Indonesian Journal of Medical Linguistics
Vol. 1, No. 2, June 2026, pp. 156~169
ISSN: 3163-3676, DOI:

Empathy by design? evaluating relational language and emotional safety in Indonesian mental health chatbots

Handono Fatkhur Rahman
Universitas Indonesia, Indonesia

Article Info

Article history:

Received Apr 19, 2026
Revised May 27, 2026
Accepted Jun 30, 2026

Keywords:

artificial intelligence;
cultural-pragmatic alignment;
emotional safety;
mental health chatbots;
relational empathy.

ABSTRACT

Background: Mental health chatbots are increasingly positioned as accessible support tools in Indonesia, yet their capacity to provide care depends not only on technical responsiveness but also on relational language, emotional safety, and cultural-pragmatic fit. **Objective:** This study aims to evaluate how Indonesian mental health chatbot discourse constructs empathy, manages emotional risk, and aligns with local communicative norms of help-seeking. **Method:** Using a qualitative discourse-pragmatic design, this study analysed 23 public text units from chatbot-related platforms, institutional pages, government health resources, and scholarly documents through a relational empathy matrix, emotional safety audit, and cultural-pragmatic alignment checklist. **Results:** Findings show that empathy is most visible through relational presence and autonomy support, while affective recognition, validation, and explicit non-judgmental stance appear less consistently. Emotional safety is strongest when chatbot discourse foregrounds boundary transparency, diagnosis restraint, and human-in-the-loop referral, but weaker when crisis wording, refusal style, and consent framing remain underdeveloped. **Implication:** Cultural alignment is supported by accessible Indonesian and locally credible referral pathways, although stigma, family pressure, multilingual diversity, and regional variation receive limited explicit attention. **Novelty:** This study offers a novel framework for evaluating mental health chatbot empathy as a relationally expressed, emotionally bounded, and culturally situated design practice.

Corresponding Author:

Handono Fatkhur Rahman
Universitas Indonesia
Kampus UI Depok, Jl. Margonda Raya, Depok, Jawa Barat, 16424, Indonesia
Email: handono.hfc@gmail.com

1. INTRODUCTION

Mental health support is increasingly mediated by conversational technologies at a moment when psychological distress has become both widespread and unevenly served. Globally, 970 million people were living with a mental disorder in 2019, with anxiety and depression among the most common conditions (World Health Organization, 2022). In Indonesia, the first National Adolescent Mental Health Survey reported that one in three adolescents, equivalent to 15.5 million young people, experienced a mental health problem within twelve months, while one in twenty, or 2.45 million, met criteria for a mental disorder (I-NAMHS, 2022). This demand has encouraged the uptake of mental health applications, online counselling platforms, and AI-enabled chatbots as low-threshold forms of support. Yet the same accessibility creates a linguistic and ethical problem: when distress is addressed through automated text, emotional safety depends not only on information accuracy but also on how care is worded, bounded, and relationally sustained.

Prior research has examined mental health chatbots from several perspectives, including clinical effectiveness, user engagement, therapeutic alliance, safety, empathy, and ethical governance. Reviews show that conversational agents may support psychoeducation, self-monitoring, behavioural activation, and access to care, although evidence remains heterogeneous across populations, designs, and outcome measures [1], [2], [3], [4]. More recent studies have shifted attention toward large language models, empathy generation, safety metrics, and human–AI collaboration in mental health support [5], [6], [7], [8]. However, less attention has been given to empathy as a culturally situated linguistic practice in Indonesian chatbot discourse. Existing scholarship often evaluates whether systems are usable, acceptable, or clinically promising, but rarely asks how Indonesian relational language, pragmatic norms, referral boundaries, stigma, and emotional vulnerability are managed within chatbot responses.

This study examines how Indonesian mental health chatbots construct relational language and emotional safety in publicly accessible digital mental health discourse. It asks four interrelated questions. First, how are empathy, validation, and non-judgmental stance linguistically produced in chatbot-facing mental health communication? Second, how do chatbot texts manage emotional safety when users express distress, uncertainty, shame, or care-seeking hesitation? Third, how are boundaries communicated when chatbots must avoid diagnosis, excessive reassurance, or therapeutic overclaiming? Fourth, how far do Indonesian mental health chatbot ecologies align with culturally sensitive healthcare communication, especially where mental distress is shaped by family expectations, stigma, religious language, and uneven access to professional care? To answer these questions, this study combines discourse-pragmatic analysis, emotional safety audit, and cultural-pragmatic alignment assessment across primary chatbot-related materials, secondary academic sources, and contextual public health documents.

This study advances the argument that empathy in mental health chatbots should not be treated as a surface feature of warm tone or affective wording. In automated mental health interaction, empathy is a designed relational achievement that depends on recognition, validation, linguistic restraint, referral clarity, and respect for user autonomy. Ethical concerns raised in recent literature show that chatbots may become harmful when they appear too human, obscure their limits, encourage dependency, provide unsafe reassurance, or withdraw support abruptly in moments of vulnerability [9], [10], [11], [12]. Indonesian healthcare linguistics further suggests that compassionate communication must account for comprehension, directive speech, multilingual barriers, illness narratives, and culturally patterned expressions of care [13], [14], [15]. This study therefore tests whether chatbot empathy is best understood as relational safety by design rather than as emotional simulation by default.

2. LITERATURE REVIEW

2.1. Relational empathy in mental health chatbots

Relational empathy in mental health chatbots refers to a communicative capacity to acknowledge distress, validate experience, and sustain supportive interaction without claiming human consciousness or clinical authority. In clinical communication, empathy is not identical to friendliness; it involves perspective-taking, affective recognition, and response appropriateness. In conversational AI, however, empathy is often operationalised as sentiment alignment, emotionally supportive wording, or user-perceived warmth [16], [17], [6]. This creates a conceptual tension: chatbot empathy may be designed as linguistic performance, evaluated as user experience, or treated as quasi-therapeutic presence. This study treats empathy as relational discourse, not machine feeling.

Relational empathy can be assessed through several linguistic indicators. First, affective recognition concerns whether a chatbot accurately names or reflects user emotion without exaggeration. Second, validation concerns whether distress is legitimised without minimisation or premature advice. Third, non-judgmental stance concerns avoidance of blame, moralisation, and stigma. Fourth, relational presence concerns continuity, warmth, and conversational responsiveness without false humanisation. Fifth, autonomy support concerns whether users are invited to choose next steps rather than being directed coercively. These indicators draw on work on empathic response generation, human–AI collaboration, and compassionate mental health communication [18], [8], [7], [19], [14].

2.2. Emotional safety

Emotional safety denotes conditions in which vulnerable users are not exposed to avoidable psychological, informational, or relational harm during chatbot interaction. Unlike clinical safety, which often focuses on adverse events, diagnosis, or treatment risk, emotional safety concerns how automated responses handle distress, dependency, refusal, uncertainty, and crisis escalation. Mental health chatbots raise distinctive concerns because they may appear caring while lacking accountability, contextual memory, and professional judgement [9], [10]. Studies of chatbot harms show that users may experience over-attachment, misrecognition, unsafe reassurance, or abandonment when systems fail relationally [11], [20]. Safety is therefore interactional.

