Trust Stack for Mental Health AI: A Survey of Calibration across Human, Interaction, and AI Layers

Xin Sun[1,2], Yue Su[3], Yifan Mo[3], Qingyu Meng[3], Yuxuan Li[3], Min Chen[3], Mengyuan Zhang[3], Saku Sugawara[1], Charlotte Gerritsen[3], Sander L. Koole[3], Koen Hindriks[3], Jiahuan Pei[3]

[1]National Institute of Informatics (NII), Japan
[2]University of Amsterdam, the Netherlands
[3]Vrije Universiteit Amsterdam, the Netherlands

Abstract

Language-based AI is increasingly deployed for mental health support, yet trust is evaluated in interdisciplinary but operationally misaligned ways: NLP and AI work measures robustness, safety, privacy, and explanations, while psychotherapy, HCI, and regulatory work emphasize therapeutic fidelity, lived experience, empathy, and reliance. Empathetic chatbots can elicit strong user trust without commensurate safety, while safer systems are under-trusted when their boundaries are opaque, a calibration gap no single community owns. Through a structured scoping synthesis of 61 papers, we survey this landscape into a three-layer framework separating (L1) human-oriented trust, (L2) interaction-oriented trustworthiness, and (L3) AI-oriented trustworthiness, and map five stakeholder perspectives onto these layers. We outline a research agenda for building socio-technically aligned trustworthy AI for mental health support, highlighting that the central objective should shift from maximizing perceived trust to calibrating human trust to demonstrated interaction- and AI-level trustworthiness.

1 Introduction

Building trustworthy artificial intelligence (AI) systems for mental support is an interdisciplinary problem. Recent advances in large language models (LLMs) have expanded applications ranging from screening and psychoeducation to conversational support (Na et al., 2025). Recent work spans clinical, natural language processing (NLP), human computer interaction (HCI), safety, and regulatory communities (see Appendix B, Figure 3 for a discipline-network view of 1,706-paper mapping corpus). Yet this literature is not isolated; rather, it is operationally misaligned: different communities often use similar trust language while evaluating different objects, such as users' subjective trust, interactive behavior, or system-level trustworthiness.

This misalignment matters because trust and trustworthiness are related but not equivalent. Trust is a trustor's subjective attitude and willingness to rely on a system, whereas trustworthiness refers to evaluable properties of the trustee, such as reliability, safety, privacy protection, fairness, and accountability (Jacovi et al., 2021; Lee and See, 2004). In mental health, the distinction is consequential: empathetic or human-like interaction can increase perceived trust even when AI model behavior is unstable, while strong safeguards may still fail if users cannot understand system limits.

Prior surveys provide important foundations but leave three gaps that we address. First, general human-AI trust surveys (Mehrotra et al., 2024; Regona et al., 2026) synthesize trust theories across domains but do not engage with mental-health-specific pressures such as clinical confidentiality, crisis escalation, vulnerability, or over-disclosure. Second, medical-AI surveys (Zhu et al., 2025) focus on hallucination but treat trust as a property of model accuracy rather than a multi-stakeholder problem. Third, psychotherapy-focused LLM surveys (Na et al., 2025) catalog applications and risks but do not separate subjective trust, interaction- and system-level trustworthiness, making it difficult to diagnose where trust failures originate. Our framework addresses the calibration gap by mapping evidence to three layers, allowing trust failures to be attributed to specific stakeholder responsibilities. Table 6 further details how mental-health pressures reshape general human-AI trust evaluation.

Specifically, we ask: how should the field evaluate whether users' trust in language-based mental health AI is appropriate? We address this gap with a three-layer trust framework for language-based mental health support (MHS) systems. The framework distinguishes human-oriented trust (users' subjective trust, expectations, and reliance), interaction-oriented trustworthiness (user-visible behaviors that communicate capability, boundaries,

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 731–752
October 17-21, 2026 ©2026 Association for Computational Linguistics

uncertainty, empathy, and control), and AI-oriented trustworthiness (model- and system-level properties that justify trust). We map these layers to psychotherapy, HCI, AI/NLP, safety/security, and regulatory perspectives, while treating clients and end users as the primary trustors represented in the human-oriented layer. Details of the PRISMA-guided review (Page et al., 2021) (not a database-exhaustive systematic review) and coding procedure are provided in Appendix B, including the PRISMA flow diagram (separate stratified search of 735 records; 61-paper analytic corpus).

The contributions of this survey are threefold.

- We formalize trust calibration in mental health AI by distinguishing trust, trustworthiness, and reliance across three evaluation layers.
- We conduct a PRISMA-guided scoping review of recent language-based mental health AI literature through this framework, clarifying how papers are searched, screened, coded, and mapped to stakeholder evaluation paradigms.
- We derive an agenda to examine current methodology and evaluation across layers and stakeholders, positioning the "gap" that links AI advances to interaction behavior and human trust.

2 Conceptual Framework

Trustworthy MHS AI is not a monolithic concept. It is multi-stakeholder evaluation problem in which human trust, interaction behavior, model capability, safety governance and responsibility must be examined together. We distill core trustworthiness *criteria* (e.g., robustness, faithfulness, transparency) into a three-layer trust framework, and review this landscape of involved *stakeholders* (i.e., practitioners, AI/HCI/safety researchers, regulators).

2.1 Definitions and Background

We use *trust* to refer to the trustor's subjective attitude toward, and willingness to rely on a system (Lee and See, 2004; Jacovi et al., 2021). The primary trustors are clients, patients, and end users, although clinicians and institutions may also form trust judgments. We use *trustworthiness* to refer to evaluable properties of the trustee: the AI system, interaction design, deployment workflow, or institution that may or may not deserve trust, including reliability, safety, privacy, fairness, transparency, accountability, and clinical appropriateness.

This distinction matters because trust can be warranted or unwarranted. Users may over-trust

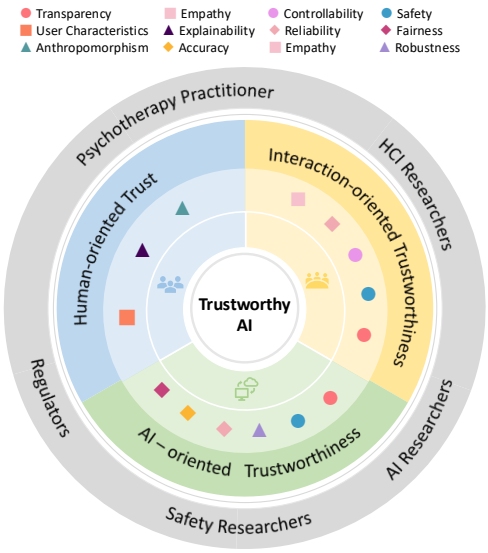


Figure 1: Based on a cross-disciplinary literature review, we identify five research and governance perspectives that shape trustworthy mental health AI. We synthesize these perspectives into three layers: human-oriented trust, interaction- and AI-oriented trustworthiness. Clients, patients, and end users are the primary trustors represented in the human-oriented layer and defined in Table 5; the outer sectors indicate evaluation perspectives, not exclusive layer ownership.

systems that communicate empathy, fluency, or human-likeness beyond actual capability, or under-trust safer systems whose safeguards are opaque. We therefore treat *reliance* as a behavioral outcome related to, but not identical with trust. The goal is appropriate reliance: human trust should be calibrated to demonstrated interaction- and system-level capabilities rather than persuasive cues.

2.2 Stakeholder Landscape

We define stakeholder groups by their primary evaluation paradigms rather than by publication venues. Venues were used only as search entry points in review process. The analytic categories below reflect what each community typically treats as evidence of trust or trustworthiness. Table 5 summarizes these definitions and clarifies how client, patient, and end-user perspectives enter the framework.

2.2.1 Psychotherapy Practitioners

Psychotherapy practitioners include licensed therapists, clinical psychologists, psychiatrists acting in therapeutic roles, and mental-health researchers. This perspective treats user trust as clinically sensitive: trust is relational, dynamic, and depends on user characteristics, system cues, oversight, vulnerability, and perceived responsibility (Chi et al., 2021; Kauttonen et al., 2025; Gille et al., 2025;

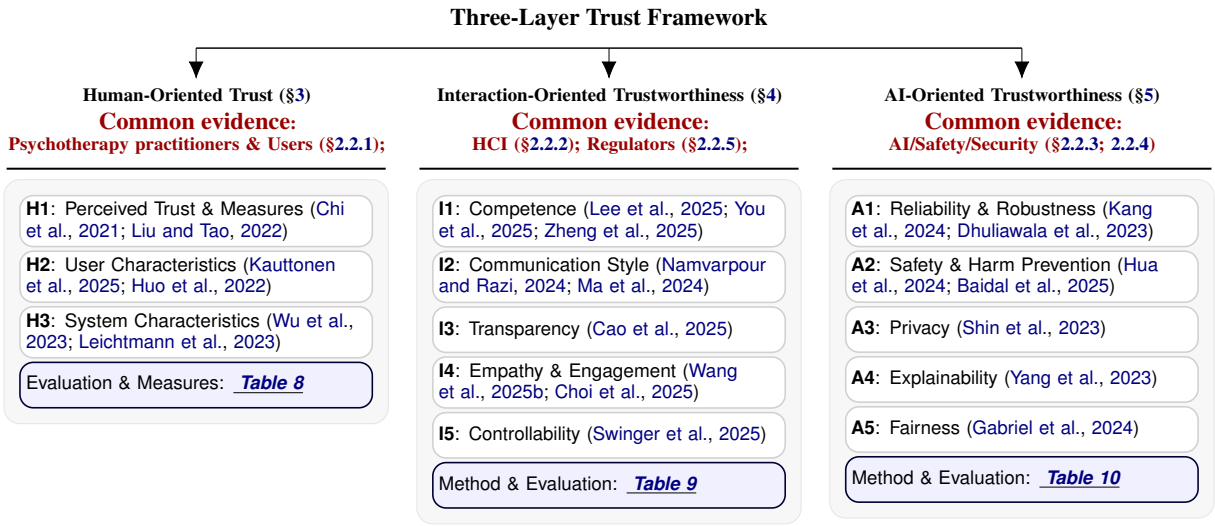


Figure 2: The proposed Three-Layer Framework. Each layer reflects dominant evaluation perspectives across stakeholders and summarizes key trust criteria with representative literature. The evidence sources shown in each column are not exclusive to layer owners; individual literature may contribute evidence to multiple layers. Examples of cross-layer evaluation signals and rubric are detailed in Table 2 and Table 3.

Rai et al., 2025). It is therefore concerned with calibrated reliance, not merely adoption or engagement. We distinguish mental-health-specific evidence from conversational support and counseling workflows (Song et al., 2025; Sun et al., 2025; Wester et al., 2024) from transferable foundations in broader high-risk decision and digital-health contexts (Leichtmann et al., 2023; Mayer et al., 2024; Woodcock et al., 2021); the latter inform explanation, disclosure, and oversight mechanisms but do not substitute for mental-health evidence.

2.2.2 HCI Researchers

HCI researchers study how trust is communicated and negotiated for mental health support through interface design, conversational behavior, and lived experience. This perspective emphasizes transparency, empathy, engagement, and controllability: interaction should surface capability, limits, and safeguards in ways users can understand and act on (Thieme et al., 2023; Namvarpour and Razi, 2024; Wang et al., 2025b; Cao et al., 2025).

