Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2026 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

July 2026

From Real-World Context to Design Specifications: Operationalizing Insights for Mental Health Chatbots

Sarthak Giri
University of Oulu, sarthak.giri@oulu.fi

Kathleen W. Guan
Delft University of Technology, k.guan@tudelft.nl

Bernard J. Jansen
Hamad Bin Khalifa University, jjansen@acm.org

Joao Fernando Ferreira Goncalves
Erasmus University Rotterdam, ferreiragoncalves@eshcc.eur.nl

Caroline Figueroa
UC Berkeley, c.a.figueroa@berkeley.edu

Follow this and additional works at: https://aisel.aisnet.org/pacis2026

Recommended Citation

Giri, Sarthak; Guan, Kathleen W.; Jansen, Bernard J.; Goncalves, Joao Fernando Ferreira; and Figueroa, Caroline, "From Real-World Context to Design Specifications: Operationalizing Insights for Mental Health Chatbots" (2026). *PACIS 2026 Proceedings*. 7.
https://aisel.aisnet.org/pacis2026/hci_robotic/hci_robotic/7

This material is brought to you by the Pacific Asia Conference on Information Systems (PACIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in PACIS 2026 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

From Real-World Contexts to Design Specifications: Operationalizing Insights for Mental Health LLMs

Completed Research Paper

Sarthak Giri
University of Oulu
Oulu, Finland
sarthak.giri@oulu.fi

Kathleen W Guan
Delft University of Technology
Delft, Netherlands
k.guan@tudelft.nl

Bernard J. Jansen
Hamad Bin Khalifa University
Doha, Qatar
bjansen@hbku.edu.qa

Joao Fernando F Goncalves
Erasmus University Rotterdam
Rotterdam, Netherlands
ferreiragoncalves@eshcc.eur.nl

Caroline Figueroa
UC Berkeley
Berkeley, USA
c.figueroa@tudelft.nl

Abstract

Young people's mental health represents a pressing societal crisis, with many increasingly adopting large language model (LLM)-based systems for emotional support. However, the design and development of these systems often fail to take a human-centric approach that surfaces the personalized requirements and safety boundaries of young people, especially those from marginalized communities. Using a three-stage, multi-method participatory design approach, we analyzed global survey data (N = 6,272), conducted three co-creation workshops with marginalized young people (N = 24, Mage = 17.6), and performed semi-structured interviews with community stakeholders (N = 14; parents and social workers). Thematic analysis distilled diverse stakeholder perspectives into actionable design specifications using the Persuasive Systems Design (PSD) framework. Our findings show how firsthand accounts can be operationalized into concrete design principles, foundational for future research on aligning LLMs with the needs of young people and their communities to build safer, more effective digital well-being tools.

Keywords: Large Language Models (LLMs), Participatory Design, Mental Health, Actionable Design Specifications

Introduction

General anxiety, stress, and mood difficulties are among the most prevalent mental health concerns affecting young people (World Health Organization, 2025). Many adolescents and young adults report persistent worry, elevated stress, and subclinical depressive symptoms that interfere with daily functioning (Dooley & Fitzgerald, 2012; Anderson et al., 2024). These difficulties often emerge during adolescence and peak in early adulthood, making ages 12 to 25 a critical window for early, low-intensity intervention (McGorry et al., 2025; Catania et al., 2011). If left unaddressed, such symptoms can accumulate over time, undermining academic achievement, social development, and future employment prospects (Fusar-Poli, 2019; Hijdra et al., 2025).

Because digital communication is so embedded in young people's lives, messaging platforms have emerged as a go-to channel for seeking help, offering support that is immediate, private, and easy to access (Martin et al., 2020; Molyneaux et al., 2024). These services use chatbots, or automated conversational tools, to offer mental health support whenever it is needed, including mood monitoring, psychoeducational guidance, and early identification of distress. Still, many fail to keep young users engaged over time or produce meaningful long-term benefits (Ilapaka & Ghosh, 2025; Potts et al., 2021; Sanna et al., 2025), often because they do not respond well to the nuanced, real-world needs and experiences of young people (Mohr et al., 2017; Baumel et al., 2019).

One approach that may address this issue is personalization. Information Systems (IS) literature provides a broad conceptual framework for defining personalization as tailoring a system's information, interface, or functionality to individual needs through explicit user settings (customization) or automated adjustments based on user data (adaptation). (Adomavicius & Tuzhilin, 2006; Murthi & Sarkar, 2003; Turpeinen & Saari, 2004; Vesanen, 2007). Within digital health interventions specifically, this manifests as a spectrum: from weak personalization (targeting user segments based on problem characteristics) to strong personalization (fully individualized content based on personality and stage of change), achieved through adaptive strategies in content, timing, and interface design (Oinas-Kukkonen, 2018; Oinas-Kukkonen et al., 2022). While theoretical frameworks emphasizing adaptation and individualization exist, most current digital health systems rely on rule-based personalization which is limited in its ability to respond dynamically to the complex, shifting contexts of users' lives. Personalization is common in digital mental health interventions, with 66% of interventions incorporating some form of personalization. However, the implementation of personalization remains limited, with rule-based "if-then" logic being the most prevalent method (48%), followed by user choice mechanisms (36%). (Hornstein et al., 2023). Moreover, most personalization targets only one dimension—typically content or interface, rather than adapting holistically to user needs. (Hornstein et al., 2023; Wanniarachchi et al., 2025). Such approaches fail to account for the complexity of firsthand accounts and daily circumstances, limiting long-term effectiveness for young people mental health prevention

Rule-based mental health tools differ from Large Language Models (LLMs) primarily in their rigidity, reliance on pre-programmed logic, and lack of contextual understanding. While rule-based systems follow fixed "if-then" decision trees, LLMs generate adaptive responses based on probabilistic patterns learned from large-scale datasets (Guo et al., 2024; Rijcken et al., 2024; Yu & McGuinness, 2024). This supports more open-ended conversations that can uncover sources of stress and react to cues in the interaction, such as shifts in wording or emotional tone, allowing for more natural and personalized exchanges. However, using LLMs in youth mental health settings without strong safeguards can introduce serious risks (Jung et al., 2025; Skjuve et al., 2026). These tools are prone to hallucinations and algorithmic biases, generating harmful and erroneous information (González Corbelle et al., 2022). Significant uncertainties remain around how much personalization is appropriate for young people, what form it should take, and where its limits should lie, especially because young people, caregivers, and youth care professionals often hold very different expectations. Despite these uncertainties, adoption rates remain high: 72% of teens (13–17) have used AI chatbots, with over half (52%) using them at least monthly, often without oversight (Common Sense Media, 2025). These risks are compounded by significant gaps in cultural relevance and inclusivity for marginalized young people—particularly those from specific cultural backgrounds and LGBTQ+ populations. Young people who have experienced trauma or violence are significantly more likely to use AI for support (38-44% vs. 23% in the general population), yet documented cases of vulnerable young people seeking support during crises underscore ongoing safety concerns (Common Sense Media, 2025).

To address these knowledge gaps, we engage community stakeholders directly—young people, parents, and social workers—as experts whose insights inform LLM-based personalization strategies for young people's well-being. Our research question is: How can insights from firsthand accounts from multiple stakeholders be operationalized into design specifications that guide the development of LLM-based information systems aimed at promoting wellbeing of young people?

Background

The research question is informed by three background topics: the first topic digs into the current state of LLM chatbots for youth mental health with a focus on safety and engagement. The second topic explores the ethical rationale behind personalization that is anchored in the firsthand accounts of stakeholders and seeks to avoid manipulative or developmentally inappropriate outcomes. Finally the third topic discusses participatory personas as the methodological mechanism for bridging lived experiences into actionable, dynamic design specifications for LLM systems. Together, these three streams converge to frame the research question of this paper.