Indicators of emotional safety include boundary clarity, crisis sensitivity, referral quality, refusal design, privacy transparency, and avoidance of therapeutic overclaiming. Boundary clarity requires chatbots to state limits without withdrawing care abruptly. Crisis sensitivity requires escalation when users disclose self-harm, hopelessness, or danger. Referral quality requires directing users toward human support without inventing unavailable services. Refusal design matters because abrupt rejection can intensify vulnerability, especially when users disclose shame or despair. Recent evaluation frameworks therefore move beyond accuracy toward safety metrics, trajectories, stress testing, and ethical safeguards [21], [5], [22], [23], [24], [25].

2.3. Cultural-pragmatic alignment

Cultural-pragmatic alignment refers to fit between chatbot language and local norms of meaning, care, authority, and help-seeking. In Indonesian mental health communication, distress may be expressed through indirectness, religious vocabulary, family obligation, shame, or somatic complaint. A chatbot that translates global therapeutic scripts into Indonesian without pragmatic sensitivity may sound polite yet fail relationally. Healthcare linguistics shows that comprehension, directive speech, bad-news delivery, multilingual barriers, and illness narratives shape how care is interpreted [26], [27], [28], [13], [15]. Cross-cultural chatbot research also suggests that emotional support varies across language communities and help-seeking cultures [29], [30].

Cultural-pragmatic alignment can be examined through language accessibility, register choice, stigma sensitivity, culturally legible empathy, and locally credible referral. Language accessibility asks whether responses avoid jargon and remain comprehensible to users with different literacy levels. Register choice concerns balance between formal Indonesian, conversational warmth, and respectful distance. Stigma sensitivity concerns whether chatbots avoid framing distress as weakness, sin, failure, or irrationality. Culturally legible empathy concerns whether responses acknowledge family, community, religion, and social pressure without stereotyping them. Locally credible referral concerns whether users are guided toward realistic support pathways. This study therefore positions Indonesian chatbot empathy as a threefold construct: relationally expressed, emotionally bounded, and pragmatically situated.

3. METHOD

The unit of analysis in this study is a publicly accessible chatbot-related document segment, including app-store descriptions, official platform pages, institutional service reports, government health pages, and scholarly descriptions of Indonesian mental health chatbot systems. This unit was selected because publicly available chatbot transcripts remain limited, while platform-facing language still reveals how empathy, safety, and service boundaries are designed for users. The corpus consists of 18 retained documents and 23 linked text units, divided into primary, secondary, and contextual data. Primary data represent Indonesian chatbot or digital mental health platforms; secondary data provide technical, linguistic, and ethical frameworks; contextual data establish national safety, referral, and health-communication benchmarks.

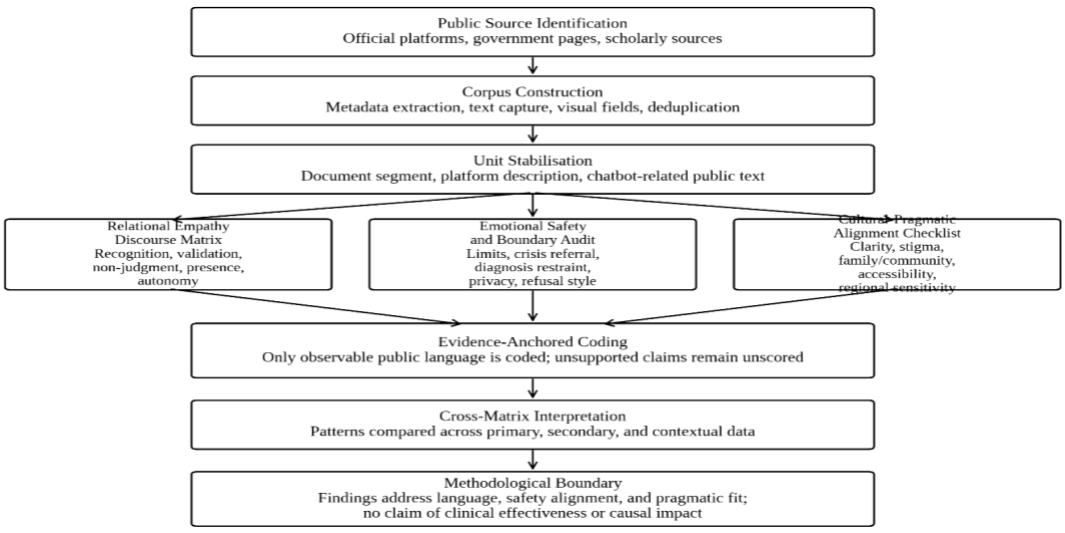

Figure 1. Operational matrix of analysis

Table 1. Dataset composition

Code	Data Type	Source	Location	Period	Number	Characteristics	Status
IND-MHCB-P01	Primary	Riliv Google Play	Indonesia	2026	1	Mental health app description, counselling, meditation, privacy language	Retained
IND-MHCB-P02	Primary	Riliv for Education	Indonesia	n.d.	1	AI Buddy, student mental-health questions, personalization claims	Retained
IND-MHCB-P03	Primary	Riliv service page	Indonesia	n.d.	1	Safe-space language, online psychologist service benchmark	Retained
IND-MHCB-P04	Primary	UNAIR/PUSPAGA PusBuddy	Surabaya	2026	1	WhatsApp chatbot assessment, psychoeducation, referral pathway	Retained
IND-MHCB-P05	Primary	THERAPi Google Play	Indonesia-language listing	2026	1	AI companion/listener description, non-medical limitation statement	Retained
IND-MHCB-P06	Primary	Ayo Sehat Kemenkes	Indonesia	n.d.	1	Official health portal context and public health language	Retained
IND-MHCB-S01	Secondary	MyCare AI article	Indonesia	2025	1	Dual-AI mental health chatbot, emotion classification, generative support	Retained
IND-MHCB-S02	Secondary	Budidarma/Garuda record	Indonesia	2024	1	LSTM-based mental health chatbot development	Retained
IND-MHCB-S03	Secondary	Rabit journal article	Indonesia	2026	1	LLM, emotion detection, retrieval-augmented generation, stress testing	Retained
IND-MHCB-S04	Secondary	JMIR Mental Health	Global	2024	1	Empathic conversational-agent design and evaluation review	Retained
IND-MHCB-S05	Secondary	JMIR	Global	2022	1	Language use in health conversational agents	Retained
IND-MHCB-S06	Secondary	Digital Health	Global	2023	1	Ethical issues in mental health chatbots	Retained
IND-MHCB-S07	Secondary	VERA-MH/ArXiv	Global	2026	1	AI safety evaluation framework for mental health	Retained
IND-MHCB-C01	Context	Kemenkes news release	Indonesia	2025	1	AI self-diagnosis risk and youth vulnerability	Retained
IND-MHCB-C02	Context	Healing119/Kemenkes	Indonesia	2025	1	Suicide prevention and emotional-support referral pathway	Retained
IND-MHCB-C03	Context	Ayo Sehat Kemenkes	Indonesia	2024	1	Adolescent mental health education and coping language	Retained
IND-MHCB-C04	Context	Ayo Sehat resource page	Indonesia	n.d.	1	Mental health resource links and referral taxonomy	Retained
IND-MHCB-C05	Context	Ayo Sehat Kemenkes	Indonesia	2025	1	Mental health maintenance and self-care language	Retained