2.2.3 AI Researchers

AI and NLP researchers define which model, retrieval, data, and evaluation properties count as evidence of system-level trustworthiness. In mental health settings, this includes robustness, fairness, faithful grounding, RAG attribution, and evaluator reliability. Because LLM behavior can vary across runs and evaluation signals can be biased or unfaithful (Cao, 2025; Liu et al., 2024; Wallat et al., 2025), model-level evidence is necessary but must be linked to interaction and human outcomes.

2.2.4 Safety and Security Researchers

Safety and security researchers ask whether systems remain safe under adversarial, high-risk, or privacy-sensitive conditions. Their evidence targets prompt injection, jailbreaks, backdoor triggers, memorization, and crisis stress tests (Yu et al., 2025; Hua et al., 2024; Wang et al., 2023; Baidal et al., 2025; Alghamdi et al., 2025). This perspective is important because mental health users disclose highly sensitive information and face severe consequences from unsafe advice or leakage.

2.2.5 Regulators and Standards Associations

Regulators and standards associations translate ethical principles into deployable requirements. We review professional frameworks in U.S., EU (NIST, 2023; AMA, 2024; EPEUCO, 2024; AIHLEG, 2019; APA, 2025; Pillay, 2025); despite different legal status, they converge on human-centricity, risk mitigation, privacy, bias auditing, accountability, and view that AI should augment rather than replace human clinical care.

2.3 Three-Layer Framework

Trust in mental health AI does not come from a single source. It is built through a calibration pathway: how the system works internally, how its capabilities and limits are communicated during interaction, and how users ultimately experience and rely on its support. To organize these views, we synthesize the following three-layer trust framework (Figure 1; 2).

Human-oriented trust. Even if a system is technically robust and interactively competent, trust ultimately depends on whether users feel the system to be trustworthy and rely on it appropriately. Human-oriented trust captures users' subjective trust toward mental health AI systems. It concerns how trust is perceived, fostered, and calibrated by users, shaped by individual characteristics, expectations, prior experiences, vulnerability, and perceived stakes. This layer reflects trust as a psychological state rather than a system property.

Interaction-oriented trustworthiness. Once users begin interacting with systems, trustworthiness becomes user-visible. Interaction-oriented trustworthiness concerns how trust is mediated through conversational behavior, feedback, uncertainty communication, transparency cues, controllability, escalation, and safety mechanisms. This layer bridges system capability and user perception: it should communicate capability and limits, not simply make the system feel supportive.

AI-oriented trustworthiness. AI-oriented trustworthiness refers to system-level criteria defined at the model, data, evaluation, and infrastructure levels, including reliability, robustness, uncertainty awareness, fairness, privacy protection, evaluation validity, and avoidance of harmful or misleading outputs. These criteria are primarily specified and assessed by AI, NLP, and security research before and during deployment, and they constrain which interaction-level trustworthiness can be warranted. Failures at this layer, such as harmful generation, privacy leakage, or biased evaluation, may undermine trust regardless of interaction quality.

Layer labels are assigned by evaluation target, not by keyword. Cross-layer assignments are expected: explainability may be evaluated as model faithfulness or explanation design; privacy as data protection or disclosure safety. Primary and secondary labels locate evidence across layers. (Table 4 is coding examples and Table 2 is aggregate assignments.) The three-layer frame trust in MHS as calibration problem that should track interaction- and AI-oriented trustworthiness. Taxonomy structures following review in Table 3, which is evaluation aids, not certification or safety guarantee.

3 Human-Oriented Trust

Human-oriented trust is understood as a subjective psychological state that shapes users' willingness to adopt, disclose to, and continue using mental health AI systems. We organize human-oriented trust around three construct-level questions: what trust is, who trusts, and how trust is perceived through user-visible cues. This layer measures trust and reliance outcomes; it does not by itself establish whether the system deserves users' trust. Table 8 summarizes the criteria and measures.

3.1 What Is Trust

Scope. This dimension targets users' subjective trust judgments toward AI-powered mental health systems. At a basic level, trust reflects a user's confidence that an AI system is reliable, benevolent, and appropriate enough to rely on in a sensitive context. This aligns with trust theories such as the ability–benevolence–integrity model (Mayer et al., 1995) and MATCH framework (Liao and Sundar, 2022), which characterize trust as a belief about an agent's capability, dependability, and communicated trustworthiness cues. Prior work measures trust in two main ways. Some studies treat trust as a *unidimensional* construct, measuring overall trust as a single judgment (Wu et al., 2023; Woodcock et al., 2021; Aoki, 2021; Kauttonen et al., 2025; Aktan et al., 2022; Liu and Tao, 2022; Mayer et al., 2024; Huo et al., 2022; Youn and Jin, 2021; Zhao et al., 2025). Others adopt a *multidimensional* view, decomposing trust into components (e.g., perceived reliability, agency, and competence (Chi et al., 2021; Luetke Lanfer et al., 2023; Leichtmann et al., 2023; Gille et al., 2025; Brunswicker et al., 2025; Rai et al., 2025)).

Evaluation. Trust is primarily measured through subjective assessments, while observed interaction behavior is better interpreted as evidence of reliance or use behavior unless paired with self-reported trust. Studies commonly employ validated scales capturing perceived agency (Luetke Lanfer et al., 2023), system reliability (Chi et al., 2021), trust propensity (Chi et al., 2021), or overall trust (Wu et al., 2023; Woodcock et al., 2021; Liu and Tao, 2022; Kauttonen et al., 2025; Aoki, 2021; Youn and Jin, 2021). Some work adopts single-item measures for direct trust assessment (Woodcock et al., 2021; Aoki, 2021; Mayer et al., 2024). Qualitative approaches examine how users interpret and justify trust judgments (Luetke Lanfer et al., 2023). Behavioral indicators such as interaction patterns, decision choices, or continued use can reveal reliance and possible miscalibration (Leicht-

mann et al., 2023; Brunswicker et al., 2025; Mayer et al., 2024; Aktan et al., 2022), but should not be treated as identical to subjective trust.

3.2 Who Trusts

Scope. This dimension targets user-side characteristics that modulate trust formation. Prior work identifies attitudes toward AI, personality traits, familiarity, prior experience, and perceived social support as important determinants of trust in mental health contexts (Kauttonen et al., 2025; Zhao et al., 2025; Huo et al., 2022). Literacy level also matters: users with limited AI knowledge often exhibit greater reliance on AI (Woodcock et al., 2021). Studies show that human-in-the-loop oversight increases trust in AI (Mayer et al., 2024; Aoki, 2021; Liu and Tao, 2022), reflecting users' preference for shared control in high-stakes contexts.

Evaluation. User characteristics are frequently measured using self-developed instruments tailored to demographic, socio-cultural factors (Aktan et al., 2022; Leichtmann et al., 2023; Zhao et al., 2025; Huo et al., 2022; Woodcock et al., 2021; Liu and Tao, 2022; Kauttonen et al., 2025), while validated scales are more commonly used for stable traits such as personality or trust propensity.

3.3 How Trust Is Perceived

Scope. This dimension targets user-visible cues that shape human trust. While many system features may influence trust, psycho- and HCI-oriented studies place particular emphasis on anthropomorphism and explainability. In conversational mental health AI, anthropomorphism refers primarily to interactive cues, such as human-like language, empathy expressions, social presence, and emotional support, rather than embodiment or VR/AR immersion effects (Wang et al., 2025b; Cao et al., 2025; Liu and Tao, 2022). Explainability refers to system's ability to communicate rationale behind its outputs and decisions in user-understandable ways (Leichtmann et al., 2023).

Evaluation. Anthropomorphism is commonly measured by validated scales (Liu and Tao, 2022). Prior work distinguishes between superficial forms (e.g., appearance or communication style) and deeper forms involving moral agency and emotions (Wu et al., 2023). Studies often manipulate empathizing and systemizing behaviors before measuring perceived anthropomorphism (Brunswicker et al., 2025). Explainability is typically studied by varying clarity and form of system explanations and assessing user comprehension, confidence, and willingness to rely (Leichtmann et al., 2023; Woodcock et al., 2021). Qualitative methods are frequently used to examine how users interpret these cues and whether the cues support calibrated trust (Wang et al., 2025b; Cao et al., 2025; Namvarpour and Razi, 2024; Choi et al., 2025). These cues affect Layer 1 trust by signaling competence and emotional understanding, but they do not by themselves establish Layer 3 capability; this is the calibration risk foregrounded by the framework.

Human-oriented trust thus captures how users perceive trust. These perceptions may diverge from interaction- and AI-level trustworthiness, which motivates evaluating calibrated trust across layers.

4 Interaction-Oriented Trustworthiness

Interaction-oriented trustworthiness asks whether system behavior supports appropriate trust during use. Unlike AI-oriented trustworthiness, it emerges in real time and is evaluated through observed conversations, interface affordances, and user comprehension. Its function is to translate system capability, uncertainty, limits, and safeguards into forms that users can understand and act on. Layer assignment is evidence-based: an NLP, AI, or health literature belongs here when its main evidence concerns user-centric behavior, control, or crisis routing. Table 9 summarizes the criteria and measures.

4.1 Competence and Reliability

Scope. Competence and reliability target whether MHS systems consistently provide accurate, contextually appropriate, and therapeutically meaningful responses during interaction. Prior work operationalizes this across application types: story-based interventions involve professional review of user-generated content for accuracy (Sien and McGrenere, 2025). Chatbots aim to recognize users' emotions and deliver evidence-based interventions (Wester et al., 2024), while LLM agents generate responses aligned with psychological principles while avoiding ineffective or harmful advice (Song et al., 2025; Sun et al., 2025). Therapist-training systems assess competence via structured evaluations of treatments (Swinger et al., 2025).

Methods and Evaluation. Literature studies competence through multiple ways. Qualitative interviews and analyses examine whether AI provide useful and therapeutically relevant responses, while

also documenting failures such as overly generic or inaccuracy (Song et al., 2025). Expertise-aligned generation ensures LLMs are controlled by expert-authored scripts, improving adherence to therapeutic principles (Sun et al., 2025). In contrast, systems for therapist training rely on protocol-based assessments, such as time-stamped behavioral analysis and checklist-style scoring aligned with the clinic manuals (Swinger et al., 2025).

4.2 Conversational Safety and Controllability

Scope. Conversational safety and controllability target how systems balance harm prevention with user agency during interaction. Safety mechanisms detect crisis, enforce boundaries, defer to human judgment in high-risk areas (Song et al., 2025; Sun et al., 2025; Swinger et al., 2025), while controllability ensures users can influence system behavior, such as selecting modules, controlling depth, or choosing system guidance level (Sun et al., 2025; Wester et al., 2024; Swinger et al., 2025).

Methods and Evaluation. To manage risk, studies commonly adopt modular or layered system designs. High-risk signals are routed to dedicated crisis modules or flagged for human intervention (Sun et al., 2025). Autonomy-in-Middle architectures further restrict AI behavior to alerting while preserving human authority over decisions (Swinger et al., 2025). Evaluation relies on user interviews and expert assessments that examine how systems handle emotionally nuanced and high-risk scenarios, often contrasting rigid filtering with more open-ended generative responses (Song et al., 2025). Crisis-oriented interaction tests should separate ordinary support, ambiguous distress, and self-harm by reporting over-refusals and escalation.

