DrukSerenity: A Culturally-Aware Multi-Agent Chatbot Framework for Mental Health Support in Bhutan

Sonam Eyden

Omdena, Global AI Collaborative
Computing Technologies Department, College of Science and Technology, Royal University of Bhutan
E-mail: sonameyden2005@gmail.com

Received: 24 May 2026; Revised: 06 July 2026; Accepted: 08 August 2026; Published: 13 August 2026

Abstract
Globally, mental health challenges represent a significant public health issue, and Bhutan similarly faces barriers to accessing mental health support due to cultural stigma and limited resources. Currently, there are mental health chatbots, but they are primarily for Western, English-speaking users. There is a need for a system that safely and culturally adapts for Bhutanese people. Therefore, this paper introduces DrukSerenity, a multi-agent chatbot that addresses crisis, offers birdirection Dzongkha-English translation and provides appropriate cultural content using data consisting of 977 Dzongkha proverbs ,alongside 22 cultural, policy, and clinical documents from related institutional sources. The system was evaluated on a synthetic benchmark of 542 messages using rule-based checks and an LLM-judge protocol, showing strong safety performance (100% crisis precision, 92.2% crisis F1) and high cultural fidelity (0.94 RAG fidelity), with an overall functional accuracy of 83.8%, where Emotion classification accuracy and processing latency remains a limitation. Although formal clinical validation with licensed practitioners has not yet been conducted, an informal usability pilot with 20 Bhutanese users received good ratings (overall experience 3.80/5) alongside concerns about proverb overuse. DrukSerenity nonetheless offers a replicable template for building culturally-aware mental health AI in other low-resource, linguistically underserved settings.

Keywords: *Multi-agent systems, mental health support, cultural adaptation, conversational AI, retrieval-augmented generation, crisis detection, Dzongkha translation, privacy architecture*

1. INTRODUCTION

Challenges in Mental health are rising in the world; about one in eight people lives with a mental health condition (WHO, 2022). There is a lifetime prevalence estimate ranging from 18.1% to 36.1% across countries (Kessler et al., 2007; WHO, 2022). In Bhutan, Mental Health is affected by cultural stigma and emotional restraint, with only six psychiatrists serving a population of over 750,000(Asia News Network; 2025; ABC News, 2024). Psychological well-being is considered a national priority within Bhutan's Gross National Happiness (GNH) framework, yet there are significant gaps in access to mental health support, especially in rural populations (Dorji, 2020). Studies show that there is low engagement and effectiveness in Western mental health models for Bhutanese people due to a lack of content related to Bhutanese values, which are rooted in Buddhist Philosophy and collective harmony (Calabrese & Dorji, 2014).

Potential is there for the existing conversational AI system, but they still have a limitation that has been clearly documented, such as inadequate cultural adaptation, simplistic emotion detection, reactive crisis handling, vulnerabilities in an individual's privacy and poor contextual continuity(Brown University,2025; Kocaballi et al. 2022). Currently, no system addresses the needs of the Bhutanese users who require their own language to interact, the cultural familiarity grounded in GNH and Buddhist philosophy, and a platform that provides a private and stigma-free environment.

DrukSerenity fills this gap. It is a multi-agent conversational AI framework that provides mental health support in the Bhutanese context through a hybrid adaptive sequential pipeline that coordinates the six core agents of the framework, integrating conditional branching, retrieval-augmented generation ,bidirectional Dzongkha-English translation, two-tier memory architecture and privacy first design. This work is important because first it addresses the limitations of Western mental health platforms that fail to provide cultural context tailored for Bhutanese people. Second, it

serves as a replicable architectural template that integrates cultural knowledge where other countries with limited resources can adapt to build conversational AI systems using their own languages, traditions, and local knowledge. Thirdly, this study provides empirical results that supports the use of a multi-agent framework for mental health systems especially in situations where responding safely during crisis and cultural appropriateness are crucial. The remainder of the paper is organized as follows: Section 2 reviews related work, Section 3 describes the methodology, Section 4 presents results, Section 5 discusses them, Section 6 concludes the paper, and Section 7 outlines future work.

2. RELATED WORK

The early system showed the possibility of AI that automates conversational support. A rule-based chatbot called Woebot delivers cognitive behavioral therapy (CBT) through daily brief conversation sessions that showed improvements in depression symptoms among college students in a randomized controlled trial(Fitzpatrick et al., 2017). Although Wysa and Replika offer comparable emotional support through empathetic responses and mood tracking, a systematic review by Abd-Alrazaq et al. (2020) of 41 mental health chatbots found that most systems lack effective crisis detection and cultural customization and exhibit limited conversational flexibility due to rule-based dialogue. Kuhail et al. (2025) also reviewed 97 published studies on mental health chatbots. They found that most research has not considered cultural adaptation and that limited attention has been given to severe mental health symptoms. A study made by Guo et al. (2024a) on large language models (LLMs) for mental health discovered perennial risks, which are inconsistent generated text, hallucinations and no ethical consideration for deployment. A similar review found a lack of unique cultural understanding and values outside Western countries, as well as the provision of clinically incorrect information (Yang et al., 2026).

Across these reviews, a recurring observation is that the majority of mental health chatbots aim for the Western, English-speaking population, which was documented in study from 2020 to 2025 across 25 AI-based mental health chatbots (Algumaei et al., 2025) confirmed by Zhang et al. (2025) in non-WEIRD (non-Western, Educated, Industrialized, Rich, and Democratic). Importance of language and cultural adaptation is is showed by Dinesh et al. (2024) through making a Spanish version of chatbot Wysa, achieving higher user engagement and emotional expression and Ankomah and Turkson (2025)'s emotion-aware chatbot in the context of Ghana's public health, which reported of being culturally relevant by 81 percent of users, showing that localization builds trust and reliability of the system, especially in low-resource settings. In Bhutan, Mental health is shaped by Gross National Happiness, Buddhist philosophy, and collective harmony, which Western therapeutic models do not consider those aspects, which would lead to low engagement and effectiveness.