This study uses a qualitative discourse-pragmatic design supported by structured content coding. This design is appropriate because the research problem concerns how empathy and emotional safety are linguistically constructed, not whether chatbots produce clinical improvement. The design follows recent work that evaluates conversational agents through language use, empathy, ethical safeguards, and safety metrics rather than through technical accuracy alone [31], [6], [5], [24]. It also draws on Indonesian healthcare linguistics, where comprehension, directive speech, compassionate communication, and patient narratives are treated as interactional conditions of care.

Information sources were selected from three layers. Primary sources include public-facing Indonesian digital mental health platforms, app-store listings, official service pages, and institutional descriptions of chatbot-enabled services. Secondary sources include peer-reviewed or indexed academic work on mental health chatbots, empathic conversational agents, language use, ethics, and safety evaluation. Contextual sources include Indonesian government health portals and referral information that define acceptable boundaries for mental health communication. This layered design prevents platform claims from being read in isolation. It allows this study to compare chatbot-facing language with broader standards of healthcare communication, crisis referral, and public mental health education.

Data collection followed a transparent public-corpus protocol. First, searches were conducted using Indonesian and English keywords such as "chatbot kesehatan mental," "AI kesehatan mental Indonesia," "mental health chatbot Indonesia," "Riliv AI Buddy," "PusBuddy," "THERAPi," "Healing119," and "self-diagnosis AI kesehatan jiwa." Second, sources were screened for public accessibility, relevance, date, institutional traceability, and connection to empathy, emotional support, referral, or safety. Third, metadata were extracted, including title, source, date, publisher, URL, data type, language, content type, and methodological role. Fourth, duplicate entries were removed through source-title-date matching. Text, quotation status, context, visual fields, and screenshot references were then stored.

Data analysis proceeded through three linked instruments. The Relational Empathy Discourse Matrix assessed affective recognition, validation, non-judgmental stance, relational presence, and autonomy support. The Emotional Safety and Boundary Audit assessed boundary transparency, crisis escalation, diagnosis restraint, refusal style, and privacy or consent framing. The Cultural-Pragmatic Alignment Checklist assessed Indonesian clarity, stigma sensitivity, family or community alignment, accessibility, and multilingual or regional sensitivity. Each code was applied only when observable textual evidence was available. This procedure keeps interpretation evidence-anchored and avoids unsupported claims about clinical effectiveness, user outcomes, or causal impact.

4. RESULTS

4.1. Relational empathy as linguistic design in Indonesian mental health chatbot discourse

Public-facing Indonesian mental health chatbot discourse constructs empathy less through sustained therapeutic exchange than through recognisable signals of availability, support, and care access. The coded corpus shows that relational presence and autonomy support each appear in seven of twenty-three text units, while affective recognition appears in six, validation in five, and non-judgmental stance in four. This distribution suggests that empathy is primarily designed as a service-facing relational promise: users are invited to approach, disclose, receive tips, access counselling, or move toward professional help. Such wording corresponds with recent work on empathic conversational agents, which treats empathy as a designed interactional feature rather than as evidence of machine feeling [6], [17], [16].

The dominant pattern is the concentration of relational language around presence and choice. Platform descriptions, institutional reports, and public health pages repeatedly use vocabulary associated with companionship, listening, safe space, referral, daily support, and professional access. By contrast, affective recognition and validation appear more cautiously, often as descriptions of system capability or support orientation rather than as observable response sequences. Non-judgmental stance is the least visible indicator, usually inferred from privacy, anonymity, and safe-space framing rather than explicit anti-stigma language. This pattern indicates that Indonesian chatbot-related discourse is more developed in presenting supportive access than in specifying how distress is linguistically received, normalised, or protected from shame.

Table 2. Distribution of relational empathy indicators in public Indonesian mental health chatbot corpus

Main Category	Subcategory	Indicator	Frequency	Percentage	Data Variation	Example/Excerpt	Data Codes	Interpretation
Relational empathy	Affective recognition	Naming or reflecting emotional state without exaggeration	6	26.1%	Primary platform descriptions; Indonesian technical articles; public health education	Emotion classification; emotion detection; sharing feelings; emotional burden	TXT-P02-01; TXT-P03-01; TXT-P05-01; TXT-S01-01; TXT-S03-01; TXT-C05-01	Recognition appears mostly as platform capability or help-seeking language, not as extended dialogic attunement.
Relational empathy	Validation	Legitimising distress without minimisation or premature correction	5	21.7%	Counselling pages; AI-listener positioning; adolescent mental-health guidance	Safe and comfortable space; listener; trusted conversation; support when overwhelmed	TXT-P01-02; TXT-P03-01; TXT-P05-01; TXT-C03-01; TXT-C05-01	Validation is present as supportive framing, although rarely supported by observable turn-by-turn response evidence.
Relational empathy	Non-judgmental stance	Avoiding blame, shame, stigma, and moralising explanations	4	17.4%	Privacy/anonymity claims; safe-space discourse; trusted-conversation advice	Anonymous/private storytelling; safe space; trusted conversation	TXT-P01-02; TXT-P03-01; TXT-C03-01; TXT-C05-01	Non-judgment is implied through privacy and safety language, but explicit anti-stigma phrasing remains limited.
Relational empathy	Relational presence	Maintaining warm supportive positioning without replacing human care	7	30.4%	App descriptions; institutional chatbot reports; companion/listener language; support services	Counselling app; AI Buddy; safe space; chatbot assessment; companion/friend/listener; Healing119 support	TXT-P01-01; TXT-P02-01; TXT-P03-01; TXT-P04-01; TXT-P05-01; TXT-C02-01; TXT-C05-01	Relational presence is most visible because public-facing sources foreground availability, companionship, and support.
Relational empathy	Autonomy support	Offering options, referral, self-care, and professional help without coercion	7	30.4%	Counselling access; daily tips; assessment/referral pathways; public health warnings	Online counselling; tailored daily tips; referral to psychologists; professional help; not a medical substitute	TXT-P01-01; TXT-P02-01; TXT-P04-02; TXT-P05-02; TXT-C01-01; TXT-C03-01; TXT-C05-01	Autonomy is strongest when supportive language is paired with referral, self-care choice, or explicit service boundaries.