4.3 Empathy and Engagement

Scope. Empathy and engagement target whether users experience the interaction as emotionally appropriate without being misled about the system's actual understanding or clinical role. Literature identifies key tasks such as generating empathetic utterance, non-judgmental tone, and balancing therapeutic alliance with therapeutic structure (Song et al., 2025; Li et al., 2024; Sun et al., 2025).

Methods and Evaluation. Systems operationalize these qualities through script- or protocol-aligned response generation (Sun et al., 2025), structured interaction designs (e.g., multi-choice interfaces or curated narratives) (Sien and McGrenere, 2025; Wester et al., 2024), and expert ratings aligned with clinical competencies (Swinger et al., 2025). Evaluation commonly uses user self-report measures of engagement or emotional experience (Sien and McGrenere, 2025; Wester et al., 2024), comparisons against rule-based or expert-aligned baselines (Sun et al., 2025), and expert coding of behaviors such as empathy and rapport in training contexts (Swinger et al., 2025).

4.4 Transparency

Scope. Transparency targets whether users can understand what the system is doing, including its limitations, decision boundaries, uncertainty, and the basis of its feedback. Literature highlights tasks such as clarifying system limitations, exposing decision logic at appropriate levels, and preventing misconceptions about human-like understanding or authority (Song et al., 2025; Sun et al., 2025).

Methods and Evaluation. Transparency is often implemented through explanations, annotations or labels that surface the rationale behind outputs (Sun et al., 2025; Swinger et al., 2025). User-facing artifacts such as journey maps or labeled strategies can improve interpretability in narrative-based and scripted settings (Sien and McGrenere, 2025; Sun et al., 2025). Evaluations consider whether transparency improves user comprehension, how it affects trust, and whether it introduces unintended interaction incentives (e.g., users adapting responses to satisfy the system rather than therapeutic goals) (Wester et al., 2024; Swinger et al., 2025).

4.5 Tensions and Trade-offs

Scope. Interaction-oriented trustworthiness is inherently multi-dimensional. Competence, safety, empathy, transparency, and controllability cannot always be optimized independently. For example, strong refusal policies can reduce unsafe advice but also damage rapport; empathetic language can support disclosure but may exaggerate perceived understanding; detailed explanations can support informed use but increase cognitive burden; and user control can preserve autonomy while weakening adherence to therapeutic protocols.

Methods and Evaluation. Interaction evaluation should report trade-offs rather than collapse criteria into single trust score. Scenario-based tests can compare ordinary support, ambiguous distress, and crisis signals; expert review assesses therapeutic

appropriateness and boundaries; user studies test if disclosures, explanations, and controls produce calibrated expectations. This aligns with Table 3: interactive cues should communicate capability and limits rather than merely persuade users to trust.

5 AI-Oriented Trustworthiness

AI-oriented trustworthiness concerns whether a mental health AI system operates reliably, safely, and accountably at the model, data, retrieval, evaluation, and infrastructure levels. This layer is not venue-bound: HCI, digital-health, or clinical-adjacent studies are coded here when their main evidence targets model behavior, data governance, privacy, infrastructure, or evaluation validity. Table 10 summarizes the criteria and measures.

5.1 Reliability and Robustness

Scope. Reliability requires consistent model behavior across semantically similar or confusing inputs, and ability to know when it does not know: the system should flag uncertainty when it cannot respond safely. Robustness extends this requirement to long-tail, out-of-distribution, and adversarial conditions where inputs are indirect, emotionally charged, dialectal, or crafted to bypass safeguards (Wang et al., 2023; Alghamdi et al., 2025). Core tasks include risk, distress, and crisis identification, where false confidence can be especially harmful (Yang et al., 2023; Cho et al., 2023).

Methods and Evaluation. Methods span expertise alignment, uncertainty calibration, selective abstention, adversarial training, and multi-run evaluation for unstable LLM outputs (Kang et al., 2024; Reuben et al., 2025; Dhuliawala et al., 2023; Srivastava et al., 2024; Qiu et al., 2025). Evaluations include robustness audits, out-of-distribution stress tests, confidence calibration, refusal/abstention analysis, and error-cost analysis for crisis and high-risk scenarios (Zhu et al., 2025). Stakeholders diverge: AI researchers prioritize consistency; practitioners prioritize avoiding risks; safety researchers prioritize resilience under manipulation (Cho et al., 2023; Liu et al., 2023).

5.2 Safety and Harm Prevention

Scope. Safety concerns the AI's capacity to avoid generating harmful, misleading, or clinically inappropriate content, especially when users disclose distress, self-harm risk, or crisis signals (Cho et al., 2023; Badawi et al., 2025; Na et al., 2025). Operationalization involves empathy, psychological-based attack, risk detection, content filters, and escalation triggers (Chen et al., 2023; Hua et al., 2024; Ozgun et al., 2025; Zhang et al., 2024).

Methods and Evaluation. Model-level safety integrates alignment, guardrails, policy-controlled refusals, escalation, and red-teaming (Chen et al., 2023; Yu et al., 2025; Hua et al., 2024; Liu et al., 2023; Qiu et al., 2025). Evaluation focuses on jailbreak resistance, crisis and ambiguous intent handling, false negatives that miss risk, and false positives that over-block supportive conversation (Alghamdi et al., 2025; Baidal et al., 2025). Stakeholder priorities diverge: safety researchers emphasize resistance, clinicians prioritize crisis recognition, and HCI researchers caution that overly restrictive refusals can undermine rapport (Wang et al., 2025b; Wester et al., 2024; Cho et al., 2023).

5.3 Explainability

Scope. Explainability enables informed use of MHS systems by helping stakeholders understand how a system produces particular outputs (Gollapalli et al., 2023; Gabriel et al., 2024).

Methods and Evaluation. Explainability can be supported through expert alignment, retrieval-augmented generation, tailored memory, and structured rationale generation (Bi et al., 2025; Zhang et al., 2025; Wang et al., 2025a; Gollapalli et al., 2023). However, explanation appearance is not sufficient evidence of trustworthiness: RAG citations may appear correct while failing to faithfully reflect evidence actually used by models (Wallat et al., 2025). Evaluations should therefore inspect rationale quality, evidence faithfulness, criterion anchoring, and whether explanations are understandable to humans (Zhai et al., 2025). In mental health settings, the calibration risk is particularly acute: plausible-sounding explanations can create false reassurance about clinical relevance, motivating faithfulness audits separate from fluency measures.

5.4 Fairness

Scope. Fairness addresses whether systems behave equitably across demographics, cultures, and languages (Gabriel et al., 2024; Lissak et al., 2024). In mental health, expressions of distress vary by culture, identity, dialect, and accessibility, so subgroup performance gaps can translate directly into uneven risk detection and uneven support quality.

Methods and Evaluation. Fairness is assessed by comparing system behavior and outcomes across groups, stress-testing for distributional shift, and auditing biases in data, retrieval, generation, and evaluation pipelines (Gabriel et al., 2024; Qi et al., 2025). Mental-health-specific fairness evaluations should additionally consider matched-risk: (i) differential crisis-detection sensitivity across subgroups; (ii) variance in empathy and refusal behavior across demographic prompts; (iii) bias in LLM-as-judges that may systematically score certain dialects or affective styles as less coherent.

5.5 Privacy and Data Protection

Scope. Privacy determines whether a system protects sensitive disclosures and prevents memorized information. In mental health contexts, perceived or actual exposure risk can undermine disclosure, reliance, and willingness to continue using MHS systems (Baidal et al., 2025). Many corpora derive from public forums or clinical notes with incomplete de-identification, making provenance and consent central to warranted trust (Cho et al., 2023; Qiu and Lan, 2025; Cabrera Lozoya et al., 2025).

Methods and Evaluation. Privacy-preserving techniques include differential privacy, memorization audits, and federated learning (Kwesi et al., 2025; Shin et al., 2023). Evaluations from stakeholders diverge: regulators focus on compliance; security researchers map leakage vectors and extraction attacks (Yu et al., 2025); clinicians expect clinical-standard confidentiality; users equate privacy with safety (Wang et al., 2025b).

Across these criteria, AI-oriented trustworthiness provides necessary but incomplete evidence. Robust models, faithful explanations, and privacy-preserving pipelines must still be translated into interactions that users can understand, such as capability disclosure, escalation, and consent practices.

6 Discussion

Cross-layer calibration and key insights. Trust fails when perceived trust and warranted trustworthiness diverge (see a worked example in Box 6.1). The reviewed corpus shows why this matters, especially for mental health. Empathy and anthropomorphic language can increase disclosure and perceived understanding while also raising over-reliance if boundaries are unclear (Song et al., 2025; Sun et al., 2025; Brunswicker et al., 2025). Crisis handling requires escalation, abstention, and false-negative analysis beyond ordinary user satisfaction (Baidal et al., 2025; Zhang et al., 2024; Alghamdi et al., 2025). Privacy functions as clinical confidentiality and emotional safety, not only data protection, because sensitive disclosures shape help-seeking (Kwesi et al., 2025; Baidal et al., 2025). Thus, mental-health trust calibration cannot be reduced to aggregate model accuracy or general HAI trust measures (Table 3; Table 6).

Implications for NLP research. The objective for trustworthy mental health AI should be calibrated trust, not trust maximization: fluent, empathic, or personalized systems should clarify support boundaries rather than appear more capable than they are. Model-level evaluations are partial evidence and should combine robustness, variability, privacy leakage, fairness, and faithful checks with interactions for ordinary support, ambiguous distress, crisis signals, privacy disclosure, and harmful-advice attempts. Human studies should then measure perceived trust, expectation alignment, and when to seek professional or emergency support. LLM-as-a-judge requires caution because evaluator bias can obscure clinically meaningful differences. LLM judge evaluations should report judge prompts, rubric anchors, bias pre-audits, and expert verification for crisis-sensitive subsets.

Box 6.1: Applying the Framework: A Worked Example

For an empathy-tuned MHS LLM (SoulChat (Chen et al., 2023)), improved empathy ratings are insufficient for trustworthy support. Our three-layer framework asks whether tuning preserves crisis handling, privacy, and robustness (AI-oriented trustworthiness); whether the interface communicates non-clinician status, boundaries, and escalation paths (Interaction-oriented trustworthiness); and whether users' trust and reliance (Human-oriented trust) remain calibrated to those limits. The key question is whether the system becomes safer or merely more trust-inducing. The same check applies to unfaithful RAG citations (Wallat et al., 2025) and to over-refusal (Wester et al., 2024).

7 Conclusion

This survey synthesizes recent language-based mental health AI literature through a three-layer framework distinguishing human-oriented trust, interaction-oriented trustworthiness, and AI-oriented trustworthiness. We show that many trust failures arise from cross-layer and stakeholder misalignment, making calibrated trust more appropriate than trust maximization. The framework supports NLP evaluation, cross-disciplinary governance, and responsible deployment, but should be read as a survey-based evaluation lens rather than a clinical validation of AI mental health system.

Limitations

This survey work has several limitations.

