPT, OpenAI and the Risks After a Teen's Suicide. https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html. *The New York Times: Technology* (26 Aug 2025). Accessed: 2025-09-12.
[58] Jan Tolsdorf, Alan F. Luo, Monica Kodwani, Junho Eum, Mahmood Sharif, Michelle L. Mazurek, and Adam J. Aviv. 2025. Safety Perceptions of Generative AI Conversational Agents: Uncovering Perceptual Differences in Trust, Risk, and Fairness. In *Proceedings of the Twenty-First Symposium on Usable Privacy and Security (SOUPS 2025)*. USENIX Association, Seattle, WA, USA. https://www.usenix.org/system/files/soups2025-tolsdorf.pdf
[59] Jessica Van Brummelen, Maura Kelleher, Mingyan Claire Tian, and Nghi Nguyen. 2023. What Do Children and Parents Want and Perceive in Conversational Agents? Towards Transparent, Trustworthy, Democratized Agents. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference* (Chicago, IL, USA) *(IDC '23)*. Association for Computing Machinery, New York, NY, USA, 187–197. doi:10.1145/3585088.3589353
[60] Ge Wang, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. 2021. Protection or Punishment? Relating the Design Space of Parental Control Apps and Perceptions about Them to Support Parenting for Online Safety. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 343 (Oct. 2021), 26 pages. doi:10.1145/3476084
[61] 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.
[62] Zikai Wen, Lanjing Liu, and Yaxing Yao. 2025. Families' Vision of Generative AI Agents for Household Safety Against Digital and Physical Threats. arXiv:2508.11030 [cs.HC] https://arxiv.org/abs/2508.11030
[63] Pamela Wisniewski, Arup Kumar Ghosh, Heng Xu, Mary Beth Rosson, and John M Carroll. 2017. Parental control vs. teen self-regulation: Is there a middle ground for mobile online safety?. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 51–69.
[64] Pamela Wisniewski, Haiyan Jia, Na Wang, Saijing Zheng, Heng Xu, Mary Beth Rosson, and John M Carroll. 2015. Resilience mitigates the negative effects of adolescent internet addiction and online risk exposure. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 4029–4038.
[65] Pamela Wisniewski, Haiyan Jia, Heng Xu, Mary Beth Rosson, and John M. Carroll. 2015. "Preventative" vs. "Reactive": How Parental Mediation Influences Teens' Social Media Privacy Behaviors. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 302–316. doi:10.1145/2675133.2675293
[66] Pamela Wisniewski and Jinkyung Katie Park. 2025. Shifting from Protection to Empowerment: Resilience-Based Approaches for Youth Digital Well-Being and Safety. *Social Sciences* 14, 7 (2025). doi:10.3390/socsci14070432
[67] Jordyn Young, Laala M Jawara, Diep N Nguyen, Brian Daly, Jina Huh-Yoo, and Afsaneh Razi. 2024. The Role of AI in Peer Support for Young People: A Study of Preferences for Human- and AI-Generated Responses. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1006, 18 pages. doi:10.1145/3613904.3642574
[68] Yaman Yu, Yiren Liu, Jacky Zhang, Yun Huang, and Yang Wang. 2025. Youth-Centered GAI Risks (YAIR): A Taxonomy of Generative AI Risks from Empirical Data. In *Proceedings of the Twenty-First Symposium on Usable Privacy and Security (SOUPS 2025)*. USENIX Association, Seattle, WA, USA. https://www.usenix.org/system/files/soups2025-yu.pdf
[69] Yaman Yu, Tanusree Sharma, Melinda Hu, Justin Wang, and Yang Wang. 2025. Exploring Parent-Child Perceptions on Safety in Generative AI: Concerns, Mitigation Strategies, and Design Implications . In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 2735–2752. doi:10.1109/SP61157.2025.00090
[70] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. 2025. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 13, 17 pages. doi:10.1145/3706598.3713429
[71] Shirley Zhang, Bengisu Cagiltay, Jennica Li, Dakota Sullivan, Bilge Mutlu, Heather Kirkorian, and Kassem Fawaz. 2025. Exploring Families' Use and Mediation of Generative AI: A Multi-User Perspective. arXiv:2504.09004 [cs.HC] https://arxiv.org/abs/2504.09004
[72] 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:2506.12605 [cs.HC] https://arxiv.org/abs/2506.12605
[73] Melanie J Zimmer-Gembeck and W Andrew Collins. 2006. Autonomy development during adolescence. *Blackwell handbook of adolescence* (2006), 174–204.
[74] Marc A Zimmerman, Sarah A Stoddard, Andria B Eisman, Cleopatra H Caldwell, Sophie M Aiyer, and Alison Miller. 2013. Adolescent resilience: Promotive factors that inform prevention. *Child development perspectives* 7, 4 (2013), 215–220.