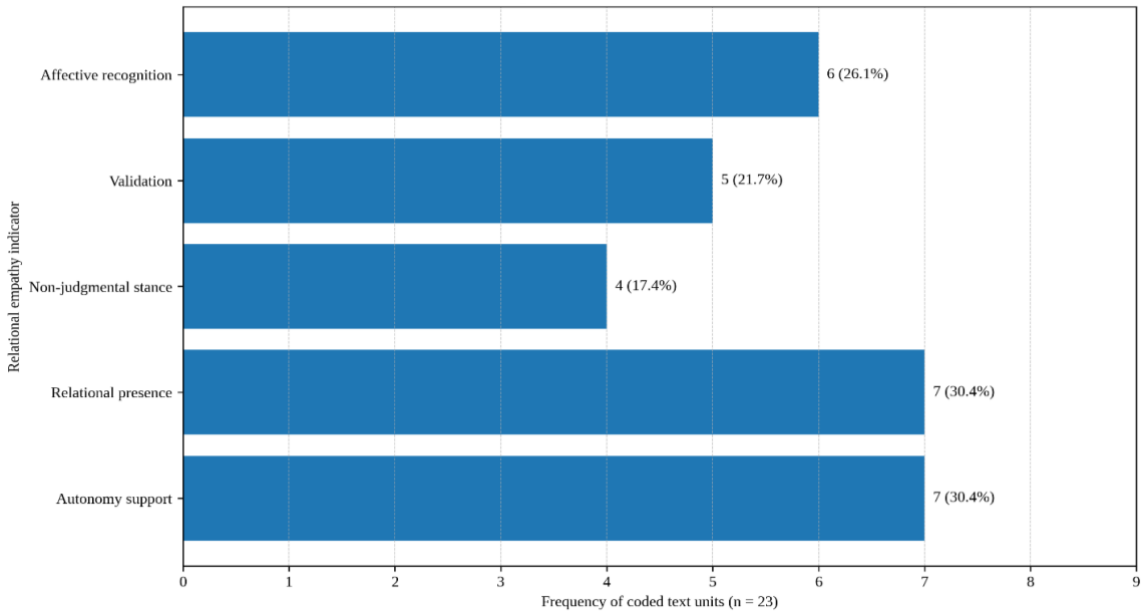

Figure 2. Relational empathy indicators

Theoretically, these findings position chatbot empathy as a layered discourse practice: it is relational when it offers presence, ethical when it preserves autonomy, and clinically cautious when it avoids replacing professional care. However, empathy remains incomplete if it is reduced to warmth, companionship, or motivational phrasing. Studies on human–AI mental health support warn that users may interpret emotionally fluent systems as more capable than they are, especially when care language is not balanced by boundary clarity and referral pathways [8], [9], [10]. This study therefore interprets Indonesian chatbot empathy as a design claim that becomes credible only when affective recognition, validation, non-judgment, relational presence, and autonomy support operate together.

4.2. Emotional safety, boundary management, and human referral in chatbot-mediated support

Emotional safety in the corpus is organised around boundary-setting rather than autonomous therapeutic authority. Boundary transparency and human-in-the-loop referral each appear in eight of twenty-three text units, followed by diagnosis restraint in seven units. This pattern indicates that Indonesian mental health chatbot discourse tends to frame AI as an entry point, companion, screening channel, or information pathway rather than as a replacement for professional mental health care. Such positioning aligns with ethical concerns in recent mental health chatbot literature, where emotionally fluent systems may become risky when they obscure their limits, overstate competence, or encourage users to substitute automated interaction for professional support [9], [10], [22].

The most dominant safety pattern is the coupling of supportive access with human referral. Public texts repeatedly mention online psychologists, operator review, official services, professional help, and trusted conversation. Diagnosis restraint is also prominent, especially where sources warn against AI-based self-diagnosis or state that an application is not a medical substitute. By contrast, crisis escalation, privacy and consent framing, and refusal style appear less frequently. Crisis escalation is visible through Healing119, PusBuddy referral, and safety-testing frameworks, but public-facing materials rarely display how a chatbot would respond linguistically to acute suicidal ideation. Refusal style is the least developed indicator because limitation statements appear as disclaimers, not as relationally sensitive interactional responses.

Table 3. Distribution of emotional safety and boundary indicators in public Indonesian mental health chatbot corpus

Main Category	Subcategory	Indicator	Primary Units	Secondary Units	Context Units	Frequency	Percentage	Data Variation	Example/Excerpt	Data Codes	Interpretation
Emotional safety	Boundary transparency	Stating service limits, non-clinical scope, or escalation boundaries	2	3	3	8	34.8%	App limitation statements; institutional referral descriptions; ethics and safety literature; public-health warnings	Not a medical substitute; should not replace professional support; safety stress testing; chatbot-to-psychologist referral	TXT-P04-02; TXT-P05-02; TXT-S03-02; TXT-S06-01; TXT-S07-01; TXT-C01-01; TXT-C03-01; TXT-C05-01	Boundary transparency is visible when AI support is framed as access, screening, or assistance rather than autonomous therapy.
Emotional safety	Crisis escalation	Directing high-risk users toward human, emergency, or specialised support pathways	1	2	2	5	21.7%	PusBuddy referral flow; safety testing frameworks; Healing119 and official mental-health resource links	Operator review; referral to psychologists; Healing119 voice call and WhatsApp chat; suicidal-ideation safety assessment	TXT-P04-02; TXT-S03-02; TXT-S07-01; TXT-C02-01; TXT-C04-01	Crisis escalation appears as referral infrastructure, but public texts provide limited evidence of real-time high-risk response wording.
Emotional safety	Diagnosis restraint	Avoiding self-diagnosis, diagnostic certainty, or replacement of professional evaluation	1	3	3	7	30.4%	Non-medical-device disclaimer; AI self-diagnosis warnings; safety and ethics sources	Not a medical device; AI self-diagnosis can mislead; professional help remains necessary	TXT-P05-02; TXT-S03-02; TXT-S06-01; TXT-S07-01; TXT-C01-01; TXT-C03-01; TXT-C05-01	Diagnostic restraint is one of the clearer safety features because several sources explicitly separate AI support from clinical judgment.
Emotional safety	Refusal and limitation style	Communicating inability or limitation without abrupt rejection of user distress	1	2	1	4	17.4%	App disclaimers; ethics literature; AI self-diagnosis warning	Not a substitute for medical help; ethical concern over professional boundaries; avoid misleading AI diagnosis	TXT-P05-02; TXT-S06-01; TXT-S07-01; TXT-C01-01	Limitation language is present, but public materials rarely show how refusal is softened in vulnerable interactional moments.
Emotional safety	Privacy and consent framing	Clarifying anonymity, private disclosure, history storage, or user-data sensitivity	4	1	0	5	21.7%	Anonymous storytelling; safe-space language; AI Buddy history storage; ethics literature	Anonymous/private storytelling; safe space; storing history; privacy and autonomy concerns	TXT-P01-01; TXT-P01-02; TXT-P02-01; TXT-P03-01; TXT-S06-01	Privacy appears most strongly in platform-facing language, although consent implications of stored interaction history require deeper scrutiny.
Emotional safety	Human-in-the-loop referral	Connecting chatbot-mediated support with psychologists, operators, official services, or trusted human support	3	0	5	8	34.8%	Online psychologist services; PusBuddy operator review; Healing119; government self-care and help-seeking guidance	Online psychologist counseling; operator review; referral to PUSPAGA psychologists; trusted conversation; professional help	TXT-P01-01; TXT-P03-01; TXT-P04-02; TXT-C01-01; TXT-C02-01; TXT-C03-01; TXT-C04-01; TXT-C05-01	Human referral is a central safety mechanism, indicating that Indonesian chatbot discourse often legitimises AI support through connection to human care.