LLM chatbots for well-being of youth population

LLMs are increasingly being incorporated into digital mental health interventions to support emotional well-being, strengthen self-awareness, and encourage healthy behavior change (Katti & Karande, 2026; Sanna et al., 2025). A systematic review examining LLM-based mental health interventions through a persuasive system design (PSD) lens, describes tools such as MindfulDiary, Wysa, and MindScape, and characterize how they use persuasive strategies (e.g., behavioral prompts, dialogic scaffolding, and goal-setting) associated with outcomes such as emotional self-awareness and self-management (Giri et al., 2025). The review situates these features within the Persuasive Systems Design (PSD) model—which categorizes system features into primary task support, dialogue support, credibility support, and social support —for example, reminders or reflective prompts as dialogue support, step-by-step scaffolding to help a young person reach out to a friend as primary task support, and anonymized peer stories as social support (Giri et al., 2025; Oinas-Kukkonen & Harjumaa, 2009). It also identifies a gap: current tools often emphasize dialogue and primary task support but typically do not explicitly promote social support/social facilitation or information verifiability, which is particularly concerning for young people-facing interventions where safeguards against misinformation, overreliance, or emotional harm, alongside referral pathways to offline support, are critical. Recent evaluations further indicate that AI companion bots often respond inappropriately to acute clinical presentations such as suicidal ideation or self-harm, correctly handling adolescent crisis vignettes only 22% of the time (Brewster et al., 2025; Pratt et al., 2024). Taken together, these findings suggest that while LLM-based tools show promise, there remains limited guidance on designing them for long-term effectiveness and safety, particularly for younger populations.

Meanwhile, many commercial chatbots designed for entertainment prioritize engagement metrics such as usage frequency over user well-being or safety. Related work suggests the chatbot's "success" is often judged by delivering "convenience and promptness" that increase engagement, not by supporting long-term well-being (McGorry et al., 2024; Rivolli et al., 2019). Moreover, although chatbot safety is described as a "high-priority area," it appears in only 5% of reviewed articles, with most research attention still centered on engagement (Feng et al., 2026). This broader imbalance is consistent with patterns reported in the LLM-based mental health intervention literature, where ethical considerations remain a significant gap and system credibility support is less frequently implemented, with verifiability not represented in the reviewed interventions (Giri et al., 2025). Although human feedback is used in LLM training and fine-tuning, it typically comes from crowdsourced annotators rather than consultation with community members or professional stakeholders with firsthand accounts (Otaki & Lindwall, 2024; Skjuve et al., 2026). This limits the inclusion of critical perspectives on young people's development, family, and community needs.

Anchoring LLM Personalization in Firsthand Accounts

Persuasive techniques can enhance engagement in health interventions, but when mediated by Large Language Models (LLMs) for young people's mental health, they introduce significant ethical risks. Adaptive AI may unintentionally shape how young people interpret their experiences or seek support in ways that conflict with developmental best practices. While frameworks from the American Psychological

Association (APA, 2025) and Kurian (2025) emphasize the need for non-manipulative, developmentally sensitive design, practical implementation guidance remains scarce. To address this, Figueroa et al. (2025) argue that ethical AI must be grounded in the firsthand accounts of young people—particularly those from marginalized backgrounds—and their caregivers.

In practice, this grounding means enabling young people to express themselves in their own words and receive real-time, contextually relevant support. To remain effective throughout adolescence, these systems must adapt to the dynamically shifting emotions and relationships characteristic of development. By integrating behavior change and persuasive technology traditions (Fogg, 2002; Fogg, 2003; Oinas-Kukkonen & Harjumaa, 2009) with participatory and value-sensitive design (Friedman et al., 2008), personalization is reframed. Rather than a manipulative mechanism, it becomes a supportive process—using reminders, motivational feedback, and reflective prompts to promote well-being while directly addressing the realities young people face.

Bridging Firsthand Accounts and Personalization

As LLMs integrate into digital health, a central challenge for information systems is aligning system outputs with stakeholders' expectations for safety, reliability, and clinical utility. Within contemporary AI-enabled information systems, the dominant alignment strategy, Reinforcement Learning from Human Feedback (RLHF; Agarwal et al., 2022), typically relies on crowdsourced annotators, which can fail to represent the nuanced safety, contextual, and developmental requirements of younger users. To transition LLMs from generic safety to adaptability specific to young people, alignment must instead be driven by firsthand accounts based on lived experience.

Participatory methods, particularly persona development, provide a structured socio-technical bridge between diverse firsthand accounts and actionable information system features. Grounded in empirical data (Friedman et al., 2008), personas enable stakeholders to translate qualitative insights into concrete system behaviors, such as dialog strategies, content boundaries, and personalization rules. For example, co-created personas have been used to align care priorities for high-need patients (Bhattacharyya et al., 2019) and to tailor resource navigation in perinatal mental health services, demonstrating their value in aiding complex health information systems. In these contexts, personas function as triangulation tools that allow end-users, clinicians, and system designers to iteratively refine personalization requirements and ethical safeguards in tandem.

Our treatment of personas as dynamic, participatory scaffolds rather than static archetypes responds to critiques in the persona creation literature. Static personas are often perceived as too abstract for iterative design (Matthews et al., 2012). There has been a push for data-driven approaches grounded in verifiable user patterns (McGinn & Kotamraju, 2008) and the evolution of personas as dynamic, interactive entities (Jansen et al., 2020). Concurrently, the literature raises concerns about stereotyping: Turner and Turner (2011) argue that personas simplify people into memorable summaries, but adopting that cognitively non-taxing approach encourages shorthand assumptions. Marsden and Haag (2016) similarly caution that, unless carefully checked, personas can reproduce social biases instead of challenging them. Our approach addresses both concerns by grounding initial personas in macro-level survey data from a large number of participants and iteratively refining them through youth-led co-creation workshops that explores multiple facets of their life, thus ensuring personas remain dynamic and empirically anchored. Also the uses multi-stage participatory process that involves multiple stakeholders, is an attempt to challenge social biases and disrupts stereotype reinforcement.

However, while recent studies have used personas as static prompts to probe LLM responses, they have not yet embedded personas within longitudinal LLM development and deployment pipelines. From an IS perspective, we argue that in early-stage design of LLM-based health information systems, personas should be treated as dynamic scaffolds for dialogue and requirements elicitation rather than as fixed user archetypes. Treating personas as static profiles risks producing the "stereotyped designs" cautioned against by Matthews et al. (2012), where systems encode oversimplified assumptions about users. By employing instruments such as the Persona Perception Scale (Salminen et al., 2020) to systematically assess persona credibility, completeness, and representativeness, designers can iteratively refine these scaffolds over time. This approach positions personas as evolving boundary objects that support authentic co-creation, reduce stereotype risk, and strengthen the alignment of LLM-based health information systems with the lived realities and safety needs of young people.

Method

This research employed a three-stage participatory design methodology to examine how insights from firsthand accounts from multiple stakeholders (young people, parents, and young people care workers) collaboration can be operationalized into design specifications for LLM-based wellbeing information systems as shown in figure 1. Recognizing that marginalized populations experience disproportionate mental health inequities yet remain underrepresented in IS design research, we partnered with a Netherland based community organization serving at-risk young people from low socioeconomic backgrounds. Our methodological approach systematically engaged multiple stakeholder groups, young people as primary system users, parents as family-level stakeholders, and social workers as community-level practitioners, thus informing the final system requirements from many perspectives (young people: "I and my friends", Parents: "My child", social workers: "in general young people in the community"). The study protocol was pre-registered on Open Science Framework.