First, our analysis is grounded in academic, peer-reviewed research literature. Real-world deployment practices, proprietary systems, patient advocacy perspectives, insurance and healthcare-system constraints, and the lived experiences of clinicians and users may reveal additional trust concerns that are not fully captured in published studies. Publicly reported incidents involving deployed mental-health or companion-chatbot systems, such as Koko's AI-mediated support for young people [1], companion chatbots like Woebot [2] and Character.AI [3], illustrate why model-level evaluation is insufficient. These cases point to cross-layer failures: unsafe or poorly governed system behavior, inadequate disclosure or boundary communication, inappropriate user reliance, and weak consent or safeguarding mechanisms. They are outside our analytic corpus and serve only as grey literature motivation. Still, they reinforce our central claim: trust failures in mental-health AI often emerge between layers, not within any single layer alone.

Second, despite broad coverage, trustworthy AI literature is rapidly evolving. Relevant work may have been missed or published outside our review window. Our 2021–2025 window was designed for the scoping corpus, while older foundational trust theory is used conceptually rather than exhaustively reviewed. The selected venues were intended to capture major publication channels for language-based mental-health AI trust evidence, but the review may miss relevant work in clinical informatics, psychiatry, broader digital medicine, and general machine-learning venues. Because the search is venue-stratified rather than database-exhaustive and query interfaces varied across sources, the review should be read as a PRISMA-guided scoping synthesis rather than an exhaustive clinical systematic review or meta-analysis.

Third, our synthesis reflects how trust and trustworthiness are operationalized in existing research, which varies across disciplines. Some dimensions may appear more prominent because they are easier to benchmark or measure, rather than because they are intrinsically more important in practice. Coding was conducted as an author-team scoping synthesis using rules and examples reported in Appendix B. We therefore report coding rules, aggregate layer counts, representative examples, and criterion-level synthesis tables, plus a double-coded subset check, , rather than treating the review as a full confirmatory double-coded systematic review. Accordingly, layer counts should be interpreted as descriptive indicators of emphasis within scoped corpus rather than confirmatory estimates of full literature.

Lastly, our three-layer framework is conceptual rather than predictive: it organizes stakeholder perspectives and evaluation requirements but does not offer guarantees about system performance, clinical effectiveness, or user outcomes. Additional validation with clinicians, patient advocates, and deployment partners is needed before framework can be treated as an operational certification tool.

[1] https://kokocares.org/

[2] https://woebothealth.com/

[3] https://character.ai/

Ethical Consideration

This work does not introduce new human-subject experiments or systems for mental health support. However, by synthesizing existing work, it aims to support more ethically grounded development, design, evaluation, and governance of trustworthy AI systems for mental health support. The survey should not be read as clinical validation of current AI mental health systems.

There are several ethical concerns and potential risks to consider. First, AI systems in mental health raise heightened ethical risks due to user vulnerability and the sensitivity of personal disclosures. A primary concern is miscalibrated trust: systems may appear clinically competent or emotionally supportive beyond their actual capabilities, leading to over-reliance, delayed help-seeking, or inappropriate substitution for professional care. Anthropomorphic and empathetic interaction cues can further amplify this risk if not carefully constrained.

Second, privacy and data protection are critical. Mental health AI often processes highly sensitive personal information, creating risks of data leakage, unintended memorization, secondary use without consent, or misuse. Safety validated in offline settings may not hold under real-world deployment.

Third, bias and uneven performance across populations present additional ethical concerns. Models trained on limited or unrepresentative data may underperform for certain groups, dialects, cultures, or clinical presentations, potentially reinforcing existing mental health disparities.

Lastly, survey and evaluation practices themselves pose ethical risks. Synthesizing existing

literature can reproduce disciplinary blind spots if lived experience, non-English work, proprietary deployment failures, or clinical expertise are underrepresented. Over-reliance on automatic metrics or LLM-based evaluators can obscure failure modes and create a false sense of safety. Transparent reporting, human oversight, and clear communication of system limitations are therefore essential.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This work was supported by JST in Japan through CREST (JPMJCR2562), K Program (JPMJKP24C2), FOREST (JPMJFR232R), and BOOST (JPMJBY24D9) grants, and partly by the European Union (EU) under Horizon Europe grant 101120393 (Sec4AI4Sec) and the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) under grants KIC1.VE01.20.004 (HEWSTI) and NWA.1215.18.006 (Theseus).

References

AIHLEG. 2019. Ethics guidelines for trustworthy AI. Technical report.

Mehmet Emin Aktan, Zeynep Turhan, and İlknur Dolu. 2022. Attitudes and perspectives towards the preferences for artificial intelligence in psychotherapy. *Computers in Human Behavior*, 133:107273.

Emad A. Alghamdi, Reem Masoud, Deema Alnuhait, Afnan Y. Alomairi, Ahmed Ashraf, and Mohamed Zaytoon. 2025. AraTrust: An evaluation of trustworthiness for LLMs in Arabic. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8664–8679, Abu Dhabi, UAE. Association for Computational Linguistics.

AMA. 2024. Augmented intelligence development, deployment, and use in health care. Technical report.

Naomi Aoki. 2021. The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence: Evidence from an online experiment. *Computers in Human Behavior*, 114:106572.

APA. 2025. Ethical guidance for ai in the professional practice of health service psychology.

Abeer Badawi, Md Tahmid Rahman Laskar, Jimmy Huang, Shaina Raza, and Elham Dolatabadi. 2025. Position: Beyond assistance – reimagining LLMs as ethical and adaptive co-creators in mental health care. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Miguel Baidal, Erik Derner, and Nuria Oliver. 2025. Guardians of trust: Risks and opportunities for LLMs in mental health. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 11–22, Vienna, Austria. Association for Computational Linguistics.

Guanqun Bi, Zhuang Chen, Zhoufu Liu, Hongkai Wang, Xiyao Xiao, Yuqiang Xie, Wen Zhang, Yongkang Huang, Yuxuan Chen, Libiao Peng, and Minlie Huang. 2025. MAGI: Multi-agent guided interview for psychiatric assessment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24898–24921, Vienna, Austria. Association for Computational Linguistics.

Sabine Brunswicker, Yifan Zhang, Christopher Rashidian, and Daniel W. Linna. 2025. Trust through words: The systemize-empathize-effect of language in task-oriented conversational agents. *Computers in Human Behavior*, 165:108516.

Daniel Cabrera Lozoya, Eloy Hernandez Lua, Juan Alberto Barajas Perches, Mike Conway, and Simon D'Alfonso. 2025. Synthetic empathy: Generating and evaluating artificial psychotherapy dialogues to detect empathy in counseling sessions. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 157–171, Albuquerque, New Mexico. Association for Computational Linguistics.

Hongliu Cao. 2025. Writing style matters: An examination of bias and fairness in information retrieval systems. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM '25, page 336–344, New York, NY, USA. Association for Computing Machinery.

Jiashuo Cao, Yun Suen Pai, Chen Li, Simon Hoermann, and Mark Billinghurst. 2025. "can i have my friend attending with me?": Design implications for using virtual supporters in remote psychotherapy. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183, Singapore. Association for Computational Linguistics.

Oscar Hengxuan Chi, Shizhen Jia, Yafang Li, and Dogan Gursoy. 2021. Developing a formative scale to measure consumers' trust toward interaction with artificially intelligent (AI) social robots in service delivery. *Computers in Human Behavior*, 118:106700.

Young Min Cho, Sunny Rai, Lyle Ungar, João Sedoc, and Sharath Guntuku. 2023. An integrative survey

on mental health conversational agents to bridge computer science and medical perspectives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11346–11369, Singapore. Association for Computational Linguistics.

Ryuhaerang Choi, Taehan Kim, Subin Park, Jennifer G. Kim, and Sung-Ju Lee. 2025. Private yet social: How llm chatbots support and challenge eating disorder recovery. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Shehzaad Dhuliawala, Vilém Zouhar, Mennatallah El-Assady, and Mrinmaya Sachan. 2023. A diachronic perspective on user trust in AI under uncertainty. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5567–5580, Singapore. Association for Computational Linguistics.

EPEUCO. 2024. Artificial intelligence act.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can AI relate: Testing large language model response for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2206–2221, Miami, Florida, USA. Association for Computational Linguistics.

Felix Gille, Laura Maaß, Benjamin Ho, and Divya Srivastava. 2025. From theory to practice: Viewpoint on economic indicators for trust in digital health. *Journal of Medical Internet Research*, 27.

Sujatha Gollapalli, Beng Ang, and See-Kiong Ng. 2023. Identifying Early Maladaptive Schemas from mental health question texts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11832–11843, Singapore. Association for Computational Linguistics.

Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434.

Wenyue Hua, Xianjun Yang, Mingyu Jin, Zelong Li, Wei Cheng, Ruixiang Tang, and Yongfeng Zhang. 2024. TrustAgent: Towards safe and trustworthy LLM-based agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10000–10016, Miami, Florida, USA. Association for Computational Linguistics.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, and 52 others. 2024. Position: Trustllm: Trustworthiness in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20166–20270. PMLR.

Weiwei Huo, Guanghui Zheng, Jiaqi Yan, Le Sun, and Liuyi Han. 2022. Interacting with medical artificial intelligence: Integrating self-responsibility attribution, human–computer trust, and personality. *Computers in Human Behavior*, 132:107253.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 624–635, New York, NY, USA. Association for Computing Machinery.

Migyeong Kang, Goun Choi, Hyolim Jeon, Ji Hyun An, Daejin Choi, and Jinyoung Han. 2024. CURE: Context- and uncertainty-aware mental disorder detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17924–17940, Miami, Florida, USA. Association for Computational Linguistics.

Janne Kauttonen, Rebekah Rousi, and Ari Alamäki. 2025. Trust and acceptance challenges in the adoption of AI applications in health care: Quantitative survey analysis. *Journal of Medical Internet Research*, 27.

Jabari Kwesi, Jiaxun Cao, Riya Manchanda, and Pardis Emami-Naeini. 2025. Exploring user security and privacy attitudes and concerns toward the use of general-purpose llm chatbots for mental health. In *Proceedings of the 34th USENIX Conference on Security Symposium*, SEC '25, USA. USENIX Association.

Jamie Lee, Kyuha Jung, Erin Gregg Newman, Emilie Chow, and Yunan Chen. 2025. Understanding adolescents' perceptions of benefits and risks in health ai technologies through design fiction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80.

Benedikt Leichtmann, Christina Humer, Andreas Hinterreiter, Marc Streit, and Martina Mara. 2023. Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*, 139:107539.

Roy J. Lewicki, Daniel J. McAllister, and Robert J. Bies. 1998. Trust and distrust: New relationships and realities. *Academy of Management Review*, 23(3):438–458.

Anqi Li, Yu Lu, Nirui Song, Shuai Zhang, Lizhi Ma, and Zhenzhong Lan. 2024. Understanding the therapeutic relationship between counselors and clients in

online text-based counseling using LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1280–1303, Miami, Florida, USA. Association for Computational Linguistics.

Q.Vera Liao and S. Shyam Sundar. 2022. Designing for responsible trust in ai systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1257–1268, New York, NY, USA. Association for Computing Machinery.

Shir Lissak, Nitay Calderon, Geva Shenkman, Yaakov Ophir, Eyal Fruchter, Anat Brunstein Klomek, and Roi Reichart. 2024. The colorful future of LLMs: Evaluating and improving LLMs as emotional supporters for queer youth. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2040–2079, Mexico City, Mexico. Association for Computational Linguistics.