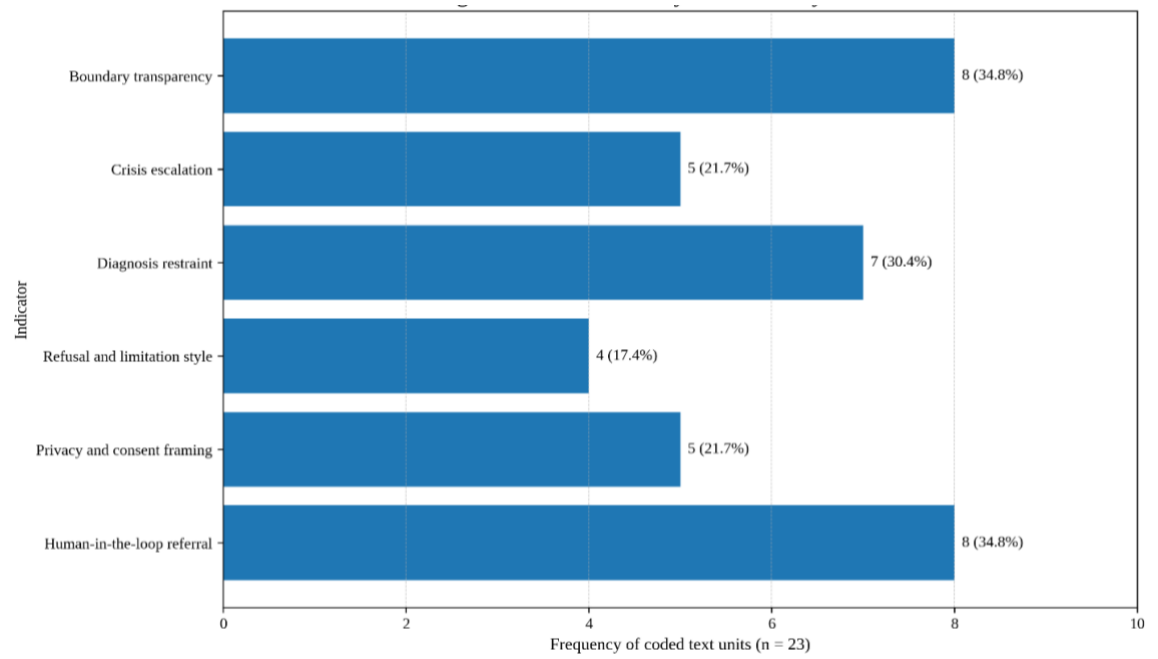

Figure 3. Emotional safety and boundary indicators

These patterns suggest that emotional safety is treated as an infrastructural and ethical boundary rather than a purely affective quality of chatbot language. This finding extends work on AI mental health safety, which argues that evaluation must include escalation, privacy, diagnosis restraint, and relational harm, not only conversational fluency or user satisfaction [32], [5], [23], [24], [25]. For this study, emotional safety is therefore understood as a boundary-mediated form of care: a chatbot becomes safer not by sounding more human, but by knowing when to defer, disclose limits, protect user agency, and connect distress to accountable human support.

4.3. Cultural-pragmatic alignment in Indonesian mental health chatbot discourse

Indonesian mental health chatbot discourse aligns most strongly with culture through accessible language and locally credible referral. Each of these indicators appears in eight of twenty-three text units, showing that cultural fit is achieved less through explicit cultural explanation than through recognisable pathways of help-seeking. Public-facing sources tend to avoid dense psychiatric terminology and instead rely on ordinary Indonesian expressions such as counselling, tips, support, safe space, trusted conversation, and professional help. This pattern matters because healthcare communication research in Indonesia emphasises that comprehension, pragmatic clarity, and compassionate wording shape whether users can interpret care as usable and legitimate [26], [13], [14].

A second pattern concerns moderated warmth. Register moderation and family or community relevance each appear in six text units, while stigma-sensitive framing appears in five. These indicators show that chatbot discourse often recognises Indonesian users as socially embedded rather than purely individual help-seekers. Student support, family assessment, trusted conversation, and official referral pathways position distress within everyday relational settings. However, explicit attention to stigma, shame, moral judgement, or culturally specific family pressure remains limited. Multilingual and regional sensitivity appears in only three text units, indicating a notable gap between Indonesia's sociolinguistic diversity and the relatively standardised Indonesian used in public chatbot-facing materials.

Table 4. Distribution of cultural-pragmatic alignment indicators in public Indonesian mental health chatbot corpus