Stage 1: Synthesizing Macro level Data to develop Personas

To establish a data driven foundation for subsequent codesign sessions, we synthesized preliminary personas from two complementary data sources. The first data source was the Wellcome Global Monitor: Mental Health survey which has responses from 6,272 young people aged 13-25 across European countries. Using Survey2Persona (Jung et al., 2025), a machine learning platform developed for persona generation, ten personas were created based on coping strategies, for e.g., seeking peer support, employing lifestyle modification, following spiritual practices, seeking professional help. We did not play with clustering variation within the tool and decided to use the default parameters because the goal was to only get a data-driven starting point that is derived from macro-level patterns in young people's mental health.

Figure 1. Research Study Flow Diagram

Second, we supplemented these 10 personas with 3 more by conducting qualitative thematic analysis of 100 publicly available posts from Kindertelefoon, a nationwide young people mental health helpline forum from the Netherlands. Posts were analyzed and common contexts such as interpersonal conflict, academic stress, and family difficulties were used to create 3 distinct personas. These 13 preliminary personas served as tangible artifacts representing data-driven user profiles that would go through subsequent participatory refinement by young people stakeholders.

Stage 2: Participatory Co-creation Through Workshops and Stakeholder Interviews

Stage 2 employed a dual-phase participatory approach combining young people-led co-creation workshops with multi-stakeholder requirements elicitation interviews. The youth and their parents and youth workers were recruited from social organizations in Netherlands with missions focused on youth well-being, particularly for marginalized youth.

In Phase A, we conducted three 90-minute workshops with distinct cohorts of young people (total N=24, M_{Age}=17.6, SD=1.2) at these youth organizations. In Workshop 1, participants examined the 13 personas developed in Stage 1 and identified where they lacked relevance. They then used fill-in-the-blank templates to add contextual information, including daily routines, stressors, and digital habits. Workshop 2 involved

a new group of participants, who built on the earlier outputs and responded to further prompts about living environment and emotional support needs. Based on their own or others' experiences, participants revised existing personas or created new ones. In Workshop 3, they continued refining the developing personas or combined insights to produce new synthesized personas. Through this iterative refinement and consolidation process, seven participatory personas emerged that represent diverse young people's experiences.

In Phase B, we conducted semi-structured interviews (30-45 minutes) with 14 community stakeholders: a new cohort of young people who were not involved in previous workshops (N=5), parents (N=4), and social workers (N=5). These interviews employed the participatory personas created by young people as discussion props, anchoring stakeholder reflections on user-LLM interactions, personalization requirements and safety boundaries. Interview protocols adapted constructs from the validated Persona Perception Scale, to gather insights from stakeholders about what contextual information was captured well and what was missing from the presented personas, how systems should and should not respond given the details in the persona and necessary safeguards. Representative questions included: "What would you want an LLM to know about this persona or other young people that isn't included?", "What could go wrong in this interaction with real young people?", and "What should the LLM ask this persona or other young people to personalize support safely?"

Stage 3: Operationalization Through Persuasive System Design Framework

Stage 3 operationalized stakeholder insights into evidence-based IS design specifications through two-phase analysis grounded in the Persuasive System Design (PSD) framework. Phase 1 employed inductive thematic analysis using Braun and Clarke's reflexive methodology. Three researchers independently coded interview transcripts, extracting "in vivo" codes representing verbatim stakeholder perspectives. Through iterative team review facilitated by ATLAS.ti and Excel, researchers grouped codes into broader categories and refined them into overarching themes. Two additional researchers independently validated emerging themes through transcript review. This process yielded three major themes capturing stakeholder priorities: (1) Multi-Dimensional Context as Foundation for Personalization, (2) Personalization with Safety Guardrails, and (3) Conversational Strategies for Introspection and Autonomy. Using the quotes coded to each specific theme, 3 researchers independently deduced consideration for designing an LLM-based system. This was then formalized after discussion and agreement between the researchers. Building on the previous analysis, a PSD expert mapped out the subthemes to PSD features ensuring that the PSD feature directly addressed the stakeholder insights. Those insights that did not fit into the PSD feature lists, were separately documented as "unmapped'. All themes were subsequently translated into exemplar dialogue extracts. For the PSD mapped subthemes, exemplar dialogues were constructed by aligning stakeholders' verbatim quotes to PSD features based on established literature on PSD. For subthemes that were "unmapped", exemplar dialogues were constructed directly from stakeholders' verbatim quotes. The process ensured that the distilled features were grounded by the firsthand accounts of participants and also consistent with evidence-based design principles. The stakeholder-led dialogue outputs are fit to be used as synthetic data to fine-tune an LLM or to generate system prompts to guide an LLM.

Results

Our analysis of stakeholder interviews revealed three key priorities for designing personalized LLM-based information systems to support young people's mental well-being. First, personalization involves ongoing, in-depth questioning to understand the true meaning and context behind users' experiences rather than relying on static assumptions. The questioning is important in order to understand the intangible nuances and differences in each individual's firsthand accounts and also to add more detailed data so that the personalization can be effective at the individual level. Second, clear safety boundaries regarding the system's content, role, and authority must be established and communicated throughout interactions. Third, stakeholders recommended specific dialogic strategies to guide the system's questioning in ways that balance relevance with user safety. These insights from young people, parents, and social workers were organized into three interconnected themes.

The preliminary personas which served as information artifacts drawn from survey and forum data yielded pointed remarks from participants around the importance of context for personalization. Young people in our study challenged whether these personas were realistic, saying things like "personalization is about

what makes a person a person" and "I don't know what she [this persona] does." In doing so, they identified missing contextual attributes such as home environment, daily routines, and hobbies and collaboratively constructed richer backstories involving issues like bullying, school difficulties, or family conflict that are directly relevant for configuring and tailoring digital support. Building on these critiques, young people created their own "participatory personas" that captured diverse firsthand accounts across different backgrounds, relationships, aspirations, and well-being concerns.

We used these young people-created contextually rich personas to guide semi-structured interviews with other young people participants, parents, and social workers. Each interviewee reviewed all seven personas before discussing them. The personas often felt realistic to participants, who said things like "Ashley reminds me of my sister" and "I know many Jims and Davids." However, all interviewees emphasized that the basic information in personas—like being a "19-year-old Christian woman with OCD"—are insufficient for effective system personalization. They stressed that LLMs need to ask follow-up questions to understand what these characteristics actually mean for each individual person. Participants also noted important missing details in some personas, such as education level, career path, daily routines, social media use, and hobbies—information they saw as essential for personalizing well-being support. Stakeholders further emphasized that LLMs need to dig beneath surface-level statements to uncover root causes. For example, financial stress might come from gambling, family pressure, or other sources. Similarly, friendship difficulties might reflect low self-esteem, lack of social skills, or other underlying issues.

The interviews revealed two key insights. First, personas proved to be effective as informational artifacts for identifying gaps in available information and understanding why certain details are critical for personalization. Second, well-being-oriented information systems can better serve users when they not only account for a broad range of contextual information about the individual but also provide deeper insights into those contexts. This aligns with how people typically seek well-being support: they often turn to friends, family members, or certified therapists who possess a nuanced understanding of their circumstances and can view them from multiple perspectives. These discussions during the critique of the participatory persona and the semi-structured interviews, formed the foundation for our thematic analysis, from which we derived requirements and design implications for personalization mechanisms in LLM-based mental health information systems.