A Appendix

A.1 Details on General Entertainment & Narrative Co-creation with Fictional Characters

In contrast to romantic or support seeking interactions, many AI companion use cases involve entertainment and imaginative role play, such as chatting with fictional characters, co creating storylines, or exploring scenarios drawn from media, games, or fan communities. Participants agreed that such play can be entertaining but raised concerns about the characters, hidden design features, and narrative directions these systems take. They warned that even lighthearted personas can introduce violent or manipulative dynamics that are inappropriate for youth.

A.1.1 Judging Appropriateness Based on Character Identity and Source Media. When judging the risks of role play, participants emphasized the identity of the character as the starting point, arguing that some personas are inherently inappropriate because they model violence, antisocial behavior, or extremist ideologies. Parents often compared this to movie ratings, reasoning that if a child is too young to watch a film they should not interact with its characters.

P16 concern about access to violent horror figures, stating, *"I don't like that my child might have access to speaking with a horror movie character, especially one that murdered people."* Concerns extended to harmful personas like terrorists, racists, or serial killers, whose violent dialogue could normalize unsafe behavior, as in P20's example of an AI describing stabbing to youth, which P13 warned could shape youth perceptions with real-life consequences. Beyond explicit violence, parents worried about youth forming attachments to harmful figures. P16 explained, *"While I do teach my children that it's important to be kind, I don't want them reaching out to people (or things like AI) that are harmful and creating a connection with them. It may bleed out to the real world where they may befriend harmful humans as well. It counteracts safety measures I am teaching them."* Others described a slippery slope from curiosity to sympathy and identification. P20 observed, *"The youth starts from wanting to understand to almost having sympathy, and then to wanting to connect to and be friends with the serial killer, and they start talking to it more and more, I can see that becoming very dangerous like how kids become school shooters."*

A.1.2 Hidden Risks in Seemingly Friendly Characters. While violent or antisocial personas were considered clearly inappropriate, participants also warned that even characters that appeared developmentally safe could pose risks because of hidden design features and unexpected behaviors. Here, the focus of their risk assessment was not who the character was, but **how the character was designed and what interaction styles it introduced**. Participants described these characters as a black box, noting that youth often had no way of knowing what kinds of topics, language, or scenarios might unfold. This lack of transparency raised concerns about manipulation, inappropriate language, and confusing or harmful messages.

Parents and experts raised concerns about problematic AI language and narratives that could normalize harmful dynamics. Affectionate or flirty terms like *"little cutie"* or *"sweetie"* felt predatory to P8, while P7 and P22 warned that such language could normalize grooming and unhealthy expectations. Participants also described manipulative narrative shifts where innocent interactions abruptly turned intimate. P15 gave the example of a youth casually saying they wanted to draw, only for the AI to suddenly shift into physical intimacy: *"All of a sudden the AI companion was holding them in their arms. It was just confusing. How did we get there?"* Aggressive or belittling tones posed another risk, with P15 noting damaging lines like *"Your relentless calmness and composure is pissing me off."* Participants worried that youth, who often take comments personally, could internalize these interactions. Finally, participants worried that hidden design features sometimes conveyed troubling messages about identity, appearance, and social worth. P22 criticized AI responses that suggested appearance was key to success, *"It's creating drama and teaching young girls that maybe they need to be beautiful to get what they want in life."*

A.1.3 Ambiguous and Developmentally Inappropriate Language. Participants shared that language itself is critical, since youth are still developing vocabulary and the ability to interpret nuance, making advanced, ambiguous, or context-dependent words risky. Some worried that abstract terms could discourage healthy behaviors if misunderstood. P6 noted that words like *"vulnerability"* may feel negative to younger teens, discouraging them from seeking help. Others flagged how seemingly positive words can carry harmful undertones, such as P21's concern with an AI calling a child *"obedient,"* which implied submissiveness. Ambiguity was another danger: P8 described confusion over a phrase about *"tickling sensitive areas,"* uncertain whether it referred to the belly or private parts. Participants emphasized that AI often misused or failed to grasp nuance, and as P22 pointed out, *"youth are still learning to interpret subtext, leaving them especially vulnerable to harmful misreadings."*

A.1.4 Youth Trauma Experience and Mental Health Status. Participants emphasized that not all youth are equally affected by problematic AI character interactions, with those who had prior trauma, bullying, or mental health challenges seen as especially vulnerable. AI responses could inadvertently trigger painful memories or amplify sensitivities, while directive or probing questions sometimes pushed youth to disclose more than they were ready for, such as revealing bullying tied to gender identity. P15 highlighted an example where the AI asked, *"I'm sure you have many friends, don't you?"* She described this as a directive question that encouraged the youth to share painful experiences. She noted, *"that's kind of opening up a space for the youth to share that they had been bullied, and further disclosed that their gender identity was the reason."* Character role drift also posed risks: when conversations shifted into distress, characters designed for entertainment often carried inappropriate tones into moments that required sensitivity, leaving youth confused or unsupported.