Kaifeng Liu and Da Tao. 2022. The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services. *Computers in Human Behavior*, 127:107026.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment. In *Socially Responsible Language Modelling Research*.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, and Maarten de Rijke. 2024. Robust information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 3009–3012, New York, NY, USA. Association for Computing Machinery.

Hanna Luetke Lanfer, Doreen Reifegerste, Annika Berg, Paula Memenga, Eva Baumann, Winja Weber, Julia Geulen, Anne Müller, Andrea Hahne, and Susanne Weg-Remers. 2023. Understanding trust determinants in a live chat service on familial cancer: Qualitative triangulation study with focus groups and interviews in germany. *Journal of Medical Internet Research*, 25.

Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z. Gajos. 2024. Evaluating the experience of lgbtq+ people using large language model based chatbots for mental health support. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Carlotta J. Mayer, Julia Mahal, Daniela Geisel, Eva J. Geiger, Elias Staatz, Maximilian Zappel, Seraina P. Lerch, Johannes C. Ehrenthal, Steffen Walter, and Beate Ditzen. 2024. User preferences and trust in hypothetical analog, digitalized and AI-based medical consultation scenarios: An online discrete choice survey. *Computers in Human Behavior*, 161:108419.

Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *The Academy of Management Review*, 20(3):709–734.

D. Harrison McKnight and Norman L. Chervany. 2001. While trust is cool and collected, distrust is fiery and frenzied: A model of distrust concepts. In *AMCIS 2001 Proceedings*, page 171.

Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. A systematic review on fostering appropriate trust in human-ai interaction: Trends, opportunities and challenges. *ACM J. Responsib. Comput.*, 1(4).

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376, Vienna, Austria. Association for Computational Linguistics.

Mohammad Namvarpour and Afsaneh Razi. 2024. Uncovering contradictions in human-ai interactions: Lessons learned from user reviews of replika. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW Companion '24, page 579–586, New York, NY, USA. Association for Computing Machinery.

NIST. 2023. Nist AI risk management framework (AI RMF 1.0).

Mithat Can Ozgun, Jiahuan Pei, Koen Hindriks, Lucia Donatelli, Qingzhi Liu, and Junxiao Wang. 2025. Trustworthy ai psychotherapy: Multi-agent llm workflow for counseling and explainable mental disorder diagnosis. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 2263–2272, New York, NY, USA. Association for Computing Machinery.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, and 7 others. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.

Yegan Pillay. 2025. Ethical decision-making guidelines for mental health clinicians in the artificial intelligence (ai) era. *Healthcare 2025, Vol. 13,*, 13:3057.

Zhiyang Qi, Takumasa Kaneko, Keiko Takamizo, Mariko Ukiyo, and Michimasa Inaba. 2025. KokoroChat: A Japanese psychological counseling dialogue dataset collected via role-playing by trained counselors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12424–12443, Vienna, Austria. Association for Computational Linguistics.

Huachuan Qiu and Zhenzhong Lan. 2025. PsyDial: A large-scale long-term conversational dataset for mental health support. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21624–21655, Vienna, Austria. Association for Computational Linguistics.

Jiahao Qiu, Yinghui He, Xinzhe Juan, Yimin Wang, Yuhan Liu, Zixin Yao, Yue Wu, Xun Jiang, Ling Yang, and Mengdi Wang. 2025. EmoAgent: Assessing and safeguarding human-AI interaction for mental health safety. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11741–11756, Suzhou, China. Association for Computational Linguistics.

Ansh Rai, Meghan E. Hurley, John Herrington, Eric Alan Storch, Casey J. Zampella, Julia Parish-Morris, Anika Sonig, Gabriel Lázaro-Muñoz, and Kristin Kostick-Quenet. 2025. Stakeholder criteria for trust in artificial intelligence–based computer perception tools in health care: Qualitative interview study. *Journal of Medical Internet Research*, 27.

Massimo Regona, Tan Yigitcanlar, Carol Hon, and Melissa Teo. 2026. Building trust in artificial intelligence: A systematic review through the lens of trust theory. *ACM Comput. Surv.*, 58(9).

Maor Reuben, Ortal Slobodin, Idan-Chaim Cohen, Aviad Elyashar, Orna Braun-Lewensohn, Odeya Cohen, and Rami Puzis. 2025. Assessment and manipulation of latent constructs in pre-trained language models using psychometric scales. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2433–2444, Vienna, Austria. Association for Computational Linguistics.

Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. 1998. Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, 23(3):393–404.

Jaemin Shin, Hyungjun Yoon, Seungjoo Lee, Sungjoon Park, Yunxin Liu, Jinho Choi, and Sung-Ju Lee. 2023. FedTherapist: Mental health monitoring with user-generated linguistic expressions on smartphones via federated learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11971–11988, Singapore. Association for Computational Linguistics.

Sang-Wha Sien and Joanna McGrenere. 2025. A gentle introduction to mental health through storytelling: Design and evaluation of digital human library. *Proc. ACM Hum.-Comput. Interact.*, 9(2).

Inhwa Song, Sachin R Pendse, Neha Kumar, and Munmun De Choudhury. 2025. The typing cure: Experiences with large language model chatbots for mental health support. *Proc. ACM Hum.-Comput. Interact.*, 9(7).

Aseem Srivastava, Smriti Joshi, Tanmoy Chakraborty, and Md Shad Akhtar. 2024. Knowledge planning in large language models for domain-aligned counseling summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17775–17789, Miami, Florida, USA. Association for Computational Linguistics.

Xin Sun, Jan de Wit, Zhuying Li, Jiahuan Pei, Abdallah El Ali, and Jos A. Bosch. 2025. Script-strategy aligned generation: Aligning llms with expert-crafted dialogue scripts and therapeutic strategies for psychotherapy. *Proc. ACM Hum.-Comput. Interact.*, 9(7).

Nathaniel Swinger, Cynthia M. Baseman, Myeonghan Ryu, Saeed Abdullah, Christopher W. Wiese, Andrew M. Sherrill, and Rosa I. Arriaga. 2025. There's no "i" in teammait: Impacts of domain and expertise on trust in ai teammates for mental health work. *Proc. ACM Hum.-Comput. Interact.*, 9(2).

Anja Thieme, Maryann Hanratty, Maria Lyons, Jorge Palacios, Rita Faia Marques, Cecily Morrison, and Gavin Doherty. 2023. Designing human-centered ai for mental health: Developing clinically relevant applications for online cbt treatment. *ACM Trans. Comput.-Hum. Interact.*, 30(2).

Roel W. Visser, Tobias M. Peters, Ingrid Scharlau, and Barbara Hammer. 2025. Trust, distrust, and appropriate reliance in (X)AI: A conceptual clarification of user trust and survey of its empirical evaluation. *Cognitive Systems Research*, 91:101357.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not faithfulness in retrieval augmented generation attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, page 22–32, New York, NY, USA. Association for Computing Machinery.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Manias Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. Decodingtrust: a comprehensive assessment of trustworthiness in gpt models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Ming Wang, Peidong Wang, Lin Wu, Xiaocui Yang, Daling Wang, Shi Feng, Yuxin Chen, Bixuan Wang,

and Yifei Zhang. 2025a. AnnaAgent: Dynamic evolution agent system with multi-session memory for realistic seeker simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23221–23235, Vienna, Austria. Association for Computational Linguistics.

Yimeng Wang, Yinzhou Wang, Kelly Crace, and Yixuan Zhang. 2025b. Understanding attitudes and trust of generative ai chatbots for social anxiety support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Joel Wester, Henning Pohl, Simo Hosio, and Niels van Berkel. 2024. "this chatbot would never...": Perceived moral agency of mental health chatbots. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Claire Woodcock, Brent Mittelstadt, Dan Busbridge, and Grant Blank. 2021. The impact of explanations on layperson trust in artificial intelligence–driven symptom checker apps: Experimental study. *Journal of Medical Internet Research*, 23(11).

Min Wu, Nanxi Wang, and Kum Fai Yuen. 2023. Deep versus superficial anthropomorphism: Exploring their effects on human trust in shared autonomous vehicles. *Computers in Human Behavior*, 141:107614.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Dana You, Yuwon Kim, and Kyoungwon Seo. 2025. Tailored virtual agent guidance for stress management: Comparing directive and non-directive approaches by alexithymia status. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Seounmi Youn and S. Venus Jin. 2021. "in a.i. we trust?" the effects of parasocial interaction and technopian versus luddite ideological views on chatbot-based customer relationship management in the emerging "feeling economy". *Computers in Human Behavior*, 119.

Miao Yu, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pan, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, Bo An, and Qingsong Wen. 2025. A survey on trustworthy llm agents: Threats and countermeasures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 6216–6226, New York, NY, USA. Association for Computing Machinery.

Wei Zhai, Nan Bai, Qing Zhao, Jianqiang Li, Fan Wang, Hongzhi Qi, Meng Jiang, Xiaoqin Wang, Bing Xiang Yang, and Guanghui Fu. 2025. MentalGLM series: Explainable large language models for mental health analysis on Chinese social media. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13588–13603, Suzhou, China. Association for Computational Linguistics.

Linhai Zhang, Ziyang Gao, Deyu Zhou, and Yulan He. 2025. Explainable depression detection in clinical interviews with personalized retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9927–9944, Vienna, Austria. Association for Computational Linguistics.

Zaibin Zhang, Yongting Zhang, Lijun Li, Hongzhi Gao, Lijun Wang, Huchuan Lu, Feng Zhao, Yu Qiao, and Jing Shao. 2024. PsySafe: A comprehensive framework for psychological-based attack, defense, and evaluation of multi-agent system safety. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15202–15231, Bangkok, Thailand. Association for Computational Linguistics.

Jiukai Zhao, Yuqi Yang, Juanxia Miao, Xue Wang, Dianjun Qi, and Shuang Zang. 2025. Factors associated with the level of trust in health information robots among the general population from a socioecological model perspective: Network analysis. *Journal of Medical Internet Research*, 27.

Xi Zheng, Zhuoyang Li, Xinning Gui, and Yuhan Luo. 2025. Customizing emotional support: How do individuals construct and interact with llm-powered chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, Qingqing Long, Yefeng Zheng, and Xian Wu. 2025. Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769, Vienna, Austria. Association for Computational Linguistics.

Appendix

A AI Usage Disclosure

We used AI tools in a supportive and limited role. Specifically, GPT-5 was used for language editing (e.g., improving clarity and conciseness). All findings and figures are based on our own data and results. The survey scoping, literature review, data analysis, and final content were conducted and verified by the authors.

B Methodology of Literature Review

To provide a transparency for a cross-disciplinary survey, we conducted a PRISMA-guided scoping review (Page et al., 2021). We use PRISMA principles to structure screening and reporting, but the study is not a database-exhaustive clinical systematic review or meta-analysis. Instead, it targets recent peer-reviewed work on language-based AI systems for mental health support and uses an auditable screening and coding procedure to synthesize how trust and trustworthiness are operationalized across research communities.