For complex healthcare tasks, specialized LLM-based agents coordinated by an orchestrator, called a multi-agent system, offer potential where According to Guo et al. (2024b) , LLM-based multi-agent systems are more advantageous than single-agent systems since they can specialize roles and coordinate dynamically. Further, Li et al. (2025) showed that CARE-AD, a multi-agent system that predicts Alzheimer's disease, surpasses a single-model baseline, showing that multi-agent systems are feasible for clinical prediction. However, the application of multi-agent systems is scarce, most of which address diagnostic functions rather than therapeutic support. Fan et al. (2024) proposed a multi-agent medical interaction simulation called AI Hospital, acknowledging notable gaps in the performance of multi-turn clinical interaction. Currently, there is no prior work that combines a multi-agent system with cultural domain knowledge for mental health support in a non-Western context, where there still remains a gap between the theoretical potential of multi-agent systems and how it is actually applied in mental health care, especially in culturally rich settings. DruksSerenity is designed to address this.

Getting accurate information from the Large Language models (LLMs)'s internal memory is often not reliable as they are mostly static. Hence, Retrieval-Augmented Generation (RAG), an AI framework, is used to prevent hallucination in LLMs by retrieving relevant documents from an external knowledge base, giving grounded, verified responses which prevent risky inaccurate information (Lewis et al., 2024). Prior work

applied this in mental health, where Shah et al. (2025) proposed a RAG system for mental health, enhancing the factual aspect and Yang et al. (2025) developed CascadeRCG, an online mental health support RAG system that enhances knowledge depth and professional responses. Although persistent challenges such as retrieval noise, poor domain adaptability and model opacity is there in naïve, advanced, and modular designs of RAG systems given by Neha et al.(2025), RAG offer domain accurate knowledge that is required in mental healthcare where responses can be grounded with therapeutic protocols, cultural wisdom, and local resources rather than generic harmful advice. Still, no RAG system offers cultural knowledge in mental health.

In addition, Bhutan has specific needs that current tools may not fully address. Bhutan Broadcasting Service reported that the PEMA Secretariat estimated more than 19,000 people in Bhutan have moderate to moderately severe depressive symptoms and more than 14,000 have moderate to severe anxiety, while its current screening system remains questionnaire-based rather than conversational. It also reported that the DrukCare Mental Health Assistant is still under development and is intended to provide culturally sensitive support for people in Bhutan. Similarly, Gurung et al. (2025) found that customary healing methods and Buddhist philosophy strongly shape Bhutanese views of mental health, reinforcing the need for culturally relevant design in the proposed DrukSerenity framework. Although the Bhutan government has promoted mental health initiatives such as the Helping Adolescents Thrive Programme, no currently described system combines customary cultural knowledge, Dzongkha support, and crisis detection in a single conversational agent

3. METHODOLOGY

3.1 System Architecture Overview

Fig. 1 shows the system architecture of DrukSerenity. It features a translation layer customized for both Dzongkha and English users, a reasoning and response layer made up of six agents which are coordinated by an orchestrator using a hybrid adaptive sequential pipeline. This pipeline structure can be useful in allowing operation of the system in a modular and adaptive fashion as a function of the type of message the user sends. The user messages are then stored in memory with privacy protection. The custom multi-agent framework system is built with the OpenAI SDK; the agent reasoning is taken care of by the GP-5.4-mini model, the RAG (reasoning and composing) are done by the GPT-4o-mini model to balance performance and speed, and retrieval is done by the text-embedding-3 small model.

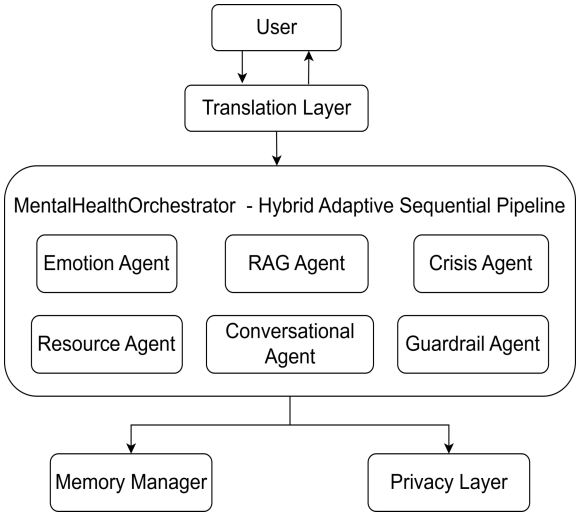

Fig. 1:*DrukSerenity system architecture*

3.2 Data Collection and Preparation

Domain-specific knowledge is applied by DrukSerenity to provide culturally relevant, institutionally validated, clinically relevant responses to facilitate retrieval-augmented generation, reducing the need for parametric knowledge. The knowledge base has been divided into two corpora, mostly because of the differences in their structure, which means that the retrieval of one knowledge base will not dilute the other. First is a corpus of proverbs containing 977 Dzongkha proverbs (including romanised transliterations and English meanings) that consist of Bhutanese moral and emotional wisdom. Second, cultural, institutional and clinical text corpora with 22 documents that include the documents related to GNH policy, national mental health strategy, child protection guidelines, organizational reports, and peer-reviewed literature on wellbeing and compassion focused therapy. Tables 1 and 2 below provide an overview of all data gathered and

examples of the proverbs in two formats

Table. 1: *Knowledge base corpus summary.*

Category	Docs	Prov.	Words
GNH policy	5	-	187,904
MH policy & strategy	7	-	73,418
Child protection	2	-	8,062
RENEW reports	2	-	19,462
Academic/clinical	4	-	32,473
Proverb corpus	2 files	977	11,493
Total	22 sources	977	~332,812

Table. 2:*Sample proverb entries (Format 2)*

#	Dzongkha (script)	literal	English meaning
1	མིའི་དཀར་རོའི་གཙོ་ལྟ་བ་འདྲི། །སས་གནས་པའི་གཙོ་ལྟ་དགོ་དྲག།	Looking at a white person's face is like looking at black earth, which is more important.	Actions and character matter more than appearance.
2	ཕྱིན་ཚེ་ཟེར་བ་དག་ཅ་མི་དགོ། །ཐད་ཚེ་ཟེར་བ་འཇོག་མི་དགོ།	Don't be happy when told "given"; don't be afraid when told "met."	Don't react prematurely to promises or threats.