Main Category	Subcategory	Indicator	Primary Units	Secondary Units	Context Units	Frequency	Percentage	Data Variation	Example/Excerpt	Data Codes	Interpretation
Cultural-pragmatic alignment	Language accessibility	Using clear Indonesian public-health language rather than technical or diagnostic jargon	3	1	4	8	34.8%	App descriptions; institutional service pages; government health education; language-use literature	Online counselling; mental-health tips; public education on coping; accessible help-seeking guidance	TXT-P01-01; TXT-P02-01; TXT-P03-01; TXT-S05-01; TXT-C01-01; TXT-C03-01; TXT-C04-01; TXT-C05-01	Accessible Indonesian is a major alignment feature because public texts avoid dense psychiatric terminology.
Cultural-pragmatic alignment	Register moderation	Balancing warmth, respectful distance, and service credibility in user-facing language	4	0	2	6	26.1%	Safe-space pages; AI Buddy descriptions; companion/listener framing; government advice	Safe and comfortable space; AI Buddy; companion/friend/listener; trusted conversation	TXT-P02-01; TXT-P03-01; TXT-P05-01; TXT-P01-02; TXT-C03-01; TXT-C05-01	Register is generally moderated through friendly but not overtly clinical language, although some companion terms may risk over-personalisation.
Cultural-pragmatic alignment	Stigma-sensitive framing	Avoiding language that frames distress as weakness, irrationality, moral failure, or personal defect	2	1	2	5	21.7%	Safe-space language; anonymous disclosure; mental-health education; compassionate communication literature	Anonymous/private storytelling; safe space; trusted conversation; mental-health maintenance	TXT-P01-02; TXT-P03-01; TXT-S04-01; TXT-C03-01; TXT-C05-01	Stigma sensitivity is present indirectly through safety and privacy language, but explicit anti-stigma wording is less developed.
Cultural-pragmatic alignment	Family and community relevance	Recognising that distress and help-seeking may involve family, community, school, or caregiving relations	2	1	3	6	26.1%	Student support; family mental-health assessment; adolescent and public-health education	Student mental-health questions; family assessment; trusted conversation; support pathway	TXT-P02-01; TXT-P04-01; TXT-S03-01; TXT-C02-01; TXT-C03-01; TXT-C05-01	Family and community appear as care contexts, but public texts rarely elaborate culturally specific family pressure or shame.
Cultural-pragmatic alignment	Locally credible referral	Connecting users with Indonesian services, psychologists, operators, or official support channels	3	0	5	8	34.8%	Online psychologist platforms; PusBuddy referral; Healing119; Kemenkes-linked resources	Online psychologist counselling; operator review; referral to PUSPAGA psychologists; Healing119; professional help	TXT-P01-01; TXT-P03-01; TXT-P04-02; TXT-C01-01; TXT-C02-01; TXT-C03-01; TXT-C04-01; TXT-C05-01	Local referral is a strong cultural-pragmatic anchor because care is made credible through Indonesian service pathways.
Cultural-pragmatic alignment	Multilingual and regional sensitivity	Acknowledging linguistic diversity, uneven access, or regional variation in healthcare communication	0	2	1	3	13.0%	Healthcare language-barrier literature; public health accessibility context	Language barriers; multilingual patients; healthcare delivery; public access to mental-health information	TXT-S05-01; TXT-S08-01; TXT-C04-01	Multilingual sensitivity is the least visible alignment feature, despite its relevance for Indonesian healthcare access.

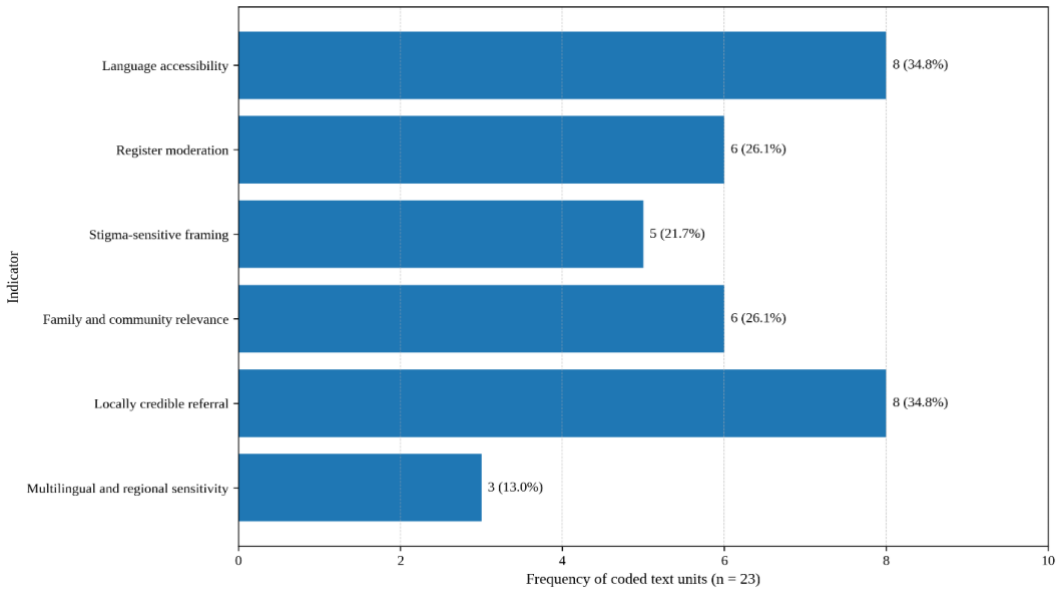

Figure 4. Cultural-pragmatic alignment

Theoretically, these findings suggest that cultural-pragmatic alignment is not achieved simply by translating mental health chatbot discourse into Indonesian. Alignment depends on whether language is comprehensible, socially credible, stigma-aware, and connected to realistic support infrastructures. This extends scholarship on chatbot-based emotional support, which shows that supportive systems must be adapted to cultural context, language practice, and local help-seeking norms rather than evaluated only through generic empathy or usability metrics [29], [30], [31]. For this study, cultural alignment functions as a boundary condition of empathy: a response cannot be emotionally safe if it sounds caring but remains pragmatically distant from users' social worlds.

5. DISCUSSION

The first implication of this study is that empathy in Indonesian mental health chatbot discourse functions primarily as an access-oriented promise rather than as a fully interactional therapeutic practice. The stronger visibility of relational presence and autonomy support indicates that public-facing chatbot materials are effective in inviting users to approach, disclose, seek tips, or move toward professional help. This is not insignificant, particularly in a mental health context where stigma, cost, distance, and hesitation may delay care-seeking. However, this same pattern also reveals a dysfunction: empathy is more often presented as availability, companionship, or supportive branding than as sustained affective attunement. This finding partly supports research on empathic conversational agents, which shows that warmth and responsiveness can improve perceived support [6], [17], but it also complicates studies that treat empathy as a measurable response quality. For this study, empathy is not merely present or absent; it is unevenly distributed across linguistic functions.

The underlying structure behind this unevenness lies in the difference between empathy as platform positioning and empathy as discourse practice. Public-facing chatbot materials are designed to reduce entry barriers, persuade users that support is available, and establish service credibility. Consequently, they privilege relational presence and autonomy support because these indicators are easier to communicate in promotional, institutional, and public health texts. By contrast, affective recognition, validation, and explicit non-judgment require more fine-grained interactional evidence, usually found in turn-by-turn dialogue rather than static platform descriptions. This explains why this study finds supportive language without necessarily finding deep dialogic attunement. Sharma et al. (2021, 2022) [18], [8] argue that empathic mental health support depends on context-sensitive response choices, while Chin et al. (2025) [19] show that empathetic replies vary across help-seeking queries. The present pattern therefore reflects the structural limits of evaluating empathy from publicly visible discourse: it captures design claims and communicative orientation, but not full therapeutic responsiveness.

The second implication concerns emotional safety. This study shows that Indonesian mental health chatbot discourse is strongest when it frames AI-mediated support as bounded, referential, and connected to human care. Boundary transparency, diagnosis restraint, and human-in-the-loop referral function as

protective mechanisms: they prevent chatbots from being positioned as autonomous therapists and help users understand when professional assistance is necessary. This supports ethical scholarship warning that mental health chatbots become risky when they obscure limits, overstate competence, or encourage emotional dependency [9], [10], [22]. Yet this protective function is incomplete. Refusal style, crisis escalation wording, and consent framing appear less developed, suggesting that safety is often institutional rather than conversational. In other words, sources may state that professional care matters, but they rarely show how a distressed user would be relationally held during a high-risk moment. Emotional safety therefore remains a boundary achievement, not merely a disclaimer.