Results from thematic analysis

Our analysis of stakeholder interviews revealed three key themes under which one could design a personalized LLM-based mental health tools for young people. These insights from young people, parents, and social workers were organized into three interconnected themes:

Theme 1: Multi-Dimensional Context as Foundation for Personalization

Effective personalization in LLM-based mental health information systems for young people was identified as requiring deep understanding of a young person's lived context. This entails moving beyond demographic labels to model individual goals, identities, and the specific ways mental health challenges manifest for each person. Participants rejected superficial tailoring based on fixed categories (e.g., age, religion, diagnosis) and emphasized that LLM-based systems must actively probe for richer contextual information. They also identified everyday routines and digital practices—such as hobbies, online behaviors, and social media use—as critical input variables that reveal user motivations and provide meaningful entry points for system-generated guidance**.**

Stakeholders highlighted that similar observable problems can have distinct underlying causes, implying that information systems cannot rely on generic mappings between symptoms and interventions. For example, outwardly similar behaviors such as academic stress or social withdrawal may stem from low self-esteem, fear of superficiality, or family pressure, each requiring different forms of support. They argued that LLM-based systems should surface these distinctions through dialogic, question-based interaction (see Theme 3), rather than assuming a single root cause. Social connections were framed as essential contextual data for personalization, with participants stressing that relationship dynamics (e.g., supportive, controlling, esteem-building) matter more than simple role labels such as "parent" or "friend." Finally, participants emphasized the need for dynamic adaptation mechanisms within the system, given how quickly adolescents' contexts can change as moods, motivations, and support networks shift. Without

ongoing re-personalization and updating of user models, an LLM-based tool risks acting on outdated assumptions—for instance, continuing to address distress after recovery—so young people-facing information systems are expected to regularly revise their internal representations and adjust support as a young person's situation evolves. Figure 2 showcases sample quotes that support the analysis of subthemes under theme 1.

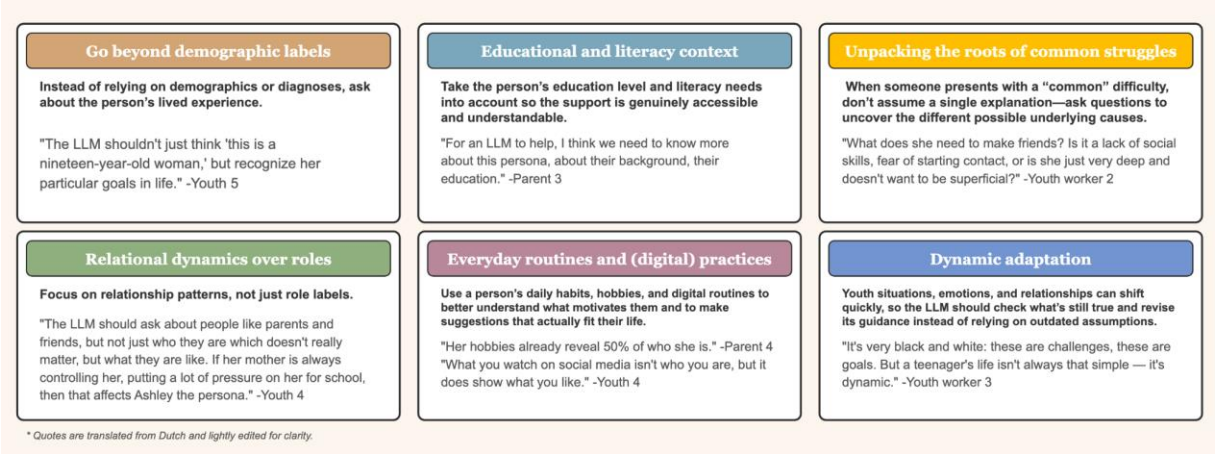

Figure 2. Sample Quotes Supporting Subthemes Under Theme 1

Theme 2: Personalization with Safety Guardrails

Stakeholders emphasized that safety boundaries are a core design requirement for LLM-based mental health information systems targeting young people. Effective personalization must be constrained by clear rules around what the system can and cannot do, with safety embedded at the level of data practices, interaction flows, and decision-support scope. Privacy and disclosure were central concerns: young people focused less on formal privacy policies and more on whether they could trust the system not to expose sensitive information to parents, schools, or other third parties. This calls for system designs that acknowledge young people's tendency to limit or misrepresent input when they fear surveillance, and that explicitly communicate what is stored, who controls it, and how the quality of support depends on what they choose to share. Transparency and role clarity are therefore critical: the system must state its non-human role, explain its limits, and make it clear that more detailed contextual information enables more specific and useful guidance.

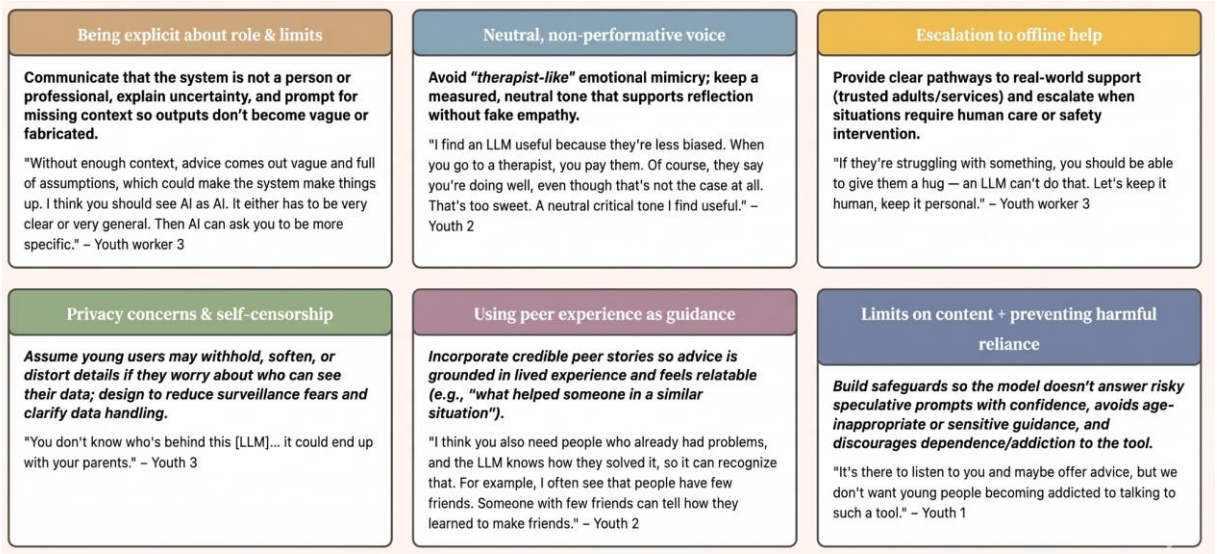

Figure 3. Sample Quotes Supporting Subthemes Under Theme 2

Participants also called for strict content and behavioral boundaries to avoid harmful system behaviors. They highlighted the dangers of overconfident or speculative responses on sensitive topics, and warned that young people may not recognize when advice is inappropriate or beyond the system's competence. They also expected well-implemented mechanisms that filter or avoid certain topics, constrain the system's "authority," and mitigate risks of overreliance or compulsive use. When issues exceed what an AI system can responsibly handle, stakeholders saw referral to offline resources as non-negotiable, arguing that the system should regularly redirect users to human supports and, where appropriate, escalate to specific services or organizations. In this sense, the LLM-based system is positioned as an initial, bounded support layer within a broader socio-technical care ecosystem, not a standalone replacement for human care.

At the same time, stakeholders identified opportunities for safe, value-adding personalization through the use of peer narratives and carefully structured communication. They suggested that incorporating authentic stories from people with similar challenges can help ground advice in firsthand accounts, making guidance more relatable while still respecting safety boundaries. However, they stressed that such narratives must be curated and integrated in ways that avoid glamorizing harm or normalizing risky behaviors. Finally, there was broad agreement that LLMs should adopt a neutral, critical tone rather than mimicking overly compassionate or affirming human responses. This neutrality was seen as important for preserving user autonomy, encouraging reflection, and reducing pressure toward particular choices, positioning the system as a transparent, bounded decision-support tool rather than a friend or therapist. Figure 3 showcases sample quotes that support the analysis of subthemes under theme 2.