Documents were collected based on their relevance to mental health and the cultural context of the Bhutanese. There are various sources such as publications from the Centre for Bhutan Studies, proverb listing in the Secretariat of the People's Education Movement (PEMA), reports of the Programme on Education, Migration and Asylum (RENEW), an official Dzongkha proverb collection published by the Dzongkha Development Commission (Format 2) and an online proverb compilation under the title “Bhutanese Proverbs” (Format 1) which lists proverbs in Dzongkha with English translation. To further enrich the format 2 proverbs with Romanized transliterations and English meanings, a large language model was used, and the results were manually checked by the author who is a native Dzongkha speaker. If there is a similarity or overlap on the content, then format 2 is preferred because it has a richer representation in the form of a Dzongkha script, literal meaning and English meaning.

All the documents were subsequently cleaned, Unicode normalised and chunked into chunks of up to 500 words per paragraph, or sentence break for longer documents. Proverbs were saved in both formats at atomic units to ensure proper alignment. The preservation has an important functional value, when a message that needs cultural guidance is needed, the RAG component will retrieve relevant proverbs and will give them directly to the conversational agent. This means that it is not paraphrasing and hallucinating, but it ensures the faithful reproduction resulting in a more culturally unique and authentic response relatable to indigenous Bhutanese people. The chunks and the proverbs are then embedded with text-embedding-3-small, and stored for cosine-similarity-based retrieval during inference.

3.3 Agent Design

3.3.1 Emotion Agent

The Emotion Agent detects emotion, crisis and evaluates the risk in one pass (Algorithm 1) with the taxonomy of more than 320 fine-grained emotions categories and the crisis-confidence threshold of 0.5. If a crisis is detected but the emotion label isn't risky, then the final emotion label is consistent with the type of crisis for consistency.

Algorithm 1 Emotion and crisis detection

Input: m, H, S; **Output:** e, c, k, t, r, g
e, c ← Classify(m, H, S) over 150+ categories
`if e ∉ TaxonomySet then e ← KeywordFallback(m)`
k, t, c_k ← DetectCrisis(m, e, Last3Turns(H))
`if k ∧ c_k ≥ 0.5 ∧ e ∉ HighRiskSet then e ← MapToEmotion(t)`
r ← RiskScore(e, c, k, c_k, S)
g ← TonePolicy(e, r, k)
return e, c, k, t, r, g

3.3.2 RAG Agent

The RAG Agent makes decisions on providing cultural guidance. It bypasses this step if a crisis is detected, the message is too short, the message is a simple greeting or the message doesn't pass relevance criteria(Algorithm 2). It expands the query based on culturally and proverb-related keywords that were earlier built based on an emotion to theme mapping conducted by the

Orchestrator if appropriate. It then fetches similar documents and proverbs and formulates a question to produce the answer.

Algorithm 2 Cultural guidance retrieval

```
Input: m, e, k, n; Output: guidance or Null
if k ∨ Length(m) < 10 ∨ IsGreeting(m) then return Null
if HasCulturalKW(m) ∨ (Relevant(e) ∧ n > 2) ∨ WordCount(m) > 5 then
    q ← EnrichQuery(m, e)
    D ← TopK(q, Docs, 3, 0.25); P ← TopK(q, Proverbs, 2, 0.25)
    if D = ∅ ∧ P = ∅ then Broaden(q); Retry
    return Compose(m, e, D, P)
else return Null
```

3.3.3 Crisis Resource Manager

It is triggered when a crisis is detected which skips the standard path and focuses on emergency support.

Algorithm 3 Crisis resource handling

```
Input: t, m, e; Output: crisis response
R ← RetrieveResources(t, m, e)
R ← Filter(R, exclude=NonActionableStats)
if ¬Has(R,112) ∨ ¬Has(R,"1098 ") then R ← R ∪ CriticalDefaults
R ← DedupPrioritize(R)
C ← DefaultCopingSteps(t)
return Template(open=Empathetic, contacts=Top(R,3), steps=C[1:2], close=Supportive)
```

3.3.4 Resource Agent

If the Orchestrator detects a non-crisis message that needs support, the Resource Agent retrieves coping strategies from RAG, then from LLM generation, and lastly falls back to the web search, where it always includes general Bhutanese resources, and only uses hardcoded coping strategies when all three retrieval methods fail(Algorithm 4).

Algorithm 4 Coping strategy retrieval

```
Input: e, i, m; Output: resources R
R ← RAGResources(e, m); if |R| ≥ 3 then return R ∪ General(e)
L ← LLMStrategies(e, i, m); R ← R ∪ L
if |L| ≥ 2 then return R ∪ General(e)
W ← WebSearch(e, i); R ← R ∪ W
if R = ∅ then R ← Fallback(e)
return R ∪ General(e)
```

3.3.5 Conversational Agent

The Conversational Agent generates the final response by combining emotion, tone, crisis status, retrieved resources, cultural context, and relevant conversation history. Before incorporating them into the prompt, when summaries from previous sessions are available, it first determines relevancy to the current conversation using keyword matching ,followed by LLM decision making if the first method doesn't work. For Dzongkha responses, additional simplification instructions are included to improve readability. If generation fails or produces an insufficient response, the system falls back to a predefined template or, if context assembly failed earlier, a crisis-aware emergency response.

Algorithm 5 Response composition

```
Input: m, e, H, cs; Output: response
temporal ← HasPriors(cs) ? ContextScope(m,e,H,cs) : ∅ // keywords; LLM if ambiguous
ctx ← BuildContext(e, tone, crisis, resources, guidance, History(H), temporal)
sys ← SystemPrompt ⊕ (TargetLang=dz ? SimplifyRules : ∅)
response ← Generate(sys, ctx)
if Empty(response) ∨ TooShort(response) ∨ GenError then response ← Fallback(e, resources, tone, H)
if PreCtxError then response ← EmergencyFallback(crisis)
return response
```

3.3.6 Guardrail Agent

In the non-crisis workflow, the validated responses (Algorithm 6) are not sent to the user until they have passed a combined validation process that includes all of the following criteria: diagnosis, boundary, harm, and privacy, after which a more appropriate response will be offered instead of a blocked reply. A validator parsing failure will default to considering the original response to be safe and bypass this step entirely in crisis responses.