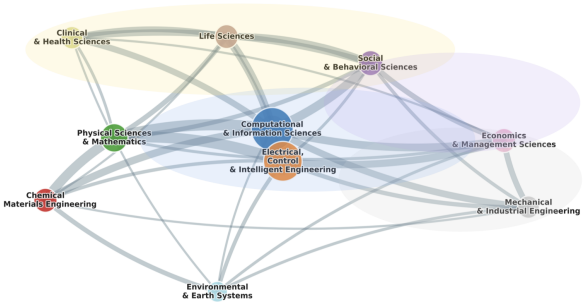


Figure 3: Discipline network from a broad 1,706-paper mapping corpus (2021–2025) related to trust in mental health AI. The network visualizes the interdisciplinary landscape. Visualization source code is released in an anonymous repository.

Scope and Search Strategy. The scoping corpus covers papers published between Jan 2021 and Nov 2025. The date range applies only to corpus selection: it was chosen to capture the recent shift toward LLM-powered conversational AI and rapidly changing evaluation practices in mental health AI. Foundational trust theory before 2021 is still used for conceptual grounding throughout the paper.

The 1,706-paper corpus in Figure 3 is a broader bibliometric mapping corpus used to visualize disciplinary connections around trust and mental health AI. It is not identical to the analytic corpus used for coding. The PRISMA-guided scoping review below uses a narrower stratified search of 735 records and retains 61 unique papers that meet the language-based mental-health AI and trust-related inclusion criteria. Thus, Figure 3 motivates the interdisciplinary landscape, whereas this appendix defines the corpus used for synthesis.

We use mental health support as an umbrella term for screening, psychoeducation, conversational support, coaching, therapy-related training, and psychological self-help. These applications are not clinically equivalent: low-risk psychoeducation primarily requires transparency and privacy safeguards, whereas crisis-adjacent or psychotherapy-adjacent systems require stronger escalation, abstention, human oversight, and accountability.

We searched across three disciplinary strata motivated by the network in Figure 3: digital health and behavioral science outlets (JMIR, CHB), NLP/AI/security venues (ACL, EMNLP, NAACL, COLING, IJCAI, USENIX), and HCI venues and journals (ACM CHI, ACM CSCW, TOCHI, IJHCS). These strata were search entry points for retrieval coverage, not analytic categories or proxies for trust layers (Table 1). Search terms combined mental-health concepts with AI-system and trust-related concepts, including variants of *AI*, *LLM*, *dialogue system*, *chatbot*, *mental health*, *psychological support*, *safety*, *trust*, *trustworthiness*, *privacy*, *reliability*, and *explainability*. Searches were conducted over titles, abstracts, and main text when available, excluding reference sections.

Search Implementation and Reproducibility. Searches were implemented as venue-stratified queries rather than a single database-wide query. We searched title, abstract, and available full-text fields, excluding reference sections. The query templates combined three concept groups: (mental health OR psychological support OR therapy OR counseling OR wellbeing) AND (AI OR LLM OR chatbot OR dialogue system OR conversational agent) AND (trust OR trustworthiness OR safety OR privacy OR reliability OR robustness OR fairness OR explainability OR transparency). Because search interfaces varied across proceedings sites, publisher portals, and manual venue browsing, query implementation was standardized at the concept-group level rather than as byte-identical query strings across all sources. We document the search strata, query templates, eligibility criteria, screening counts, final corpus size, coding rules, and representative coding examples to support a PRISMA-guided scoping synthesis. This

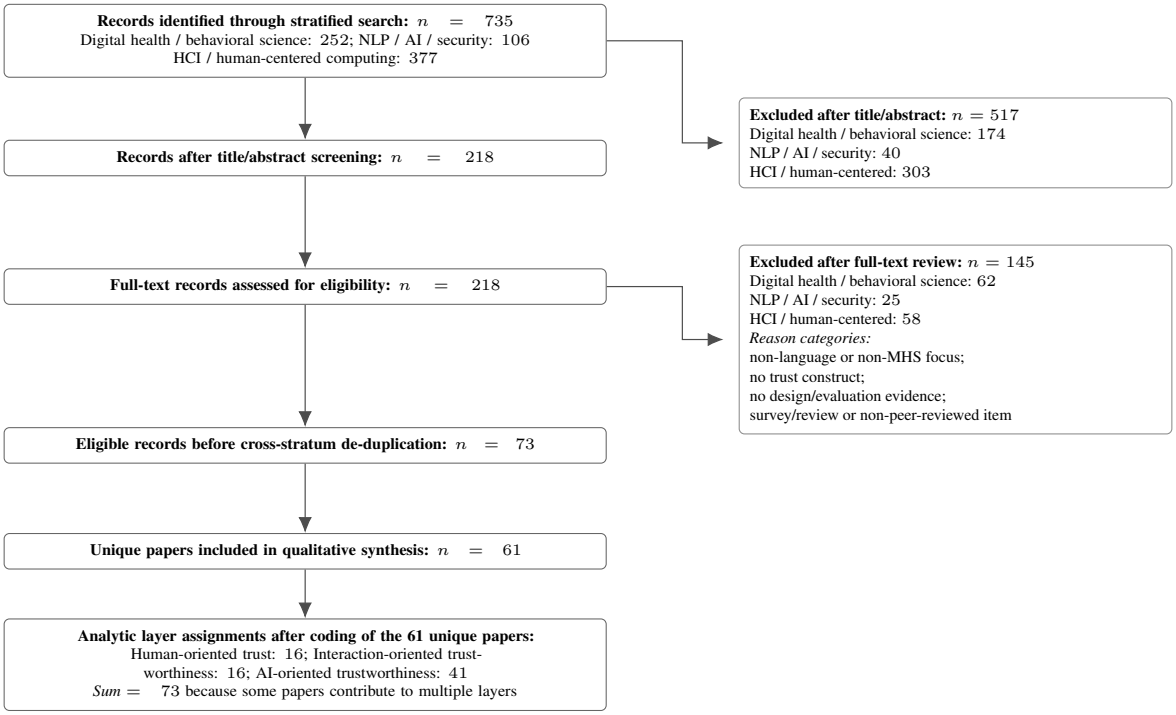


Figure 4: PRISMA-style flow for the PRISMA-guided scoping review. Search strata were used only for retrieval coverage and were not analytic layer labels. The full-text exclusion box lists eligibility categories applied during review. The final corpus contains 61 unique papers; layer assignments are reported as 16/16/41 because some papers contributed evidence to more than one layer.

Search entry stratum	**Venues / sources**	**Purpose in retrieval**
Digital health / behavioral science	JMIR, CHB	Capture health, disclosure, user trust, perceived competence, and clinical-adjacent evidence relevant to mental health support.
NLP / AI / security	ACL, EMNLP, NAACL, COLING, IJCAI, USENIX	Capture model, retrieval, robustness, safety, privacy, fairness, and evaluation evidence from language-based AI systems.
HCI / human-centered computing	ACM CHI, ACM CSCW, TOCHI, IJHCS	Capture interaction design, lived experience, transparency, controllability, crisis routing, and conversational safety evidence.

Table 1: Search entry strata used for retrieval coverage. It reports where papers were searched from. These strata are not analytic layer labels and do not determine how papers are assigned to the three-layer framework.

Analytic layer	**Assignments**	**Assignment criterion**	**Example constructs**
Human-oriented trust	$n = 16$	The primary evaluation target is a user's subjective trust state, expectation, or reliance outcome.	Perceived trust, trust propensity, disclosure, user characteristics, reliance behavior.
Interaction-oriented trustworthiness	$n = 16$	The primary evaluation target is user-visible system behavior or interaction design.	Empathy, transparency, controllability, crisis routing, uncertainty communication, boundary setting.
AI-oriented trustworthiness	$n = 41$	The primary evaluation target is a model-, data-, retrieval-, evaluation-, privacy-, or infrastructure-level property.	Robustness, safety, privacy, fairness, explanation faithfulness, evaluator reliability.

Table 2: Analytic layer assignments are produced after coding. The final corpus contains 61 unique papers; assignments sum to 73 because secondary labels were used when a paper provided substantive evidence for more than one layer. This corresponds to 12 additional layer assignments beyond the 61 unique papers, not 12 additional papers. These counts describe the distribution of evidence within our scoped corpus rather than field-wide prevalence estimates; they are used to identify recurring emphases and gaps, not to make statistical claims about the literature.

Layer	Core evaluation question	Typical methods	Risk if omitted	Implication for NLP systems
Human-oriented trust	Do users' trust and reliance match what the system can safely provide?	Validated trust scales, interviews, reliance behavior, expectation checks, longitudinal use and disengagement indicators.	Persuasive systems may create over-trust, while opaque but safer systems may create under-trust or abandonment.	Report user-facing outcomes, reliance, and expectation calibration rather than only model scores.
Interaction-oriented trustworthiness	Do interaction behaviors communicate capability, limits, uncertainty, and safety boundaries appropriately?	Scenario-based dialogue tests, expert review, crisis-routing evaluation, transparency and controllability studies, user comprehension checks.	Empathy, fluency, or anthropomorphic cues may inflate trust without corresponding capability; excessive refusals may damage rapport.	Evaluate responses in realistic mental-health scenarios, including uncertainty communication and boundary setting.
AI-oriented trustworthiness	Does the underlying model or pipeline deserve trust under normal, long-tail, and adversarial conditions?	Robustness and fairness audits, calibration and abstention tests, privacy leakage tests, RAG faithfulness checks, LLM-evaluator bias audits.	Benchmark performance may overstate safety, faithfulness, privacy, or clinical reliability.	Treat RAG, LLM-as-a-judge, and aggregate metrics as partial evidence that must be linked to interaction and user outcomes.

Table 3: Integrated trust evaluation rubric. The rubric synthesizes reviewed practices and identifies evaluation gaps so that evaluations target calibrated trust rather than trust maximization. It is an evaluation checklist, not an empirically validated scoring instrument, certification instrument, or guarantee of safety. In practice, calibration can be measured as the gap between user trust or reliance and demonstrated capability, e.g., rates of following incorrect advice (over-reliance) and rejecting correct support (under-reliance) in scenario tests, making the rubric usable as a three-layer report card. A minimal *calibrated-trust* report should include one signal from each layer: *L1:* humans perceived trust or expectation-alignment measure; *L2:* boundary comprehension, uncertainty disclosure, and escalation behavior under ordinary, ambiguous, and crisis-like scenarios; *L3:* robustness/abstention, privacy leakage, fairness, and faithful-grounding checks. Claims of trustworthy support should report cross-layer consistency rather than a single trust score or metric.