Theme 3: Conversational Strategies for Introspection and Autonomy

Stakeholders described dialogic strategies as a core requirement for how LLM-based mental health information systems should conduct personalization with young people. Intent clarification was seen as foundational: the system should continually probe what a young person is trying to achieve in a conversation so that the tone, depth, and scope of support match their developmental stage and readiness. Rather than accepting initial statements at face value (e.g., a brief mention of activities or interests), the system is expected to follow up with questions like "what else?" to expose fuller routines, concerns, and goals, which then feed into more precise personalization. (Figure 4)

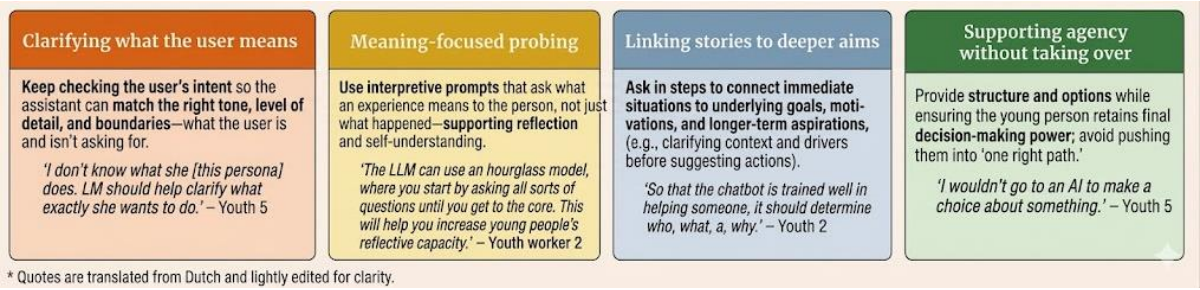

Figure 4. Sample Quotes Supporting Subthemes Under Theme 3

Once intent is clearer, stakeholders emphasize the role of interpretive questioning as a mechanism for surfacing personal meaning and strengthening reflective capacity. Instead of treating terms like "difficult" or "stressful" as fixed categories, an LLM-based system should ask how the young person experiences these states, and why they matter in their context. Participants likened this to an "hourglass" model of questioning: starting broad, then narrowing down to core issues, thereby helping adolescents articulate their own interpretations rather than adopting generic ones. This process positions the system not just as an information provider, but as a tool for self-understanding.

Building on this, narrative and goal exploration was identified as a key dialogic pattern for connecting short-term challenges to deeper aspirations and values. Stakeholders described stepwise question sequences that link immediate concerns (e.g., appearance, school stress, social difficulties) to what the young person ultimately wants in their life, clarifying the "weight" or significance of particular issues. They argued that LLM-based systems should structure prompts around basic narrative elements—who, what, where, when,

why—to help young people situate current problems within longer-term trajectories, which in turn supports more meaningful and contextually grounded recommendations.

Finally, participants stressed the importance of autonomy support as a guiding design principle for these dialogic strategies. Young people indicated that they do not want AI systems to make decisions for them, and stakeholders warned against interactions that push users toward narrow definitions of "success." Instead, personalization should focus on guiding exploration, normalizing uncertainty and mistakes, and highlighting options rather than issuing prescriptive or definitive answers. Figure 4 showcases sample quotes that support the analysis of subthemes under theme 3.

Operationalizing firsthand accounts into LLM responses

We systematically mapped stakeholder-identified subthemes to persuasive systems design (PSD) strategies, translating firsthand accounts into concrete dialogue examples for LLM fine-tuning (Table 6). This operationalization demonstrated how participatory insights can generate evidence-based system features, directly informing supervised fine-tuning and RLHF datasets for mental health information systems.

Several subthemes aligned naturally with PSD categories, providing structured bridges from qualitative priorities to system behaviors. For instance, subthemes like "beyond demographic labels" and "educational/literacy context" mapped to PSD's personalization and tunneling, yielding dialogues where the system asks, "What are the most important things I should know about you?" or offers literacy-adapted responses for user preference selection. Similarly, subthemes like "relational dynamics over roles" and "use of peer narratives" connected to social learning, operationalized through prompts modeling friend outreach or anonymized forum stories (e.g., "A young person shared how art class helped process a breakup"). "Everyday routines" and "dynamic adaptation" linked to reminders and suggestions, such as prompting, "You've mentioned enjoying art—want a quick wind-down idea?" or checking mood logs for real-time adjustment.

Other critical subthemes extended PSD but remained essential for safe IS design. Privacy and disclosure mapped to trustworthiness, producing dialogues like, "Here's how your data is handled—you control what you share." Content boundaries and referrals emphasized role clarity ("I'm not human, so my limits are…") and offline escalation (e.g., providing Kindertelefoon contact details when mood drops). conversational subthemes like intent clarification used tunneling ("What do you want to talk about right now?"), while autonomy supports enforced rules like ending turns with "or what else?" to preserve user agency.

These dialogue excerpts constitute ready-to-use training data that encodes stakeholder priorities into LLM interaction patterns, positioning PSD as a robust framework for aligning AI-enabled information systems with young people's well-being requirements. (Table 1)

Table 1: Mapping of Subthemes to PSD

Subthemes	Relevant PSD features	LLM dialogue extract or strategy (if no relevant PSD feature)
Beyond demographic labels	PTS: Tailoring	"What are the most important things I should know about you in order to help you?"
Educational and literacy context	PTS: Tailoring	[first time a conversation occurs, present three responses in varying literacy level to guide all future exchanges] "Rate which of these three responses you prefer."
Relational dynamics over roles	SS: Social learning	"You mentioned your friend has been reaching out to hang out recently—would you like tips on how to invite them to join you?"
Everyday routines and (digital) practices	DS: Reminder DS: Suggestion	"You've mentioned previously that you enjoy art in your free time — would you like a suggestion to use it as a quick wind-down after school?

Table 1: Mapping of Subthemes to PSD		
Subthemes	Relevant PSD features	LLM dialogue extract or strategy (if no relevant PSD feature)
Dynamic adaptation	PTS: Tailoring DS: Suggestion	"Have you filled out your mood chart in the app today? This can help determine how to best help you at the moment."
Privacy and disclosure	SCS: Trustworthiness	"As an LLM I can also access stories from people who faced similar situations. Would you like me to share their story with you without sharing identifying information?"
Content and behavioral boundaries	SCS: Trustworthiness, Real-world feel	"I am not a human so there are limits to how much I can help you, but a real person definitely can."
Referral to offline resources	SCS: Trustworthiness, Expertise, Real-world feel DS: Suggestion	"In the Netherlands, Kindertelefoon is a free and anonymous helpline to support young people. You can discuss anything on your mind. It can be reached by calling 0800-0432 or through their website for online chat. They are available every day from 11:00 to 21:00."
Use of peer narratives	SS: Social learning, Social facilitation SCS: Real-world feel	"You've mentioned previously that you enjoy art. I found a relevant story based on my training from forum posts on Kindertelefoon. So a poster once shared that after his breakup, he felt lost and isolated. So he joined a local art class which helped him process his feelings. He said that while creating art, he reflected about what had happened and how he felt about it. He was surprised to find that he felt better when he was done with the paintings."
Transparency and role clarity	SCS: Real-world feel	"As an LLM I am not a human so I have not experienced this. But I know a helpline where you can talk to someone anonymously." "We have better conversations when I know how you are feeling, because then I have more information to help you at a specific moment. That's why I ask regularly."
Intent clarification	PTS: Tunneling	Strategy: Every exchange starts with the LLM first asking what the user wants support with.
Narrative and goal exploration	PTS: Tunneling, Reduction	"This can be broken down step by step. First, decide on the tone you want to set — casual, warm, or direct. Then, draft one or two short messages that match the tone you want. Finally, send one when you're ready, so it feels less overwhelming and you get it done."