Algorithm 6 Response validation

1**Input:** y; **Output:** validated y′
2 if Crisis then return y
3 result ← Check(y, {NoDiagnosis, Boundaries, NoHarm, Privacy})
4 if ParseError(result) then return y
5 if result.isSafe then return y
6 else return result.corrected or DefaultSafe

3.4 Hybrid Adaptive Sequential Pipeline

The orchestrator executes the workflow as illustrated in Figure 2 through a hybrid adaptive sequential pipeline where Standard message executes the seven steps if any conditional branches are fired, while detected crisis branches straight to the Crisis Resource Manager for immediate emergency response.

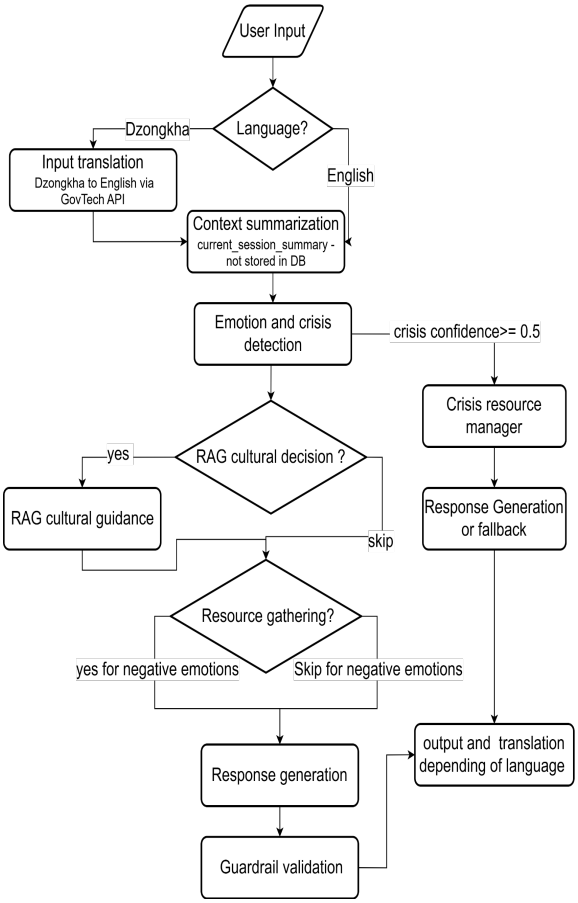

Fig. 2: *Hybrid adaptive sequential pipeline*

3.5 Memory Architecture

The two-tier summarisation strategy is divided into two streams: session level and cross-session context (Figure 3). A temporary summary is re-generated approximately and If the session was of a minimum length, at the end of the session a final summary is generated just once and stored persistently for future sessions.

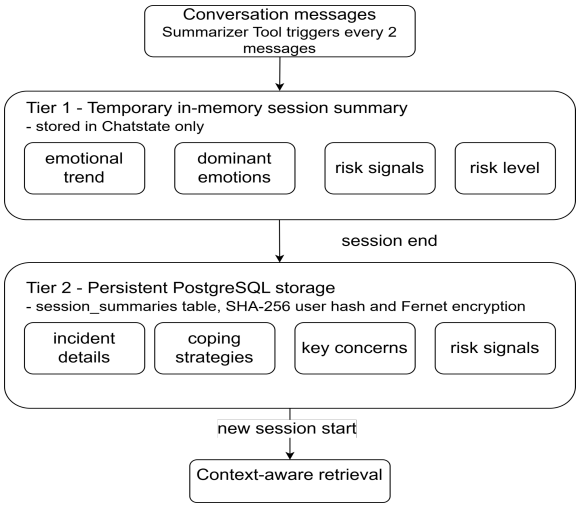

Fig. 3: *Two-tier memory architecture*

During a resuming session the system will retrieve up to 5 summaries of past sessions (excluding the current one).

3.6 Privacy Architecture

It is designed in a dual mode such that each user identifier is hashed by using the cryptographic hash algorithm of SHA-256 in the authenticated mode to obfuscate their identity. It also uses a Fernet symmetric encryption that locks and makes the text unreadable without a key. For anonymous mode, since no account is created, each user is given a Universally Unique Identifier (UUID) with structured metadata so that temporary data stored within the session can be used properly to give an appropriate response and is discarded after discontinuation.

3.7 Evaluation Methodology

The absence of a benchmark dataset for mental health conversational AI from Bhutan, means that testing on Bhutanese vulnerable users, particularly for crisis scenarios assessed here, represents an ethical risk, and is not within the scope for evaluation during the development stage. Thus, synthetic generation ensures a controlled, repeatable of various scenarios without putting actual users at risk. This approach was demonstrated by Park and Smith(2024) to be a viable approach for using LLM approaches for safety evaluations in such systems, but it should

not replace the real participation and licensed clinicians.

Special care was taken to generate synthetic data in response to various scenarios. In total, with 542 messages across 6 categories (see Table 3), a total of 6 batches were used with GPT-5.4-mini using a temperature of 0.3 to avoid overlap between batches and ensure scenario diversity. The token limits for each category were from a minimum of 10000 tokens for shorter positive or neutral messages to 17000 tokens for dzongkha messages to make up for the cost of muti-byte Tibetan Unicode which resulted in 42 of the 50 messages being under this limit. The dataset was then stored as a JSON file, using as an input for the evaluation pipeline.

Table 3*: Synthetic evaluation dataset by category*

Category	Bat.	Msgs	Focus
Emotional support	4	100	Distress, fear/anger, guilt, isolation, burnout
Crisis	4	100	Suicidal ideation, self-harm, harm-to-others, panic, negation traps
Cultural / RAG	4	100	GNH, proverbs, Buddhist practice, community values
Positive / neutral	4	100	Joy, gratitude, confidence, hope, peace, neutral
Edge cases	4	100	Medical boundary, mixed emotions, aggression, dissociation
Dzongkha	2	50 (42)	Emotional support and crisis, in Dzongkha script

Each time a message traversed the pipeline, it records the state of each agent in the pipeline including their detected emotion and their confidence level, the crisis flag and its category, whether the agent activated the RAG flag or not and the associated proverb(s) or resources retrieved, the branch of the workflow the agent followed, and the final response. These outputs were recorded as a structured trajectory record for each of the messages that were produced, providing full traceability of the system. The overall evaluation process using synthetic data from synthetic data generation to final scoring is summarized in Figure 4.