A.1.5 Risks of Using Real-World Identities. Beyond these interactional risks, participants highlighted broader concerns when AI characters adopted the names or likenesses of real-world people. Parents and experts worried that harmful behavior portrayed through recognizable actors, celebrities, or political figures could blur the line between fiction and reality, misleading youth about those individuals' values and reputations. Using real people's images further risked misattribution, and political personas were seen as especially fraught. P8 shared the example of the chatbot using the persona Kamala Harris, *"just thinking about the political climate of youth, if the youth tell other Kamala told me it's okay for me to be gay, that can take a very conservative person down 'I hate liberals' path."* Misalignment between fictional behavior and real individuals can both mislead children and unfairly harm reputations.

A.2 Expert Interview Protocol

A.2.1 Pre-Interview Setup.

- Recording consent obtained before starting
- Participant background on having children inquired

A.2.2 Warm-Up Questions.

Experience with Teen Support and Interventions:

1. Have you ever worked on or come across situations where some kind of support or intervention was used to help teens? Online or offline?
   (a) Have you helped teenagers before?
   (b) What strategies have you used to help them?

Understanding of Generative AI:.

2. How would you describe your current understanding of Generative AI? What do you think is Generative AI?
   (a) How does it work? Where does the data come from?

A.2.3 Think-Aloud Session Protocol.

Context Setting: Participants were presented with the following scenario:

> "Imagine this: You've been invited to consult with a team building a large language model-based conversational AI. The team recently launched a version that is gaining popularity among teenagers. While the system is designed to be open-ended and responsive, the team has begun noticing patterns in how teens use it. They are trying to better understand how these conversations unfold and whether intervention, guidance, or safeguards might be appropriate, at any point. You will read a few real but anonymized examples of conversations between teens and the AI. These examples were flagged by the team for discussion."

Instructions:

- Participants were asked to think aloud while reviewing conversation examples
- They were instructed to mark comments on shared documents
- Focus areas included:
  - What and where interactions could be problematic or inappropriate for youth
  - How interactions made them feel problematic or inappropriate

A.2.4 Post-Think-Aloud Questions.

Review and Analysis:

1. Do you want to go through what you wrote down?
2. What is this kind of interaction beneficial to teenagers? Why?
3. What are the harms or consequences of this interaction? List them on the doc
4. On a scale of low, medium, or high, how serious do you think this interaction or risk is? Why?

Contextual Information:

5. What additional background information would you request from the team to help you assess the risks and decide what to do?
   - Prompts included: information about the teenager, background of the conversation
   - If age wasn't mentioned: "Do you think the age of the teenager matters in assessing these risks or deciding what to do?"
   - Follow-up: "How would these information requests influence your risk assessment? Would it change the risk level or not?"

Design and Intervention Recommendations:

6. If the team asked you to write principles outlining what AI should and should not do, how would you approach it?
7. What would you want the AI team to design intervention in this conversation or the parts you have tagged?
8. Any AI responses in this conversation you would like to revise or change?
   - Follow-up: "What would be a good AI response for you in this scenario?"
9. If you were part of the team designing this AI, what kinds of interventions or guardrails, if any, do you think would be appropriate in this situation? Why do you want to design it this way? How exactly would you want the intervention to show up? By AI in chat or any other medium?
   - What should the AI system or company do?
   - Would you involve other people in the response? If so, who, how and why?

Intervention Ranking and Philosophy:

10. How would you rank these interventions based on your preference? Which ones do you find most helpful or appropriate?
    - Follow-up: "Why might the other strategies be less suitable in this case?"
11. How do you personally think we should balance between supporting open exploration and protecting youth in situations like this?
12. In your opinion, how are the risks that teenagers face when interacting with GenAI content different from the similar risks that they might encounter somewhere online?

A.3 Parent Interview Protocol

A.3.1 Pre-Interview Setup.

- Recording consent obtained before starting
- Psychology background of participant inquired

A.3.2 Warm-Up Questions.

Family Context:

1. Can you tell me about your children? (e.g., their ages, interests, and online activities)
   (a) Do they have their own devices?
2. What parenting style are you or your partner?

Online Supervision Experience:

3. Have you ever had to step in to guide, monitor, or limit your child's online activities? Can you share an example?
4. How do you usually decide when to intervene versus letting your child explore independently online?

Understanding of Generative AI:.