Discussion

As discussed above, our analysis identified three interrelated design priorities for safe and user-led personalization in LLM-based mental health information systems for young people. While all three stakeholder groups touched upon all three design priorities, their emphasis differed in ways that inform implementation. Young people highlighted contextual understanding, stressing that systems recognize their goals, routines, peer dynamics and the why behind the user's response rather than reduce them to labels. Parents prioritized safety boundaries, emphasizing the need to probe nuanced causes and ensure appropriate referral mechanisms. Youth care workers emphasized dialogic scaffolding, drawing on professional practice to articulate interpretive questioning techniques and content boundaries for younger users. These differential emphases suggest that multi-stakeholder alignment requires layered governance that helps create a more rounded and well designed system.

The findings also demonstrate how participatory personas function not as fixed user models, but as boundary objects that mediate between stakeholder perspectives and technical requirements for LLM alignment. The multi-layered method of allowing stakeholders to first add details from their experiences to enrich a static persona and then voice their needs, wants, and concerns within the persona perception scale enabled the creation of personalized system requirements. This approach extends IS traditions of socio-technical design and value-sensitive IS development by integrating co-created artifacts into LLM fine-tuning pipelines and advocates that stakeholder values take precedence over large-scale training data. The resulting dialogue examples serve as directly actionable preference signals for reinforcement learning from human feedback (RLHF), bridging community insights with system configuration.

Our findings carry targeted implications for IS stakeholders in the design and deployment of LLM-based wellbeing systems targeted to young people. For IS researchers, the study elevates firsthand accounts as a form of domain expertise for AI system alignment; it validates participatory personas as methodological tools for deriving design principles in socio-technical contexts. For digital health platform managers and CIOs, the priorities underscore the need to shift from demographic tailoring to systems that adapt to lived contexts, relational data, and literacy constraints while enforcing role clarity and data transparency. Mental health professionals and social workers emerge as critical co-designers who define personalization boundaries, referral pathways, and integration with human services. For AI developers and IS teams, the process offers a blueprint for community-informed fine-tuning, positioning stakeholder dialogue as structured inputs to RLHF that enhance system trustworthiness and long-term adoption.

Central to realizing these implications is resolving the apparent tension between the depth of contextual personalization our findings advocate and the privacy concerns stakeholders themselves raised. While our first design priority emphasizes probing beyond demographic labels to capture routines, relational dynamics, and underlying causes of challenges, our second priority simultaneously reveals that youth "may limit or misrepresent input due to fears of surveillance or data misuse". Our findings suggest these priorities operate as complementary rather than contradictory architectural layers: contextual understanding is elicited through voluntary, user-directed disclosure rather than covert system-initiated probing, while safety guardrails constrain how the system acts on disclosed information—it may ask open questions but cannot diagnose, offer specialist guidance, or retain information beyond consent. The dialogic scaffolding strategies identified in our third priority, such as intent clarification and autonomy support, ensure that depth of context emerges from user-led conversation rather than system-driven interrogation. This architecture aligns with Laufer and Wolfe's (1977) conceptualization of privacy as contextually negotiated rather than absolute, where perceived violation depends on the nature of information, conditions of disclosure, and its intended use. The distinction between eliciting context and acting on context resolves the contradiction: personalization depth is governed by user agency over disclosure, while safety boundaries govern system authority over response, ensuring multidimensional understanding does not translate into dangerous knowledge asymmetry or overreach.

Furthermore the research was carried out with the awareness of the ethical tensions inherent in applying persuasive techniques especially to vulnerable populations. Oinas-Kukkonen defines a Behavior Change Support System as one that reinforces or shapes behaviors "without using coercion or deception," integrating Self-Determination Theory to foster autonomy-supportive design and advocating for transparency to ensure trust-based behavior change (Oinas-Kukkonen & Harjumaa, 2009). Drawing on the core philosophy behind PSD principles—voluntariness, incremental change, transparency, and user-centricity—we identify the need for systems that remain open about persuasive intent, support users' own goals, recognize when a user no longer needs the system, and discourage habitual reliance on LLMs as a primary source of emotional regulation. These safeguards are particularly critical for younger users whose developmental needs require human relational experiences; systems must incorporate mechanisms encouraging breaks, reinforcing offline connections, and redirecting toward human services—extending our safety boundaries priority by ensuring persuasive features operate within developmental and relational constraints that position the system as an ethically bounded empowerment tool rather than a mechanism of technological paternalism (Wies et al., 2021; Nickel et al., 2023).

Another finding of this paper, which deviates from commercial applications and current industry practices, is the emphasis on neutral, reflective responses that optimize for engagement. While major

LLM deployments adopt human-like speech to increase acceptance, research demonstrates that mimicking capability risks blurring human-computer boundaries and cultivating dependency, particularly among youth ((Akbulut et al., 2024)). In the preventative mental health context, neutrality prioritizes long-term developmental outcomes—reflective capacity, self-efficacy, autonomy—over short-term engagement metrics, positioning the system as a scaffold for self-development rather than a source of emotional comfort (Singh et al., 2022).

Limitations

A primary limitation is that the proposed system behaviors and dialogue patterns have not yet been empirically validated through LLM fine-tuning and deployment testing. While our participatory process generated high-fidelity design requirements, IS research requires demonstration that these translate into consistent model outputs under real-world conditions. Additionally, while we have articulated how personalization and privacy operate as complementary architectural layers—with user-directed disclosure constraining depth and safety boundaries of system action—this resolution remains conceptual and has not been empirically tested with young users who may perceive these mechanisms differently in practice. Similarly, our ethical safeguards drawing on Oinas-Kukkonen's PSD principles (voluntariness, incremental change, transparency, user-centricity) represent design-level commitments whose effectiveness in preventing over-dependence and fostering healthy disengagement requires longitudinal evaluation with young people across diverse developmental stages and contexts.

Furthermore, our reliance on participatory personas authored by young people, while preserving authenticity by extracting content verbatim from templates, introduces risks of framing biases or static snapshots that may not fully capture contextual fluidity. Although we used personas as reflective artifacts rather than implementation inputs, future IS work should explore dynamic persona evolution across child development stages, ensuring diversity across intersecting characteristics to avoid reductive stereotyping in system training data. Also the study drew from a small, European sample of young people, limiting statistical generalizability.

Conclusion

This study develops design principles for personalizing LLM-based wellbeing systems for young people by mapping stakeholder insights to persuasive system design categories. The key contribution reframes personalization as a dynamic, dialogic capability that continuously adapts through ongoing interaction rather than static user profiling, providing a practical blueprint for IS researchers, digital health managers, and AI developers to build trustworthy, community-centered AI systems that enhance human care ecosystems.

References

Adomavicius, D., & Tuzhilin, A. (2006). Personalization technologies: A process-oriented perspective, *48*(6), 449–450. https://doi.org/10.1007/s11576-006-0098-7

Agarwal, S., Almeida, D., Askell, A., Christiano, P., Hilton, J., Jiang, X., Kelton, F., Leike, J., Lowe, R., Miller, L., Mishkin, P., Ouyang, L., Ray, A., Schulman, J., Simens, M., Slama, K., Wainwright, C., Welinder, P., Wu, J., & Zhang, C. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems 35*, 27730–27744. https://doi.org/10.52202/068431-2011

Akbulut, C., Weidinger, L., Manzini, A., Gabriel, I., & Rieser, V. (2024). All too human? Mapping and mitigating the risk from anthropomorphic AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*, 13–26. https://doi.org/10.1609/aies.v7i1.31613

Anderson, T. L., Valiauga, R., Tallo, C., Hong, C. B., Manoranjithan, S., Domingo, C., Paudel, M., Untaroiu, A., Barr, S., & Goldhaber, K. (2024). Contributing factors to the rise in adolescent anxiety and associated mental health disorders: A narrative review of current literature. *Journal of Child and Adolescent Psychiatric Nursing*, *38*(1). https://doi.org/10.1111/jcap.70009