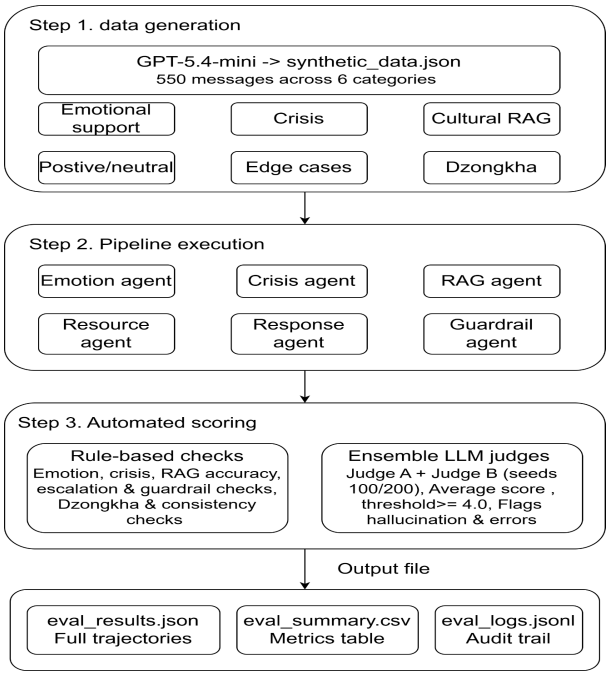

Fig.4 *DrukSerenity evaluation pipeline*

The system's performance is then evaluated against ground truth in the Rule-based check to test the accuracy of the emotional labels, correctness of crisis precision and recall, correctness of RAG activation, presence of resources, correctness of guardrail triggers, and correctness of Dzongkha Unicode validation. After this, Two independent GPT-5.4-mini judges evaluate faithfulness score based on a 4-step chain of thought protocol, ranging from 1 to 5. Judge A checks fact faithfulness or hallucination and Judge B checks cultural faithfulness and safety. Besides, both judges highlight the proverb fabrication and crisis miscategorisation. The two judges' scores were then averaged out to produce a final faithfulness score. For cases in which the system gives an appropriate response even if it does not give the precise emotion label, a functional accuracy metric was deemed if the predicted emotion label is the same as the ground truth, or the faithfulness score is 4.0 or above, with no crises miscategorized.

3.8 User Experience Pilot

To complement the synthetic evaluation, an informal usability pilot was conducted on a live deployment of DrukSerenity. A Google Form was distributed to a convenience sample of Bhutanese users, with explicit safety instructions not to disclose real personal crises and not to test self-

harm scenarios, which were reserved for the synthetic evaluation. Twenty individuals accessed the prototype and completed the consent and background items; nineteen completed every item of the usability scale. Participants were asked to try at least three scenario types: an emotional or stress-related disclosure, a request for a Bhutanese proverb or cultural perspective, and a casual positive check-in. Usability was measured with a ten-item, five-point Likert scale (1 = strongly disagree, 5 = strongly agree) followed by open-ended questions on strengths, points of confusion, cultural accuracy, and technical issues. A deployment constraint affected testing: the GovTech Dzongkha-English translation API requires VPN-restricted access and could not run on the public Render/Vercel deployment, so live bidirectional translation (Section 3.1) was unavailable during the pilot. This was a non-clinical pilot; participants were not screened for mental health status, and it complements rather than replaces the clinician-reviewed pilot proposed in Section 7

4. RESULTS

4.1 Safety-Critical Performance

Safety metrics are important in a mental health AI system. To ensure high accuracy of crisis detection, the system did not indicate any false-positive messages where all 542 were accurate as shown in figure 5. A high crisis F1-score (harmonic mean of precision and recall) and crisis recall was obtained. Accuracies for high-risk emotion escalation (whether the message included hopelessness, worthlessness or despair always resulted in a resource response) were 100%. For content related to agitation-state grounding, the grounding content was present in 98.1% of the relevant responses, while the accuracy of the guardrails that covered violations of the medical boundary, privacy, and scope of authority reached 96.3%. The accuracy of isolation connections metrics percent shows that the system did not always include a social-connection reference in the answers to lonely or isolated users, but the answers themselves were generally empathetic, which shows the main gap being the type of resource it provided rather than a response quality.

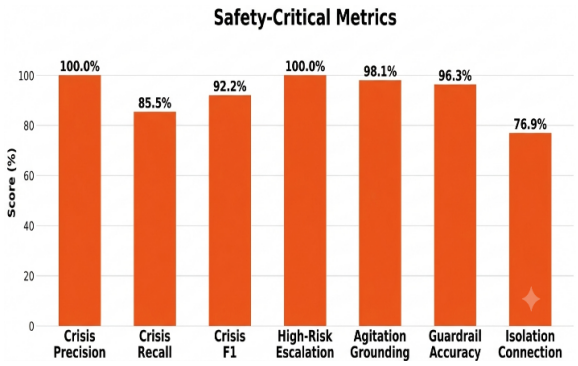

Fig. 5:*Safety-critical evaluation metrics.*

4.2 Cultural Intelligence Performance

As far as cultural grounding is concerned, DrukSernity has done well as reflected in figure 6 below. RAG fidelity, defined as the proportion of retrieved proverbs and cultural content that appeared in the response, was at 0.94, while the accuracy of the responses in Dzongkha was 97.1%, demonstrating that the translation pipeline was able to produce accurate and valid Tibetan Unicode output.The cultural RAG category had the highest functional accuracy with a high ensemble judge score, high hallucination pass rate and high proverb non-fabrication rate, showing that retrieved cultural content was integrated faithfully in most cases. The lack of accuracy on emotion classification shows GNH, proverb, and Buddhist-practice queries are not predominantly emotion-based but rather informational queries, so the level of accuracy on the former is not as important a measure of quality as the latter is in the case of functional accuracy.