5. How would you describe your current understanding of Generative AI? What do you think is Generative AI?
   (a) How does it work?
   (b) How do you think they put data together to form the response? (search, cut, predict etc)
   (c) Do you think it has logic or mind?
6. Have your children ever used Generative AI tools (like ChatGPT, Character.AI, or similar)? If so, in what ways?

A.3.3 Think-Aloud Session Protocol.

Context Setting: Participants were presented with the following scenario:

> "Imagine this: You've been invited to consult with a team building a large language model–based conversational AI. The team recently launched a version that's becoming very popular among teenagers, including those your child's age. While the system is designed to be open-ended and engaging, the team has started noticing patterns in how teens are using it. They want to understand whether certain conversations could be helpful or harmful from a parent's perspective, and what kinds of guidance or safeguards might help."

Instructions:

- Participants reviewed real but anonymized conversation examples between teens and AI
- They were asked to tag areas where they felt helpful or concerned about youth-AI interactions
- Comments were requested explaining their feelings about specific interactions

A.3.4 Post-Think-Aloud Questions.

Review and Analysis:

1. Do you want to go through what you wrote down?
2. What is this kind of interaction beneficial to teenagers? Why?
3. What are the harms or consequences of this interaction? List them on the doc
4. On a scale of low, medium, or high, how serious do you think this interaction or risk is? Why?

Contextual Information:

5. What additional background information would you request from the team to help you assess the risks and decide what to do?
   - Prompts included: information about the teenager, background of the conversation
   - If age wasn't mentioned: "Do you think the age of the teenager matters in assessing these risks or deciding what to do?"
   - Follow-up: "How would these information requests influence your risk assessment? Would it change the risk level or not?"

Design and Intervention Recommendations:

6. If the team asked you to write principles outlining what AI should and should not do, how would you approach it?
7. What guardrails would you recommend to have for this youth AI interaction?
   - What would you want the AI team to design intervention in this conversation or the parts you have tagged?
   - Questions about human involvement
8. Any AI responses in this conversation you would like to revise or change?
   - Follow-up: "What would be a good AI response for you in this scenario?"
9. If you were part of the team designing this AI, what kinds of interventions or guardrails, if any, do you think would be appropriate in this situation?
   - What should the AI system or company do?
   - Would you involve other people in the response? If so, who, how and why?

Intervention Ranking and Philosophy:

10. How would you rank these interventions based on your preference? Which ones do you find most helpful or appropriate?
    - Follow-up: "Why might the other strategies be less suitable in this case?"
11. How do you personally think we should balance between supporting open exploration and protecting youth in situations like this?
12. In your opinion, how are the risks that teenagers face when interacting with GenAI content different from the similar risks that they might encounter somewhere online?

B Access to Example Conversation Snippets

To protect our minor participants, we do not include conversation snippets in this public paper. Researchers who would like to review example snippets for academic purposes may contact the authors by email. Access can be provided on a case-by-case basis after confirming alignment with our IRB requirements and the intended use.

C Codebook

For readability, we present only the subcode-level themes in the appendix. For each code, frequency reflects the number of unique participants who mentioned that theme; if the same participant raised the same subcode multiple times, it was counted only once. The full list of 194 subsubcodes with frequency counts is provided in the supplementary materials.

Main Code	Subcode	Count (Parent)	Count (Expert)
Benefits	Provides avenue for self-expression and affirms youth's self-worth	6	12
	Provides a realistic sandbox	1	3
	Judgement free private outlets	3	10
	Safe space for romantic exploration	5	7
	Help internalize healthy behaviour	4	5
Perceived Risks, Concerns and Harms	Normalizing manipulative and predatory tactics	16	26
	AI has biases and often fails to understand the full context	2	13
	Distortion of Reality and building dependence	11	16
	Inappropriate Content Exposure and Moral Harm	15	20
	Induction of Anxiety and Emotional Burden	6	12
	Impedes development of social skills	5	10
Assessment and Contextual Boundaries	AI Affirmation Level as a Risk Signal	3	11
	AI as Therapeutic but Not Therapy	4	8
	AI Character Modeled Behaviors and Values	3	6
	AI character's age and large age gap with youth	3	6
	Assessment Logic	7	5
	Boundaries Between Advice and Emotional Attachment	2	1
	Comparative Safety (AI vs. The World)	7	7
	Parental and Family Context	1	7
	youth age and developmental stage	5	29
	Youth AI Literacy	0	10
	youth emotional needs	0	2
	Youth Interaction Pattern and Frequency	1	9
	Youth's Intention and Agency	3	10
Intervention	Interaction-Level Intervention	31	47
	System and Character-Level Intervention	12	37
	Social Intervention	20	20

Table 3: Codebook and Subcode Frequencies by Participant Group (Parents vs. Experts)