The reason this boundary pattern matters is that mental health risk emerges at the intersection of language, vulnerability, and system limitation. A chatbot can be technically limited yet relationally safe if it communicates limits with care, offers concrete referral, and avoids abandoning the user. Conversely, it can sound warm but become unsafe if it gives false reassurance, implies diagnostic certainty, or refuses abruptly. Recent safety frameworks increasingly recognise this problem by moving beyond accuracy toward stress testing, trajectory-based safety, and harm-sensitive evaluation [32], [5], [24], [25]. This study extends that position by showing that emotional safety in Indonesian chatbot discourse is structured through two simultaneous logics: care expansion and risk containment. AI is made socially useful because it expands low-threshold access, but it becomes ethically defensible only when its language maintains diagnostic restraint, privacy awareness, and credible referral. This is why safety should be assessed as communicative governance, not only as system performance.

The third implication is that cultural-pragmatic alignment determines whether chatbot empathy can be interpreted as meaningful care in Indonesia. Accessible Indonesian and locally credible referral appear as the most visible alignment features, indicating that public-facing chatbot discourse often avoids excessive psychiatric jargon and connects users to recognisable support channels. This finding aligns with Indonesian healthcare linguistics, where comprehension, pragmatic clarity, compassionate language, and patient voice shape the legitimacy of care [26], [13], [14], [15]. However, this study also reveals a dysfunction: cultural fit remains relatively shallow when stigma, family pressure, shame, multilingual diversity, and regional access are only indirectly addressed. Chatbot discourse may therefore sound clear and supportive while still failing to recognise the social conditions through which distress is expressed. This partially supports cross-cultural chatbot research [29], [30], but it also shows that localisation cannot be reduced to language translation.

The underlying structure of this cultural-pragmatic gap is the dominance of standardised mental health communication over socially embedded distress discourse. Indonesian chatbot materials tend to adopt accessible national-language framing because it is scalable, institutionally safe, and broadly comprehensible. This structure benefits public communication but may underrepresent users whose distress is shaped by regional language, family hierarchy, religious interpretation, gendered expectations, or fear of stigma. Healthcare communication studies in Indonesia have shown that directive speech, medical explanation, bad-news delivery, and language barriers affect how patients interpret care [27], [28], [13]. This study brings that insight into AI-mediated mental health communication by arguing that emotional safety requires pragmatic proximity to users' social worlds. The implication is not that every chatbot must reproduce all local cultures, but that culturally thin empathy may remain performative. For Indonesian mental health chatbots, empathy by design must therefore be relationally expressed, emotionally bounded, and culturally situated.

6. CONCLUSION

This study demonstrates that empathy in Indonesian mental health chatbots is not a stable attribute of artificial intelligence, but a relational achievement shaped by language, boundaries, and cultural fit. Its central lesson is that emotionally supportive design becomes credible only when affective recognition, validation, non-judgment, autonomy, referral clarity, and pragmatic sensitivity operate together. By integrating discourse-pragmatic analysis, emotional safety audit, and cultural-pragmatic alignment, this study contributes a more nuanced framework for evaluating mental health chatbot communication beyond usability, accuracy, or general user satisfaction. It refreshes current debates by shifting attention from whether chatbots appear empathetic to how empathy is linguistically and ethically designed.

This study is limited by its reliance on publicly available platform texts, institutional descriptions, public health documents, and secondary literature rather than private user–chatbot transcripts or clinical outcome data. Consequently, its findings address communicative design and safety alignment, not therapeutic effectiveness, symptom reduction, or causal impact on user well-being. Future research should examine real interactional exchanges under strict ethical safeguards, compare multiple Indonesian-language chatbot systems across risk scenarios, and include user perspectives from diverse linguistic, regional, and help-seeking backgrounds. Further mixed-methods work could also test whether relationally safer wording improves trust, comprehension, escalation behaviour, and willingness to seek professional mental health support.

ACKNOWLEDGMENTS

The author thanks the reviewers and editorial team for their constructive feedback on this manuscript.

FUNDING INFORMATION

Author states no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

Handono Fatkhur Rahman: conceptualization, methodology, formal analysis, investigation, data curation, writing – original draft, writing – review and editing.

CONFLICT OF INTEREST STATEMENT

Author states no conflict of interest.

AI USE DECLARATION

The author used artificial intelligence tools to assist with the generation of illustrative figures and with the formatting and layout of this manuscript. AI was not used to generate the substantive content, conceptualization, data, analysis, or conclusions of the research, all of which remain the sole responsibility of the author.

INFORMED CONSENT

Not applicable – this study did not involve direct human participants. Data consisted solely of publicly available Indonesian mental health chatbot discourse materials.

ETHICAL APPROVAL

This research did not involve human or animal subjects; it analyzed publicly accessible Indonesian mental health chatbot discourse materials, and therefore no institutional review board approval was required. Where research does involve human subjects, the author affirms that such research would be conducted in full compliance with all relevant national regulations and institutional policies, in accordance with the tenets of the Declaration of Helsinki, and would be approved by the author's institutional review board or equivalent committee.

DATA AVAILABILITY

The data (publicly available Indonesian mental health chatbot discourse materials and coding materials) analyzed in this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] E. M. Boucher, N. Harake, H. E. Ward, S. Stoeckl, J. Vargas, J. D. Minkel, A. Parks, and R. D. Zilca, "Artificially intelligent chatbots in digital mental health interventions: A review," *Expert Review of Medical Devices, 18, 37–49*, 2021, doi: 10.1080/17434440.2021.2013200.

[2] Y. He, L. Yang, C. Qian, T. Li, Q. Zhang, and X. Hou, "Conversational agent interventions for mental health problems: Systematic review and meta-analysis of randomized controlled trials," *Journal of Medical Internet Research, 25*, 2022, doi: 10.2196/43862.

[3] H. Li, R. Zhang, Y.-C. Lee, R. E. Kraut, and D. Mohr, "Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being," *NPJ Digital Medicine, 6*, 2023, doi: 10.1038/s41746-023-00979-5.

[4] R. Balan, A. Dobrean, and C.-R. Poetar, "Use of automated conversational agents in improving young population mental health: A scoping review," *NPJ Digital Medicine, 7*, 2024, doi: 10.1038/s41746-024-01072-1.

[5] Z. Guo, A. Lai, J. Thygesen, J. Farrington, T. Keen, and K. Li, "Large language models for mental health applications: Systematic review," *JMIR Mental Health, 11*, 2024, doi: 10.2196/57400.

[6] R. Sanjeewa, R. Iyer, P. Apputhurai, N. Wickramasinghe, and D. Meyer, "Empathic conversational agent platform designs and their evaluation in the context of mental health: Systematic review," *JMIR Mental Health, 11*, 2024, doi: 10.2196/58974.

[7] A. Howcroft, A. Bennett-Weston, A. Khan, J. Griffiths, S. Gay, and J. Howick, "AI chatbots versus human healthcare professionals: A systematic review and meta-analysis of empathy in patient care," *British Medical Bulletin, 156*, 2025, doi: 10.1093/bmb/ldaf017.