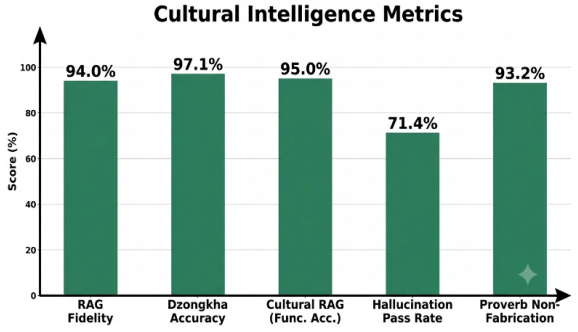

Fig. 6:*Cultural intelligence evaluation metrics.*

4.3 Emotion Classification Performance

The overall accuracy of emotion detection was 52.2% on a taxonomy of over 150 fine-grained categories designed to capture nuanced, culturally-

specific emotion states than could not be captured by a typical taxonomy with six to eight categories like Woebot, Wysa, or Replika would have. Many of the apparent errors are more nuanced differences in the correct emotional neighbourhood for example, hopelessness often became despair or emptiness. Functional accuracy (yes/no if the response was appropriate, no matter what the label was) was 83.8%. The rate of diagnostic emotions that were missed but still outputted as OK was 31.5%, meaning that a third of all emotions received a good answer although the correct emotion label was not given. This suggests that the Conversational Agent appears to be reasonably robust in the face of misdiagnoses by the user on the input side. The emotion accuracy (23.8%) of the Dzongkha category is the lowest, as expected, because the messages in this category are machine-translated from Dzongkha to English before classification and then back to Dzongkha, which leads to an increase in translation error in generating the emotions.

4.4 Per-Category Analysis

Table 4 depicts the emotion accuracy, functional accuracy and ensemble judge score by emotion. The emotional support category had a reasonable score, which confirms that the emotional support is one of the important cases that the system is able to cope with. The system does not over-escalate non-distressed users with positive and neutral messages as the functional accuracy is almost perfect (98%). Since there is a gap between the accuracy of emotion and functional accuracy of the RAG category, informational aspects of the type of questions in this category are suggested (Section 4.2). The most noteworthy trend is that the emotion of the category (crisis) has an accuracy of 34% while the functional and crisis category has an accuracy of 95% and 64% respectively. There's not something that they did wrong, it's an intentional architectural artifact because the crisis detection will trigger the Crisis Resource Manager that will give a structured answer to safety, in which ensemble evaluators are not as generous in scoring an empathetic response. In general, the Dzongkha category is the most difficult category with only being 50% functional and a category ensemble judge score of 3.24/5 due to the reported problem of semantic degradation. This is one of the reasons for further research on native Dzongkha processing (see Section 7).

Table 4*:Per-category evaluation results across key metrics.*

Category	Emotion Acc.	Function. Acc.	Crisis Acc.	Judge (/5)	DZ Acc.
Emotional Support	74.0%	94.0%	100.0%	4.29	N/A
Cultural / RAG	36.0%	95.0%	100.0%	4.38	N/A
Positive /Neutral	83.0%	98.0%	100.0%	4.25	N/A
Crisis	34.0%	64.0%	95.0%	3.50	N/A
Edge Cases	46.0%	82.0%	99.0%	4.00	N/A
Dzongkha	23.8%	50.0%	76.2%	3.24	97.1%

4.5 Hallucination and Fidelity Analysis

Overall, 71.4% passed the hallucination test which was below the targeted rate and focused on two areas, namely, crisis (41%) and Dzongkha (31%). This is primarily due to the "templated" safety-response nature of a crisis, rather than to a factual miscommunication in the response itself, and is more of a reflection of ensemble scoring than of an actual mistake on the part of the responder. In the Dzongkha category, it represents the semantic noise result of translation. The results for the cultural RAG category were similar to its high RAG fidelity (Section 4.2), as the hallucination pass rate was 94%. 6.8% of applicable responses and 5.4% of the instances of crisis miscategorisation had been fabricated as proverbs, both slightly higher than the target. The mean score of the judges for every ensemble was 4.02 out of 5.

4.6 Summary of Results

The entire set of evaluation measures are presented in Table 5. Combined, these results demonstrate a system that performs well on the safety-critical functions, the cultural grounding aspects, and that is even better on the fine-grained emotion labelling , but remains a clear target for further engineering work on the hallucination pass rate and processing latency. These results are for the system under

controlled, synthetic conditions only, and do not in themselves demonstrate clinical benefit in the real world, a distinction that is discussed further in Section 5.

Table 5:*Full evaluation metrics, targets, and observed values.*

Metric	Value
Total Messages Evaluated	542
Emotion Detection Accuracy	52.2%
Functional Accuracy	83.8%
Emotion Miss But Output OK	31.5%
Crisis Precision	100.0%
Crisis Recall	85.5%
Crisis F1	92.2%
Crisis False Positive Rate	0.0%
RAG Activation Accuracy	59.8%
RAG Fidelity Score (Mean)	0.94
Resource Activation Accuracy	68.6%
High-Risk Escalation Accuracy	100.0%
Agitation Grounding Accuracy	98.1%
Isolation Connection Accuracy	76.9%
Guardrail Accuracy	96.3%
Hallucination Pass Rate	71.4%
Ensemble Judge Score	4.02 / 5
Proverb Fabrication Rate	6.8%
Crisis Miscategorisation Rate	5.4%
Dzongkha Response Accuracy	97.1%
Avg. Processing Time	19.1s

4.7 User Experience Pilot Results

Twenty participants tested the deployed prototype. Among the 19 participants who provided demographic information, most were aged 18–24 (47.4%), followed by those aged 25–34 (21.1%), 35–44 (15.8%), and 45 or above (15.8%). Most participants (80%) had not previously used a mental health chatbot. Testing was conducted primarily on mobile devices (55%) and laptops (45%), with no participants using tablets. The mean scores for the ten usability items are presented in Table 6.