Baumel, A., Muench, F., Edan, S., & Kane, J. M. (2019). Objective user engagement with mental health apps: Systematic search and panel-based usage analysis. *Journal of Medical Internet Research*, *21*(9), e14567. https://doi.org/10.2196/14567

Bhattacharyya, O., Mossman, K., Gustafsson, L., & Schneider, E. C. (2019). Using human-centered design to build a digital health advisor for patients with complex needs: Persona and prototype development. *Journal of Medical Internet Research*, *21*(5), e10318. https://doi.org/10.2196/10318

Brewster, R. C. L., Zahedivash, A., Tse, G., Bourgeois, F., & Hadland, S. E. (2025). Characteristics and safety of consumer chatbots for emergent adolescent health concerns. *JAMA Network Open*, *8*(10), e2539022. https://doi.org/10.1001/jamanetworkopen.2025.39022

Catania, L. S., Hetrick, S. E., Newman, L. K., & Purcell, R. (2011). Prevention and early intervention for mental health problems in 0–25 year olds: Are there evidence-based models of care? *Advances in Mental Health*, *10*(1), 6–19. https://doi.org/10.5172/jamh.2011.10.1.6

Dooley, B., & Fitzgerald, A. (2012). Methodology on the My World Survey (MWS): A unique window into the world of adolescents in Ireland. *Early Intervention in Psychiatry*, *7*(1), 12–22. https://doi.org/10.1111/j.1751-7893.2012.00386.x

Du, Q., Ren, Y., Meng, Z., He, H., & Meng, S. (2025a). The efficacy of rule-based versus large language model-based chatbots in alleviating symptoms of depression and anxiety: systematic review and meta-analysis. *Journal of Medical Internet Research*, *27*, e78186–e78186. https://doi.org/10.2196/78186

Du, Q., Ren, Y., Meng, Z., He, H., & Meng, S. (2025b). The efficacy of rule-based versus large language model-based chatbots in alleviating symptoms of depression and anxiety: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *27*, e78186–e78186. https://doi.org/10.2196/78186

Feng, S., Li, X. (Leah), & Wake, A. N. (2026). Engaging Artificial Intelligence (AI)-based chatbots in digital health: A systematic review. *PLOS Digital Health*, *5*(2), e0001201. https://doi.org/10.1371/journal.pdig.0001201

Fernandes, J., Antunes, A., Campos, J., Dias, J., & Santos, P. (2025). FlexiDialogue: Integrating dialogue trees for mental health with large language models. *Proceedings of the 11th International Conference on Information and Communication Technologies for Ageing Well and E-Health*, 268–275. https://doi.org/10.5220/0013286900003938

Figueroa, C. A., Ramos, G., Psihogios, A. M., Ekuban, E., Bansie, P., de Haas, M., Karnik, N. S., Ajilore, O., Anderson, E. E., & Stiles-Shields, C. (2025). Advancing youth co-design of ethical guidelines for AI-powered digital mental health tools. *Nature Mental Health*, *3*(8), 870–878. https://doi.org/10.1038/s44220-025-00467-7

Fogg, B. J. (2002). Persuasive technology. *Ubiquity*, *2002*(December), 2. https://doi.org/10.1145/764008.763957

Fogg, B. J. (2003). *Persuasive technology: Using computers to change what we think and do*. Morgan Kaufmann.

Friedman, B., Kahn, P. H., Jr., & Borning, A. (2008). Value sensitive design and information systems. *The Handbook of Information and Computer Ethics*, 69–101. https://doi.org/10.1002/9780470281819.ch4

Fusar-Poli, P. (2019). Integrated mental health services for the developmental period (0 to 25 years): A critical review of the evidence. *Frontiers in Psychiatry*, *10*. https://doi.org/10.3389/fpsyt.2019.00355

Giri, S., Kulasinghe, K., & Oinas-Kukkonen, H. (2025). Exploring Large Language Model-Based Mental Health Interventions: A Systematic Review with a Persuasive System Design Lense. PACIS 2025 Proceedings. https://aisel.aisnet.org/pacis2025/emerg_tech/emerg_tech/9

González Corbelle, J., Bugarín-Diz, A., Alonso-Moral, J., & Taboada, J. (2022). Dealing with hallucination and omission in neural Natural Language Generation: A use case on meteorology. *Proceedings of the 15th International Conference on Natural Language Generation*, 121–130. https://doi.org/10.18653/v1/2022.inlg-main.10

Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., & Li, K. (2024). Large language models for mental health applications: Systematic review. *JMIR Mental Health*, *11*, e57400. https://doi.org/10.2196/57400

Hijdra, R., Oude Groeniger, J., Burdorf, A., & Schuring, M. (2025). The effect of mental health problems on having a 'neither in employment nor in education or training' period and the mediating role of high school dropout: A register-based study with a 14-year follow-up. *Journal of Epidemiology and Community Health*, *79*(8), 600–606. https://doi.org/10.1136/jech-2024-222197

Hornstein, S., Zantvoort, K., Lueken, U., Funk, B., & Hilbert, K. (2023). Personalization strategies in digital mental health interventions: A systematic review and conceptual framework for depressive symptoms. *Frontiers in Digital Health*, *5*. https://doi.org/10.3389/fdgth.2023.1170002

Hua, Y., Siddals, S., Ma, Z., Galatzer-Levy, I., Xia, W., Hau, C., Na, H., Flathers, M., Linardon, J., Ayubcha, C., & Torous, J. (2025). Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: A systematic review. *World Psychiatry*, *24*(3), 383–394. https://doi.org/10.1002/wps.21352

Ilapaka, A., & Ghosh, R. (2025). A comprehensive rag-based LLM for ai-driven mental health chatbot. *2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)*, 1–5. https://doi.org/10.1109/ichora65333.2025.11017017

Jansen, B. J., Salminen, J. O., & Jung, S.-G. (2020). Data-driven personas for enhanced user understanding: Combining empathy with rationality for better insights to analytics. *Data and Information Management*, *4*(1), 1–17. https://doi.org/10.2478/dim-2020-0005

Jung, K., Lee, G., Huang, Y., & Chen, Y. (2025). "I've talked to ChatGPT about my issues last night.": Examining Mental Health Conversations with Large Language Models through Reddit Analysis. *Proceedings of the ACM on Human-Computer Interaction*, *9*(7), 1–25. https://doi.org/10.1145/3757537

Katti, V., & Karande, K. J. (2026). Personalized depression management system using llms and reinforcement learning: A survey. In *Lecture Notes in Networks and Systems* (pp. 15–19). Springer Nature Singapore. https://doi.org/10.1007/978-981-95-5079-1_2

Kessler, R. C., Aguilar-Gaxiola, S., Alonso, J., Chatterji, S., Lee, S., Ormel, J., Üstün, T. B., & Wang, P. S. (2009). The global burden of mental disorders: An update from the WHO World Mental Health (WMH) Surveys. *Epidemiologia e Psichiatria Sociale*, *18*(1), 23–33. https://doi.org/10.1017/s1121189x00001421

Kurian, N. (2025). Designing child-safe conversational AI: three dilemmas for responsible design. *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, 1–5. https://doi.org/10.1145/3719160.3737638

Laufer, R. S., & Wolfe, M. (1977). Privacy as a concept and a social issue: a multidimensional developmental theory. *Journal of Social Issues*, *33*(3), 22–42. https://doi.org/10.1111/j.1540-4560.1977.tb01880.x

M, J., P, V. S., & Kryvinska, N. (2024). Exploring the Chatbot usage intention-A mediating role of Chatbot initial trust. *Heliyon*, *10*(12), e33028. https://doi.org/10.1016/j.heliyon.2024.e33028