Representative evidence	Primary construct	Evaluation target	Layer label rule	Why not venue-based
Trust scales and user studies (Chi et al., 2021; Kauttonen et al., 2025; Liu and Tao, 2022)	Perceived trust, trust propensity, user characteristics	Subjective trust response and reliance tendency	Primary: human-oriented trust	The label follows the measured construct. A user study in any venue is human-oriented when its central evidence is users' trust or reliance.
Scripted or protocol-aligned chatbot interactions (Song et al., 2025; Sun et al., 2025; Wester et al., 2024)	Empathy, controllability, transparency, therapeutic alignment	Observable conversational behavior and user-facing controls	Primary: interaction-oriented trustworthiness	The label follows the interaction evidence, regardless of whether the paper appears in HCI, AI, or health venues.
ACL/NAACL user-facing counseling studies (Lissak et al., 2024; Qi et al., 2025)	Subgroup support quality, counseling dialogue quality, client feedback	User-visible support behavior and subgroup-sensitive response quality	Primary: interaction-oriented when evidence concerns support behavior; secondary: AI-oriented when it audits model or data properties	An ACL/NAACL paper is not automatically AI-oriented; the label follows the evaluation target, not the venue.
RAG and explanation faithfulness studies (Wallat et al., 2025; Zhang et al., 2025; Gollapalli et al., 2023)	Faithfulness, attribution quality, rationale inspection	Model or pipeline evidence for whether explanations deserve trust	Primary: AI-oriented trustworthiness; secondary: interaction-oriented when explanations are user-facing	A technical paper can receive an interaction label only when it evaluates how explanations are exposed to or interpreted by users.
Safety, jailbreak, and crisis-handling evaluations (Hua et al., 2024; Baidal et al., 2025; Alghamdi et al., 2025)	Harm prevention, crisis escalation, adversarial robustness	Worst-case system behavior and high-risk interaction outcomes	Primary: AI-oriented trustworthiness; secondary: interaction-oriented for crisis routing	The primary label reflects system safety evidence; a secondary label is added only when user-visible routing or escalation behavior is substantively evaluated.
Privacy-preserving mental-health AI (Shin et al., 2023; Kwesi et al., 2025)	Privacy, data protection, memorization risk	Sensitive-data handling and leakage resistance	Primary: AI-oriented trustworthiness	The evidence concerns data and infrastructure protection, which constrains warranted trust even when no user study is present.

Table 4: Representative coding examples showing how papers were mapped from trust-related constructs to framework layers. Primary labels follow the main evaluation target; secondary labels are used only when a paper provides substantive evidence for another layer. These examples document the mapping logic for a PRISMA-guided scoping synthesis and are not intended as a confirmatory inter-rater reliability report.

Stakeholder / perspective	Definition in this survey	Typical evidence	How users or clients enter the framework
Clients, patients, and end users	Primary trustors whose subjective trust, reliance, disclosure, and help-seeking outcomes are affected by mental health AI.	Self-report scales, interviews, participatory design, preference tasks, usage and reliance indicators.	Represented centrally in the human-oriented layer rather than treated as a publication-venue category.
Psychotherapy practitioners and clinical researchers	Licensed therapists, clinical psychologists, psychiatrists acting in therapeutic roles, and mental-health researchers concerned with therapeutic fidelity and client safety.	Therapeutic alliance, clinical appropriateness, crisis safety, user vulnerability, and human oversight evidence.	Bring client needs, therapeutic norms, confidentiality, and responsibility into the interpretation of warranted trust.
HCI researchers	Researchers who study user-centered design, interaction patterns, lived experience, and empirical use of AI-mediated mental-health support.	User studies, interviews, interface evaluations, design probes, scenario studies, and interaction analyses.	Translate lived-experience evidence into interaction-level criteria such as transparency, controllability, empathy, and boundary communication.
AI/NLP researchers	Researchers who develop or evaluate model, retrieval, generation, classification, and automatic evaluation methods.	Benchmarks, robustness tests, RAG faithfulness analyses, LLM-as-a-judge evaluations, calibration, and fairness audits.	Provide evidence about whether system capabilities can justify the level of trust encouraged by user-facing interaction.
Safety and security researchers	Researchers who examine adversarial misuse, jailbreaks, privacy leakage, memorization, and worst-case failures.	Red-teaming, jailbreak audits, membership inference, extraction attacks, and crisis scenario stress tests.	Identify failure modes that are especially consequential when vulnerable users disclose sensitive information.
Regulators and standards bodies	Institutions and professional bodies that formalize minimum requirements for risk management, accountability, privacy, and oversight.	Regulatory frameworks, professional ethics guidance, risk management standards, and accountability principles.	Define external constraints under which user trust may be considered warranted or institutionally acceptable.

Table 5: Stakeholder definitions used in this survey. Categories are based on evaluation paradigms and roles in trust calibration, not on publication venue boundaries.

General HAI issue	Mental-health-specific pressure	Representative reviewed evidence	Consequence for this framework
Reliance	Users may be distressed, vulnerable, or unsure whether to seek professional help.	User trust and reliance studies (Chi et al., 2021; Woodcock et al., 2021; Mayer et al., 2024); crisis-oriented safety work (Hua et al., 2024; Baidal et al., 2025).	Evaluate appropriate reliance, help-seeking, and disengagement, not only adoption or engagement.
Empathy	Warm language can invite disclosure while also inflating perceived competence or clinical authority.	Anthropomorphism and empathy studies (Wu et al., 2023; Liu and Tao, 2022; Brunswicker et al., 2025); scripted support systems (Sun et al., 2025; Song et al., 2025).	Test whether empathy calibrates expectations or exaggerates what the system can safely provide.
Transparency	Users may mistake AI support for therapy, diagnosis, or professional oversight.	Explanation and transparency studies (Leichtmann et al., 2023; Woodcock et al., 2021; Sun et al., 2025; Wester et al., 2024).	Require boundary disclosure, uncertainty communication, and clear escalation pathways.
Privacy	Disclosures may resemble clinical-confidential information and include trauma, stigma, or crisis details.	Privacy and leakage evaluations (Shin et al., 2023; Kwesi et al., 2025; Yu et al., 2025).	Evaluate consent, retention, leakage, memorization, and perceived privacy safety.
Robustness	Distress and self-harm cues can be ambiguous, indirect, culturally varied, or adversarially phrased.	Robustness, calibration, and adversarial tests (Kang et al., 2024; Dhuliawala et al., 2023; Alghamdi et al., 2025).	Use scenario stress tests for ordinary support, ambiguous distress, and crisis cases.
Fairness	Expressions of distress vary by culture, language, identity, and access to care.	Fairness and subgroup analyses (Gabriel et al., 2024; Lissak et al., 2024; Qi et al., 2025).	Audit differential risk detection, support quality, refusal behavior, and escalation across groups.
Explanation	Plausible explanations may create false reassurance in high-stakes support conversations.	RAG and explanation-faithfulness evaluations (Wallat et al., 2025; Zhang et al., 2025; Gollapalli et al., 2023).	Evaluate faithfulness, user comprehension, and clinical relevance rather than explanation fluency alone.

Table 6: How mental health support changes general human-AI trust evaluation. The table links domain-specific pressures to representative evidence from the reviewed corpus and explains why this survey emphasizes trust calibration across human, interaction, and AI-oriented layers rather than perceived trust alone. These pressures operate at the framework level, not only in the analysis details: in mental health, the trust relationship is itself part of the intervention, so miscalibrated trust directly harms disclosure, help-seeking, and crisis safety rather than only task accuracy. This separates our aim from general (X)AI trust and reliance surveys (Visser et al., 2025; Liao and Sundar, 2022), which target general task settings.

venue-stratified design improves cross-disciplinary coverage but may miss relevant work in clinical informatics, psychiatry, broader digital medicine, AI ethics, and general machine-learning venues.

Inclusion and Exclusion Criteria. As shown in Figure 4, we included peer-reviewed papers that met all three criteria: (1) the paper studied, evaluated, or analyzed a language-based AI system or AI-mediated workflow for mental health support, screening, coaching, therapy-related training, or psychological self-help; (2) the paper explicitly addressed at least one trust-related construct or criterion, even if it did not use the word "trust" directly; and (3) the paper provided empirical findings, system design, evaluation methodology, technical analysis, or clinically relevant design evidence. Trust-related constructs included user trust, reliance, therapeutic alliance, perceived competence, interaction quality, empathy, transparency, controllability, reliability, robustness, fairness, privacy, safety, crisis handling, and evaluation validity.

We excluded papers that focused only on non-health domains, mental health without AI, non-language or sensor-only interventions outside the scope of this survey, purely theoretical arguments without design or evaluation evidence, tutorials, extended abstracts, non-peer-reviewed articles, and prior surveys. Prior surveys were excluded from the scoping corpus but used as related work when they helped position our contribution. Broader health or digital-health studies were included only when they directly informed transferable mechanisms relevant to mental health AI trust, such as sensitive disclosure, user vulnerability, clinician accountability, or privacy expectations.

Screening and Selection. Screening followed a multi-stage process summarized in Figure 4. First, titles and abstracts were screened to remove records that clearly lacked an AI system, mental-health relevance, or any trust-related criterion. Second, full texts were reviewed against the inclusion and exclusion criteria. Full-text exclusions were applied for reasons such as non-language-based AI, no mental-health support focus, no trust-related construct, survey or review status, non-peer-reviewed or tutorial format, purely theoretical scope, or absence of design/evaluation evidence. We report these exclusion categories to clarify eligibility decisions, while treating the synthesis as a scoping review rather than a reason-by-reason exclusion audit. Third, retained papers were de-duplicated across strata and checked for cross-layer relevance. The final corpus contains 61 unique papers. Because a single paper can contribute evidence to more than one layer, the layer-level assignments are reported as Human-oriented $= 16$, Interaction-oriented $= 16$, and AI-oriented $= 41$ (sum $= 73$), corresponding to 12 additional layer assignments beyond the 61 unique papers (Table 2). This distribution is itself a finding. The AI-oriented evidence base is approximately 2.5× the human-oriented base, despite the centrality of user trust in mental health support. Even when we restrict to papers published in 2023–2025 (the post-LLM-conversational-MHS period), the imbalance persists (AI-oriented = 24, Human-oriented = 9, Interaction-oriented = 11). This asymmetry is not solely a venue-selection artifact: digital-health journals (JMIR, CHB) were searched specifically to capture human-oriented evidence, yet AI/NLP venues yielded substantially more language-based MHS trust studies. We interpret this distribution as evidence of the calibration gap our framework articulates: the community has invested heavily in measuring system-level properties (robustness, fairness, privacy) while comparatively underinvesting in measuring whether users' subjective trust and reliance track those properties in mental health contexts.

Coding and Synthesis Procedure. After selection, six authors participated in extraction and open coding, followed by author-team consensus. For each paper, we extracted the mental-health support context, AI system type, trust-related construct, evaluation target, evaluation method, stakeholder perspective, and evidence type. Initial open codes included perceived trust, reliance, therapeutic alliance, empathy, transparency, controllability, crisis handling, privacy, robustness, fairness, explanation faithfulness, and evaluation validity. We assigned a primary layer based on the paper's main evaluation target and allowed secondary labels only when a paper provided substantive evidence for another layer. For example, a study of chatbot crisis response could be coded as interaction-oriented when the central evidence concerned conversational behavior, with a secondary AI/safety label if it also evaluated adversarial or high-risk failure cases.