Table 6:*Mean usability ratings from the 20-respondent pilot survey.*

Item	Mean (/5)	n
Responses easy to understand	3.89	19
Responses relevant to input	3.75	20
Felt culturally relevant	3.90	20
Tone appropriate and respectful	3.80	20
Responses felt supportive/helpful	3.75	20
Proverbs felt natural, not forced	3.30	20
Handled the message well	3.85	20
Would trust the chatbot	3.32	19
Would use again	3.45	20
Response speed acceptable	3.50	20
Overall experience	3.80	20

Overall, participants reported a positive experience with DrukSerenity, with a mean score of 3.65/5 across the ten usability items and an overall experience rating of 3.80/5. The lowest-rated aspects were the naturalness of the Bhutanese proverbs and users' trust in them. Open-ended feedback revealed three recurring themes. First, participants appreciated the chatbot's cultural grounding, with several describing the Bhutanese proverbs as providing "a sense of comfort and home." However, others felt that the proverbs appeared too frequently, regardless of the conversation, making them seem forced rather than supportive. One participant noted that this issue would likely be more noticeable for the chatbot's target users aged 16–36. Second, several participants commented on the Dzongkha proverb text, describing some translations as overly literal, noting awkward semicolon placement, and reporting one case where the interface defaulted to the wrong language. Because the live Dzongkha-English translation service was unavailable during the public deployment (Section 3.8), these observations relate only to the static proverb corpus described in Section 3.2 rather than the bidirectional translation pipeline evaluated synthetically in Section 4.4. Third, several participants independently highlighted slow response times, consistent with the average latency reported in Section 4.6. A smaller number also felt that some responses were too generic or continued offering reassurance even after users indicated they were feeling better, such as replying with

additional comfort after the user simply responded, "okay." These qualitative findings are consistent with the quantitative results reported in Sections 4.5 and 4.6, particularly regarding proverb integration, hallucination performance, and response latency.

5. DISCUSSION

The evaluation shows that DrukSerenity achieves its core safety goals: no false alarms when handling a real crisis, no misrepresentation of Bhutanese cultural material, and structurally valid Dzongkha output. These results support the architectural decisions described in Section 3, particularly crisis branching in the orchestration pipeline (3.4) and proverb retrieval by the RAG Agent (3.3.2).

All these results, however, are based on a synthetic test developed and scored by large language models (LLMs) (Section 3.7); no Bhutanese participants or licensed clinicians reviewed the test. This is intentional: clinical and ethical safeguards are not yet in place to use real, potentially vulnerable users. Formal clinical piloting through the Omdena x GovTech partnership could not be completed within the project timeline, requiring ethical review and institutional coordination beyond the project's closure. As a partial substitute, an informal, non-clinical usability pilot was conducted with 20 Bhutanese users (Section 4.7), corroborating the synthetic evaluation's findings on proverb integration and processing latency while surfacing additional concerns, such as trust, that the synthetic benchmark could not capture. This pilot has two scope limits: it involved no clinical screening or licensed-practitioner review, so the results in Section 4 still do not establish clinical benefit in real-world deployment; and the public prototype, hosted on Render/Vercel, could not reach the GovTech translation API's VPN-restricted endpoint, so live bidirectional translation was inactive and the architecture in Section 3.1 remains validated only under the synthetic conditions of Section 4.4. Formal clinical validation and a VPN-capable deployment remain the main open questions for future research (Section 7).

Two further deployment issues warrant consideration. Each message passes through sequential LLM calls, producing a system latency of 19.1 seconds, above the conversational target of < 5 seconds; this is an engineering rather than a design problem, addressable through response caching and parallel agent execution. The dual-mode privacy design (Section 3.6) and Guardrail Agent's boundary checks (Section 3.3.6) provide a reasonable baseline of ethical safeguards, but a human-supervised clinical system would need to validate these before deployment with real users.

6. CONCLUSION

In conclusion, mental health support in Bhutan faces systemic challenges, including cultural stigma, a shortage of skilled workers, and geographical distribution. This work introduced DrukSerenity, a multi-agent conversational framework integrating crisis-adaptive orchestration, retrieval-augmented cultural grounding, and bi-directional Dzongkha-English translation in one system. On a synthetic benchmark of 542 messages, the system showed strong safety-critical performance (no false positives, accurate high-risk escalation), high cultural grounding and Dzongkha structural accuracy, alongside acknowledged limitations in fine-grained emotion labelling, hallucination pass rate, and processing latency. These results show the architecture performs as expected under controlled synthetic conditions, and an informal usability pilot with real users lends preliminary support to its cultural and conversational design; however, no clinician-reviewed evaluation has yet been conducted, so the results cannot validate the system's clinical value.Overall, these findings suggest that DrukSerenity is a promising step toward providing accessible and culturally appropriate mental health support in Bhutan, although further clinical evaluation is still needed.

7. FUTURE WORK

For future work, building on the informal usability pilot, a formal clinical pilot with Bhutanese participants and licensed clinician review will be conducted once ethical approval and institutional coordination through the Omdena x GovTech partnership are secured. This requires migrating to VPN-capable infrastructure so the GovTech Dzongkha-English translation API can be reached, allowing the bidirectional translation pipeline to be tested with real users for the first time. Beyond this, priorities include expanding Dzongkha-native language processing to reduce translation-related emotion and hallucination gaps, extending the

proverb and cultural document corpus, and reducing processing latency through response caching and parallel agent execution. Longer-term directions include user modelling for personalised responses, multimodal voice access for low digital literacy, secure referral pathways and therapist-dashboard integration, and specialised agents for goal-setting and moderated peer support.

8. ACKNOWLEDGEMENT

The author acknowledges the use of general-purpose artificial intelligence tools, including text editors, grammar checkers, and formatting assistants, in preparing this manuscript. This work was developed during the Refinement Tool Phase of the Omdena x GovTech Innovation Challenge. The author thanks Omdena and GovTech for the platform and opportunity to develop this framework, and for permission to publish this research, and thanks the College of Science and Technology for the support, guidance, and academic resources that contributed to this work's completion.