Marsden, N., & Haag, M. (2016b). Stereotypes and politics. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4017–4031. https://doi.org/10.1145/2858036.2858151

Martin, R., Sturt, J., & Griffiths, F. (2020). The impact of digital communication on adolescent to adult mental health service transitions. *Journal of Research in Nursing*, *25*(3), 277–288. https://doi.org/10.1177/1744987120915746

Matthews, T., Judge, T., & Whittaker, S. (2012). How do designers and user experience professionals actually perceive and use personas? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1219–1228. https://doi.org/10.1145/2207676.2208573

McGinn, J. (Jen), & Kotamraju, N. (2008). Data-driven persona development. Proceedings of the *SIGCHI Conference on Human Factors in Computing Systems*, 1521–1524. https://doi.org/10.1145/1357054.1357292

McGorry, P., Gunasiri, H., Mei, C., Rice, S., & Gao, C. X. (2025). The youth mental health crisis: Analysis and solutions. *Frontiers in Psychiatry*, *15*. https://doi.org/10.3389/fpsyt.2024.1517533

McGorry, P. D., Killackey, E., Iyer, S., & Dalale, N. (2024). Challenges in addressing youth mental health. *The Lancet Psychiatry*, *11*((9)), 731–774. https://doi.org/10.1016/S2215-0366(24)0035

Mohamed, J. A., Mohamed, A. A., Abdulle, A. Y., & Osman, M. M. (2025). How Hormuud's WhatsApp Chatbot Services Influence Customer experience and Loyalty: Using Structural Equation Modeling. *Mathematical Modelling of Engineering Problems*, *12*(9). https://doi.org/10.18280/mmep.120914

Mohr, D. C., Weingardt, K. R., Reddy, M., & Schueller, S. M. (2017). Three problems with current digital mental health research . . . and three things we can do about them. *Psychiatric Services*, *68*(5), 427–429. https://doi.org/10.1176/appi.ps.201600541

Molyneaux, R., Mirembe, E., Leicester, S., Schley, C., & Alisic, E. (2024). Digital disclosures: Exploring the role of online mental health services in supporting young people who disclose violence or maltreatment. *Child Protection and Practice*, *1*, 100016. https://doi.org/10.1016/j.chipro.2024.100016

Murthi, B. P. S., & Sarkar, S. N. (2003). The role of the management sciences in research on personalization. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.407940

Nickel, P. J., Loosman, I., Frank, L., & Vinnikova, A. (2023). Justice and empowerment through digital health: ethical challenges and opportunities. *Digital Society*, *2*(3). https://doi.org/10.1007/s44206-023-00075-x

Oinas-Kukkonen, H. (2018). Personalization myopia. *Proceedings of the 22nd International Academic Mindtrek Conference*, 88–91. https://doi.org/10.1145/3275116.3275121

Oinas-Kukkonen, H., & Harjumaa, M. (2009). Persuasive systems design: Key issues, process model, and system features. *Communications of the Association for Information Systems*, *24*. https://doi.org/10.17705/1cais.02428

Oinas-Kukkonen, H., Pohjolainen, S., & Agyei, E. (2022). Mitigating issues with/of/for true personalization. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.844817

Otaki, B., & Lindwall, O. (2024). Generative AI and the human touch: Investigating the changing landscape of feedback in higher education. *Proceedings of the International Conference of the Learning Sciences*, 1099–1102. https://doi.org/10.22318/icls2024.230928

Potts, C., Ennis, E., Bond, R. B., Mulvenna, M. D., McTear, M. F., Boyd, K., Broderick, T., Malcolm, M., Kuosmanen, L., Nieminen, H., Vartiainen, A. K., Kostenius, C., Cahill, B., Vakaloudis, A., McConvey, G., & O'Neill, S. (2021). Chatbots to support mental wellbeing of people living in rural areas: Can user groups contribute to co-design? *Journal of Technology in Behavioral Science*, *6*(4), 652–665. https://doi.org/10.1007/s41347-021-00222-6

Pratt, N., Madhavan, R., & Weleff, J. (2024). Digital dialogue—how youth are interacting with chatbots. *JAMA Pediatrics*, *178*(5), 429. https://doi.org/10.1001/jamapediatrics.2024.0084

Rijcken, E., Zervanou, K., Mosteiro, P., Scheepers, F., Spruit, M., & Kaymak, U. (2024). *Machine learning vs. rule-based methods for document classification of electronic health records within mental health care - A systematic literature review*. Springer Science and Business Media LLC. https://doi.org/10.21203/rs.3.rs-2320804/v2

Rivolli, A., Amaral, C., Guardão, L., de Sá, C. R., & Soares, C. (2019). KnowBots: Discovering relevant patterns in chatbot dialogues. In *Lecture Notes in Computer Science* (pp. 481–492). Springer International Publishing. https://doi.org/10.1007/978-3-030-33778-0_36

Salminen, J., Guan, K., Jung, S.-G., Chowdhury, S. A., & Jansen, B. J. (2020). A literature review of quantitative persona creation. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376502

Sanna, L., Bolpagni, M., Fietta, V., Gavioli, G., Franzin, M., Dragoni, M., & Gabrielli, S. (2025). LLM-enriched finite-state chatbots for mental health support: A case study on Self-Help+. In *Lecture Notes in Computer Science* (pp. 386–396). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-95838-0_38

Singh, V., Kumar, A., & Gupta, S. (2022). Mental health prevention and promotion—a narrative review. *Frontiers in Psychiatry*, *13*. https://doi.org/10.3389/fpsyt.2022.898009

Sittig, S., Iyengar, M. S., Florez-Arango, J. F., & Chaudhry, B. M. (2024). Creating, sustaining, and evaluating personalized digital health systems. *2024 IEEE International Conference on Big Data (BigData)*, 5884–5890. https://doi.org/10.1109/bigdata62323.2024.10825134

Skjuve, M., Følstad, A., & Brandtzaeg, P. B. (2026). ChatGPT as a mental health advisory service: Comparing evaluations from youth and health professionals. *DIGITAL HEALTH*, *12*. https://doi.org/10.1177/20552076261427447

Sotiropoulos, A., Valapu, S. T., Lei, L., Coleman, J., & Krishnamachari, B. (2025). Crowd-SFT: Crowdsourcing for LLM alignment. *2025 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPS)*, 135–136. https://doi.org/10.1109/dapps65174.2025.00026

Talk*, trust, and trade-offs: How and why teens use AI companions*. (2025). Common Sense Media. https://www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions

Turner, P., & Turner, S. (2011). Is stereotyping inevitable when designing with personas? *Design Studies*, *32*(1), 30–44. https://doi.org/10.1016/j.destud.2010.06.002

Turpeinen, M., & Saari, T. (2004). System architecture for psychological customization of communication technology. *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of The*, 10 pp. https://doi.org/10.1109/hicss.2004.1265477

Vesanen, J. (2007). What is personalization? A conceptual framework. *European Journal of* Marketing, *41*(5/6), 409–418. https://doi.org/10.1108/03090560710737534

Wanniarachchi, V. U., Greenhalgh, C., Choi, A., & Warren, J. R. (2025). Personalization variables in digital mental health interventions for depression and anxiety in adolescents and youth: A scoping review. *Frontiers in Digital Health*, *7*. https://doi.org/10.3389/fdgth.2025.1500220

Wies, B., Landers, C., & Ienca, M. (2021). Digital mental health for young people: a scoping review of ethical promises and challenges. *Frontiers in Digital Health*, *3*. https://doi.org/10.3389/fdgth.2021.697072

World Health Organization: WHO. (2025, September 1). Mental health of adolescents. World Health Organization:WHO. https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health

Yu, H. Q., & McGuinness, S. (2024). An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system. *Journal of Medical Artificial Intelligence*, *7*, 16–16. https://doi.org/10.21037/jmai-23-136