[8] A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, and T. Althoff, "Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support," *Nature Machine Intelligence, 5, 46–57*, 2022, doi: 10.1038/s42256-022-00593-2.

[9] S. Coghlan, K. Leins, S. Sheldrick, M. Cheong, P. Gooding, and S. D'Alfonso, "To chat or bot to chat: Ethical issues with using chatbots in mental health," *Digital Health, 9*, 2023, doi: 10.1177/20552076231183542.

[10] Z. Khawaja and J. Bélisle-Pipon, "Your robot therapist is not your therapist: Understanding the role of AI-powered mental health chatbots," *Frontiers in Digital Health, 5*, 2023, doi: 10.3389/fdgth.2023.1278186.

[11] L. Laestadius, A. Bishop, M. Gonzalez, D. Illenčík, and C. Campos-Castillo, "Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika," *New Media & Society, 26, 5923–5941*, 2022, doi: 10.1177/14614448221142007.

[12] J. J. Shen, D. DiPaola, S. Ali, M. Sap, H. W. Park, and C. Breazeal, "Empathy toward artificial intelligence versus human experiences and the role of transparency in mental health and social support chatbot design: Comparative study," *JMIR Mental Health, 11*, 2024, doi: 10.2196/62679.
[13] M. Nurtyas, "Language barriers in healthcare delivery: A sociolinguistic case study of multilingual patients in Indonesian urban clinics," *Indonesian Journal of Medical Linguistics, 1(1), 26–38*, 2026.
[14] Y. D. Rusita, "Constructing empathy through language: A linguistic perspective on compassionate communication in mental health services," *Indonesian Journal of Medical Linguistics, 1(1), 67–81*, 2026.
[15] M. Syaifuddin, "Narratives of illness in patient blogs: Exploring personal voices and medical identity in Indonesian digital health discourse," *Indonesian Journal of Medical Linguistics, 1(1), 52–66*, 2026.
[16] V. Sorin, D. Brin, Y. Barash, E. Konen, A. Charney, G. Nadkarni, and E. Klang, "Large language models and empathy: Systematic review," *Journal of Medical Internet Research, 26*, 2023, doi: 10.2196/52597.
[17] M. Abbasian, I. Azimi, M. Feli, A. M. Rahmani, and R. C. Jain, "Empathy through multimodality in conversational interfaces," *ArXiv, abs/2405.04777*, 2024, doi: 10.48550/arxiv.2405.04777.
[18] A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, and T. Althoff, "Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach," *Proceedings of the Web Conference 2021*, 2021, doi: 10.1145/3442381.3450097.
[19] H. Chin, G. Baek, C. Cha, and M. Cha, "Chatbots' empathetic conversations and responses: A qualitative study of help-seeking queries on depressive moods across 8 commercial conversational agents," *JMIR Formative Research, 9*, 2025, doi: 10.2196/71538.
[20] J. De Freitas, A. Uğuralp, Z. Oğuz-Uğuralp, and S. Puntoni, "Chatbots and mental health: Insights into the safety of generative AI," *Journal of Consumer Psychology*, 2023, doi: 10.1002/jcpy.1393.
[21] J. Park, M. Abbasian, I. Azimi, D. Bounds, A. Jun, J. Han, et al., "Building trust in mental health chatbots: Safety metrics and LLM-based evaluation tools," 2024.
[22] H. R. Lawrence, R. A. Schneider, S. Rubin, M. J. Mataric, D. McDuff, and M. J. Bell, "The opportunities and risks of large language models in mental health," *JMIR Mental Health, 11*, 2024, doi: 10.2196/59479.
[23] M. R. Meadi, T. Sillekens, S. Metselaar, A. Van Balkom, J. Bernstein, and N. Batelaan, "Exploring the ethical challenges of conversational AI in mental health care: Scoping review," *JMIR Mental Health, 12*, 2025, doi: 10.2196/60432.
[24] K. H. Bentley, L. Belli, A. Chekroud, E. J. Ward, E. R. Dworkin, E. V. Ark, et al., "VERA-MH: Reliability and validity of an open-source AI safety evaluation in mental health," *ArXiv, abs/2602.05088*, 2026, doi: 10.48550/arxiv.2602.05088.
[25] H. Morrin, J. Au-Yeung, Z. Agnew, S. Østergaard, and T. A. Pollak, "It is the journey, not the destination: Moving from end points to trajectories when assessing chatbot mental health safety," *JMIR Mental Health, 13*, 2026, doi: 10.2196/91454.
[26] N. Laila, "Medical jargon and patient comprehension: A linguistic analysis of informed consent practices in Indonesian hospitals," *Indonesian Journal of Medical Linguistics, 1(1), 13–25*, 2026.
[27] I. L. Nazulpa, "Doctor–patient communication in rural Indonesia: A pragmatic study of directive speech acts in clinical consultations," *Indonesian Journal of Medical Linguistics, 1(1), 1–12*, 2026.
[28] A. S. Nugroho, "Verbal strategies for delivering bad news in oncology settings: A discourse-pragmatic approach," *Indonesian Journal of Medical Linguistics, 1(1), 39–51*, 2026.
[29] H. Chin, H. Song, G. Baek, M. Shin, C. Jung, M. Cha, J. Choi, and C. Cha, "The potential of chatbots for emotional support and promoting mental well-being in different cultures: Mixed methods study," *Journal of Medical Internet Research, 25*, 2023, doi: 10.2196/51712.
[30] D. Nirupama, N. Rao, and C. Sinha, "Language adaptations of mental health interventions: User interaction comparisons with an AI-enabled conversational agent (Wysa) in English and Spanish," *Digital Health, 10*, 2024, doi: 10.1177/20552076241255616.
[31] Y. Shan, M. Ji, W. Xie, X. Qian, R. Li, X. Zhang, and T. Hao, "Language use in conversational agent–based health communication: Systematic review," *Journal of Medical Internet Research, 24*, 2022, doi: 10.2196/37403.
[32] S. Sarkar, M. Gaur, L. K. Chen, M. Garg, and B. Srivastava, "A review of the explainability and safety of conversational agents for mental health to identify avenues for improvement," *Frontiers in Artificial Intelligence, 6*, 2023, doi: 10.3389/frai.2023.1229805.

BIOGRAPHY OF AUTHOR

Handono Fatkhur Rahman is a PhD student at Universitas Indonesia, with a primary area of expertise in nursing and health sciences. He is also an Indonesian academic, lecturer, and researcher affiliated with the Faculty of Health (Fakultas Kesehatan) at Universitas Nurul Jadid in Probolinggo, East Java. His academic research and publications focus heavily on clinical nursing, community health interventions, and patient psychological well-being. Key areas of focus include Clinical Nursing & Interventions, Psychological & Mental Health, and Community & Public Health. He can be contacted at email: handono.hfc@gmail.com.