9. REFERENCES

Algumaei, A., Yaacob, N. M., Doheir, M., Al-Andoli, M. N., & Algumaei, M. Y. (2025). Symmetric therapeutic frameworks and ethical dimensions in AI-based mental health chatbots (2020–2025): A systematic review of design patterns, cultural balance, and structural symmetry. *Symmetry*, *17*(7), 1082. doi:10.3390/sym17071082

Ankomah, E., & Turkson, R. E. (2025). Emotion-aware AI chatbots for mental health support in low-resource public health systems: A case study from Ghana. *World Journal of Public Health*, *10*(3), 317. doi:10.11648/j.wjph.20251003.17

Dinesh, D., Rao, M., & Sinha, C. (2024). Language adaptations of mental health interventions: User interaction comparisons with an AI-enabled conversational agent (Wysa) in English and Spanish. *Digital Health*, *10*, 1–13. doi:10.1177/20552076241255616

Gurung, A., Timsina, T. C., & Singh, K. N. (2025). Bridging minds and medicines in Bhutan: A scoping review of community perspectives on mental health, substance use, and traditional–modern care integration. *Open Science Framework*. doi:10.17605/osf.io/n5muh

Kuhail, M. A., Al-Shamaileh, O., Farooq, S., Shahin, H., Abdelzaher, F., & Thomas, J. (2025). A systematic review on mental health chatbots: Trends, design principles, evaluation methods, and future research agenda. *Human Behavior and Emerging Technologies*, *2025*(1), 9942295. doi:10.1155/hbe2/9942295

Li, R., Wang, X., Berlowitz, D. R., Mez, J., Lin, H., & Yu, H. (2025). CARE-AD: A multi-agent large language model framework for Alzheimer's disease prediction using longitudinal clinical notes. *npj Digital Medicine*, *8*, 1–10. doi:10.1038/s41746-025-01940-4

Neha, F., Bhati, D., & Shukla, D. K. (2025). Retrieval-augmented generation (RAG) in healthcare: A comprehensive review. *AI*, *6*(9), 226. doi:10.3390/ai6090226

Shah, H. A., Islam, A., Tariq, Z. U. A., Belhaouari, S. B., & Househ, M. (2025). Retrieval augmented generation system for mental health information. *Studies in Health Technology and Informatics*, *327*, 1–5. doi:10.3233/shti250929

Yang, J., Liu, T., Luo, Y., Niu, T., Pang, P. C.-I., Ao, X., & Yang, Q. (2026). Exploring the application boundaries of LLMs in mental health: A systematic scoping review. *Frontiers in Psychology*, *16*, 1715306. doi:10.3389/fpsyg.2025.1715306

Zhang, Q., Zhang, R., Xiong, Y., Sui, Y., Tong, C., & Lin, F.-H. (2025). Generative AI mental health chatbots as therapeutic tools: Systematic review and meta-analysis of their role in reducing mental health issues. *Journal of Medical Internet Research*, *27*, e78238. doi:10.2196/78238

Bhutan Broadcasting Service. (2025, August 29). *The PEMA Secretariat expands mental health care reach by rolling out online screening system*. BBS. https://www.bbs.bt/233068/

Bhutan Broadcasting Service. (2025, October 28). *AI lab program marks first year, unveils two major tools for mental health and climate alert*. BBS. https://www.bbs.bt/234961/

Bhutan Broadcasting Service. (2026, May 29). *Bhutan expands mental health support in schools through HAT Programme*. BBS. https://www.bbs.bt/242651/

Fan, Z., Wei, L., Tang, J., Chen, W., Wang, S., Wei, Z., Huang, F., & Zhou, J. (2025). AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *Proceedings of the 31st International Conference on Computational Linguistics*, 10183–10213.

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, Survey Track*, 8048–8057. doi:10.24963/ijcai.2024/890

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

Yang, D., Zhu, J., Wu, H., Tan, M., Li, C., & Yang, M. (2025). CascadeRCG: Retrieval-augmented generation for enhancing professionalism and knowledgeability in online mental health support. *Proceedings of the ACM on Web Conference 2025*, 1465–1469. doi:10.1145/3701716.3715466

Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M., & Househ, M. (2020). Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *22*(7), e16021. doi:10.2196/16021 ABC News. (2024, April 19). Bhutan is known for being a happy country, but mental health is a hidden problem. *ABC News*. https://www.abc.net.au/news/2024-04-19/bhutan-mental-health-a-hidden-issue-counsellor-says/103681868

Asia News Network. (2025, October 13). Over half of mental health cases in Bhutan linked to anxiety and depression, report finds. *Asia News Network*. https://asianews.network/over-half-of-mental-health-cases-in-bhutan-linked-to-anxiety-and-depression-report-finds/

Brown University. (2025, October 21). New study: AI chatbots systematically violate mental health ethics standards. *Brown University News*. https://www.brown.edu/news/2025-10-21/ai-mental-health-ethics

Calabrese, J. D., & Dorji, C. (2014). Traditional and modern understandings of mental illness in Bhutan: Preserving the benefits of each to support Gross National Happiness. *Journal of Bhutan Studies*, *30*(1). https://d1i1jdw69xsqx0.cloudfront.net/digitalhimalaya/collections/journals/jbs/pdf/JBS_30_01.pdf

Dorji, T. (2020). The Gross National Happiness Framework and the health system response to the COVID-19 pandemic in Bhutan. *The American Journal of Tropical Medicine and Hygiene*, *104*(2), 441–445. doi:10.4269/ajtmh.20-1416

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, *4*(2), e19. doi:10.2196/mental.7785

Kessler, R. C., Angermeyer, M. C., Anthony, J. C., de Graaf, R., Demyttenaere, K., Gasquet, I., de Girolamo, G., Gluzman, S., Gureje, O., Haro, J. M., Kawakami, N., Karam, A., Levinson, D., Medina Mora, M. E., Oakley Browne, M. A., Posada-Villa, J., Stein, D. J., Tsang, C. H. A., Aguilar-Gaxiola, S., . . . Üstün, T. B. (2007). Lifetime prevalence and age-of-onset distributions of mental disorders in the World Health Organization's World Mental Health Survey Initiative. *World Psychiatry*, *6*(3), 168–176.

Kocaballi, A. B., Sezgin, E., Clark, L., Carroll, J. M., Huang, Y., Huh-Yoo, J., Kim, J., Kocielnik, R., Lee, Y.-C., Mamykina, L., Mitchell, E. G., Moore, R. J., Murali, P., Mynatt, E. D., Park, S. Y., Pasta, A., Richards, D., Silva, L. M., Smriti, D., . . . Zubatiy, T. (2022). Design and evaluation challenges of conversational agents in health care and well-being: Selective review study. *Journal of Medical Internet Research*, *24*(11), e38525. doi:10.2196/38525

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktaschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

World Health Organization. (2022). *World mental health report: Transforming mental health for all*. World Health Organization. https://www.who.int/publications/i/item/9789240049338