The trust/trustworthiness distinction was theory-informed, and the three-layer framework was refined through open coding and axial grouping. During coding, constructs repeatedly clustered around three evaluation targets: (1) subjective user atti-

Document	Agency	Year	Key Focus of Trustworthiness
NIST AI Risk Management Framework (AI RMF 1.0) (NIST, 2023)	National Institute of Standards and Technology	2023	Focuses on seven characteristics of trustworthy AI to manage risks and promote responsible use: Valid and Reliable, Safe, Secure and Resilient, Accountable and Transparent, Explainable and Interpretable, Privacy-Enhanced, and Fair with Harmful Bias Managed. Accountability and Transparency are emphasized as relating to all other characteristics and internal processes.
Artificial Intelligence Act (AI Act) (EPEUCO, 2024)	European Parliament and Council	2024	Promotes human-centric and trustworthy AI while ensuring a high level of protection of health, safety, and fundamental rights. Mandatory requirements for high-risk systems include Technical Robustness and Safety, Accuracy and Cybersecurity, high standards for Data Governance, and strong measures for Transparency and Accountability. It is informed by the seven AI HLEG ethical requirements.
Augmented Intelligence Development, Deployment, and Use in Health Care (AMA, 2024)	American Medical Association	2024	Mandates that health care AI be designed, developed, and deployed in a manner that is ethical, equitable, responsible, accurate, transparent, and evidence-based. Key concerns are Transparency/Disclosure of AI use at the point of care, Bias Mitigation/Equity, alignment of Liability and Accountability with the entity best positioned to mitigate harm, and robust Data Privacy and Cybersecurity.
Ethical Guidance for AI in the Professional Practice of Health Service Psychology (APA, 2025)	American Psychological Association Ethics Committee	2025	Aligns AI use with fundamental psychological ethical principles (Beneficence and Nonmaleficence, Integrity, Justice, and Respect for People's Rights and Dignity). Key areas include Transparency and Informed Consent (including the right to opt out), Mitigating Bias and Promoting Equity, Data Privacy and Security (e.g., HIPAA compliance), Accuracy and Misinformation Risks, and ensuring Human Oversight and Professional Judgment.
Ethical Decision-Making Guidelines for Mental Health Clinicians in the Artificial Intelligence (AI) Era (Pillay, 2025)	Journal of Healthcare	2025	Proposes a framework based on five pillars derived from professional ethical codes and AI guidelines: Autonomy and Informed Consent; Beneficence and Non-Malfeasance; Confidentiality, Privacy, and Transparency; Justice, Fairness, and Inclusiveness; and Fidelity, Professional Integrity, and Accountability. Emphasizes that AI must augment, not replace, human clinical care.

Table 7: Representative regulatory and professional guidance documents relevant to trustworthy AI in health and mental-health contexts.

Dimension	Criterion	What is evaluated	Typical evidence	Representative studies
What trust is	Ability / competence	Whether users believe the system can provide reliable, useful, and appropriate support.	Likert scales; interviews; behavioral monitoring.	(Chi et al., 2021; Luetke Lanfer et al., 2023; Brunswicker et al., 2025; Gille et al., 2025)
	Benevolence	Whether users perceive the system as supportive, caring, and aligned with their interests.	Qualitative interviews and focus groups.	(Luetke Lanfer et al., 2023; Rai et al., 2025)
	Integrity / reliability	Whether users view the system as honest, dependable, and behaviorally consistent.	Interviews; behavioral monitoring; trust scales.	(Luetke Lanfer et al., 2023; Gille et al., 2025; Rai et al., 2025; Leichtmann et al., 2023)
	Individual perception	How trust propensity, perceived agency, and subjective interpretation shape trust judgments.	Likert scales; qualitative inquiry; behavioral monitoring.	(Luetke Lanfer et al., 2023; Chi et al., 2021; Brunswicker et al., 2025)
Who trusts	User characteristics	Attitudes toward AI, personality traits, familiarity, prior use, and perceived social support.	Pre-/post-study questionnaires; self-developed instruments.	(Kauttonen et al., 2025; Zhao et al., 2025; Huo et al., 2022)
	AI literacy	How AI knowledge affects reliance, caution, and willingness to accept system recommendations.	Questionnaires and trust/reliance measures.	(Woodcock et al., 2021)
	Human oversight / control	Whether shared control or human-in-the-loop oversight changes perceived accountability.	Likert scales; behavioral monitoring.	(Liu and Tao, 2022; Aoki, 2021; Mayer et al., 2024)
How trust forms	Anthropomorphism	Human-like language, social presence, empathy cues, and perceived emotional understanding.	Validated scales; manipulated system behaviors; interaction logs.	(Chi et al., 2021; Wu et al., 2023; Liu and Tao, 2022; Brunswicker et al., 2025)
	Explainability	Whether explanations clarify outputs, decisions, limitations, and appropriate reliance.	Explanation manipulations; comprehension and confidence measures.	(Leichtmann et al., 2023; Woodcock et al., 2021)

Table 8: **Human-oriented trust** criteria, evidence types, and representative studies. The redesigned table groups measures by what trust is, who trusts, and how trust is formed, emphasizing that subjective trust and behavioral reliance should be interpreted separately.

Criterion	Interaction focus	Typical methods	Calibration evidence	Representative studies
Competence and reliability	Response usefulness, contextual appropriateness, therapeutic alignment, and protocol adherence.	User interviews; case analysis; expert-authored scripts; checklist scoring.	Whether the interaction supports appropriate reliance rather than merely fluent conversation.	(Song et al., 2025; Sun et al., 2025; Swinger et al., 2025; Wester et al., 2024)
Conversational safety and controllability	Crisis routing, boundary setting, user agency, module choice, and escalation to human judgment.	Layered system design; Autonomy-in-the-Middle; expert review; user feedback; usage analytics.	Whether safety mechanisms reduce harm without hiding limits or removing meaningful user control.	(Song et al., 2025; Sun et al., 2025; Swinger et al., 2025; Wester et al., 2024)
Communication style	Tone, role framing, response structure, and guidance style during sensitive disclosure.	Modular dialogue design; expert judgment; comparative response review.	Whether communication preserves autonomy and therapeutic boundaries while remaining usable.	(Swinger et al., 2025; Song et al., 2025)
Transparency	Capability disclosure, limitation statements, explanation cues, and uncertainty communication.	Structured explanations; visual or strategy cues; comprehension checks.	Whether users understand what the system can and cannot do.	(Sun et al., 2025; Sien and McGrenere, 2025; Wester et al., 2024)
Empathy and engagement	Emotionally appropriate language, non-judgmental tone, rapport, and sustained engagement.	Protocol-aligned response generation; narrative interfaces; expert behavioral coding; self-report.	Whether rapport supports disclosure without exaggerating understanding or clinical authority.	(Sun et al., 2025; Sien and McGrenere, 2025; Swinger et al., 2025; Wester et al., 2024)

Table 9: **Interaction-oriented trustworthiness** criteria and evaluation practices. The table foregrounds how user-visible behavior should communicate capability, limits, uncertainty, and safeguards instead of simply increasing perceived trust.

Criterion	Methods	Evaluation signals	Trust-calibration role	Representative studies
Reliability and robustness	Fine-tuning; prompting; calibration; uncertainty quantification; multi-run evaluation.	Accuracy, BLEU/BERTScore, out-of-distribution tests, confidence calibration, and abstention analysis.	Shows whether the system can support reliance under ordinary and long-tail inputs.	(Yang et al., 2023; Kang et al., 2024; Dhuliawala et al., 2023; Alghamdi et al., 2025)
Safety and harm prevention	RLHF/alignment training; rule-based guardrails; escalation protocols; red-teaming.	Jailbreak resistance, toxicity scoring, crisis scenario tests, refusal rates, and false positive/negative analysis.	Constrains whether supportive interaction can be considered safe in high-risk situations.	(Liu et al., 2023; Hua et al., 2024; Yu et al., 2025; Alghamdi et al., 2025)
Privacy and data protection	Differential privacy; federated learning; memorization auditing; restricted fine-tuning.	Extraction attacks, membership inference, memorization probes, and sensitive-data leakage audits.	Determines whether disclosure and continued use can be warranted in sensitive contexts.	(Shin et al., 2023; Kwesi et al., 2025; Yu et al., 2025)
Explainability and faithfulness	RAG; chain-of-thought prompting; multi-agent workflows; memory; structured rationales.	Human evaluation, rationale quality, evidence tagging, attribution faithfulness, and criterion anchoring.	Separates useful explanations from plausible but unfaithful rationales or citations.	(Yang et al., 2023; Ozgun et al., 2025; Gollapalli et al., 2023; Zhai et al., 2025; Zhang et al., 2025; Bi et al., 2025)
Fairness	Bias auditing; demographic-aware prompting; data balancing; counterfactual testing.	Demographic parity, subgroup performance, empathy variance, and group fairness metrics.	Checks whether trustworthiness evidence generalizes across users, dialects, cultures, and risk profiles.	(Gabriel et al., 2024; Lissak et al., 2024; Baidal et al., 2025; Huang et al., 2024)

Table 10: **AI-oriented trustworthiness** criteria, methods, and evaluation signals. The redesigned table connects technical evidence to calibrated trust claims rather than treating aggregate performance as sufficient proof of trustworthiness.

tudes, expectations, and reliance outcomes; (2) user-visible interaction mechanisms such as explanations, disclosures, empathy, controllability, and crisis routing; and (3) model- or infrastructure-level properties such as robustness, privacy, safety, fairness, and evaluation validity. These groups were then compared against stakeholder perspectives to identify recurring alignments and tensions. Coding was performed by the author team applying the layer rules above, with disagreements resolved through consensus discussion. To check coding consistency, a subset of the corpus was independently double-coded, and primary layer assignment reached Cohen's $\kappa = .73$, indicating substantial agreement. We provide representative coding examples in Table 4, stakeholder definitions in Table 5, an evaluation-oriented summary in Table 3, aggregate layer counts, and criterion-level synthesis tables to make the mapping logic explicit.

C Summary of regulations and ethical standards.

Table 7 summarizes key regulatory frameworks and ethical guidelines governing trustworthy AI in mental and general health contexts, highlighting how agencies and professional bodies operationalize trustworthiness through requirements on safety, transparency, accountability, privacy, and human oversight.

D Literature for "Human-oriented" Trust

Human-oriented trust captures how users perceive, form, and adjust trust toward AI that provides mental health support. Prior work operationalizes this layer along three complementary dimensions: what trust is, who trusts, and how trust is formed through perceived system cues. The literature primarily relies on subjective measures such as Likert-scale questionnaires, supplemented by interviews, stated-preference tasks, and behavioral monitoring. Table 8 summarizes the frequent evaluation approaches and representative studies in this layer.

Classic accounts define trust as a willingness to accept vulnerability based on positive expectations of ability, benevolence, and integrity (Mayer et al., 1995; Rousseau et al., 1998; Lee and See, 2004; Hoff and Bashir, 2015); these dimensions underlie the perceived-trust measures reviewed here. Distrust is a distinct construct built on active negative expectations, not the mere absence of trust (Lewicki et al., 1998; McKnight and Chervany, 2001; Visser et al., 2025): a user may trust a system for support while distrusting its privacy or crisis handling, which our layers treat as separate calibration signals rather than one trust score.

E Literature for "Interaction-oriented" Trustworthiness

Interaction-oriented trustworthiness captures how trust is shaped through observable system behavior during interaction. Across the literature, this layer is operationalized through criteria such as competence, communication style, transparency, empathy and engagement, and controllability. Studies evaluate these criteria using a combination of expert review, user-centered methods, and structured protocol assessments. Table 9 summarizes the key criteria in this layer, along with the dominant methods and evaluation practices identified in our review.

F Literature for "AI-oriented" Trustworthiness

AI-oriented trustworthiness focuses on whether mental health AI systems meet model- or system-level requirements for safety, reliability, and responsible deployment, independent of any specific interaction design. Prior work operationalizes this layer through criteria such as reliability and robustness, safety and harm prevention, privacy and data protection, explainability, and fairness. Table 10 summarizes the dominant technical methods and evaluation practices used to assess these criteria in the literature.