

Rafiq: An Arabic Conversational AI System for Early Depression Detection with Integrated Symptom Monitoring, PHQ-9 Screening, and Retrieval-Augmented Generation

Alaa Alshurafa

Islamic University of Gaza

Basem Alijla

Islamic University of Gaza

Motaz K H SAAD

`motazk.saad@unisalento.it`

University of Salento

Research Article

Keywords: Conversational AI, mental health chatbot, Arabic NLP, depression detection, PHQ-9, retrieval-augmented generation

Posted Date: October 19th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11154896/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Rafiq: An Arabic Conversational AI System for Early Depression Detection with Integrated Symptom Monitoring, PHQ-9 Screening, and Retrieval-Augmented Generation

Alaa Alshurafa¹, Basem Alijla¹, Motaz K H SAAD^{1,2,*}

¹*Faculty of Information Technology, Islamic University of Gaza, Gaza, Palestinian Territory, Occupied*

²*Department of Engineering for Innovation, University of Salento, Lecce, Italy*

Abstract

Depression is a leading cause of disability worldwide, yet Arabic-speaking communities remain underserved by clinically grounded conversational AI. This paper presents Rafiq, a conversational healthcare informatics system for early Arabic depression detection and psychoeducation, developed end to end and clinically validated. Rafiq integrates four modules: (i) an Arabic mental health knowledge base of 14,624 words curated from four authoritative clinical sources (DSM-5-TR, MSD Manual, WebTeb, and WHO); (ii) a symptom detection module that monitors user messages with GPT-4o-mini and keyword-guided prompting driven by a clinically validated 217-expression Arabic PHQ-9 lexicon; (iii) an automated three-phase PHQ-9 administration and scoring module; and (iv) a retrieval-augmented generation (RAG) pipeline grounded in the knowledge base. The modules are orchestrated by a six-node LangGraph workflow and delivered via a React Native mobile application. Six Arabic sentence embedding models were benchmarked for mental-health retrieval on a purpose-built dataset of 33 queries and 99 annotated query–document pairs; Arabic-Triplet-Matryoshka-V2 performed best (MAP = 80%, NDCG = 86%, Precision@1 = 93% at $k = 2$), while a state-of-the-art multilingual mixture-of-experts model failed to transfer (MAP = 4%). Symptom detection reached 98% accuracy and an F1-score of 0.98 under leakage-aware evaluation. RAGAS evaluation yielded faithfulness = 0.88, answer relevancy = 0.84, and context precision = 0.91, and GPT-4o-mini matched GPT-4o at lower latency and cost. Validation by five licensed psychologists across five scenarios produced a mean overall score of 4.54/5.0 with no critical safety concerns. Rafiq shows that integrating structured clinical screening with evidence-grounded conversational AI is feasible for early depression detection in Arabic.

Keywords: Conversational AI, mental health chatbot, Arabic NLP, depression detection, PHQ-9, retrieval-augmented generation

*Corresponding author: Motaz K H SAAD, Department of Engineering for Innovation, University of Salento, Lecce, Italy.

Email address: motazk.saad@unisalento.it (Motaz K H SAAD)

1. Introduction

Depression is one of the most prevalent and disabling mental health conditions globally, affecting an estimated 280 million people and accounting for a substantial proportion of years lived with disability [1]. In Arabic-speaking regions, the burden of depression is further compounded by cultural stigma, limited access to mental health professionals, and a critical shortage of clinically validated Arabic-language digital health tools [2, 3].

Conversational AI systems and mental health chatbots have emerged as promising vehicles for broadening access to psychological support, offering low-barrier, stigma-reduced, and around-the-clock availability [4, 5, 6]. However, the overwhelming majority of such systems are developed for English or other high-resource languages, leaving Arabic speakers substantially underserved [7, 3]. Existing Arabic-oriented mental health applications are typically limited to general information provision or basic chatbot interaction and do not integrate structured clinical screening protocols or evidence-grounded response generation [2].

A central clinical tool for depression screening is the Patient Health Questionnaire-9 (PHQ-9), a validated nine-item self-report instrument [8] with established Arabic reliability and validity [9]. Despite its wide clinical adoption, To the best of our knowledge, no prior Arabic-language chatbot system has been reported to integrate PHQ-9 administration within a natural conversational flow that also performs symptom monitoring and retrieval-based psychoeducation.

This paper addresses this gap by introducing *Rafiq*, an end-to-end Arabic conversational AI system for early depression detection and support. Beyond the integrated system itself, this work consolidates the underlying components into a unified research pipeline. The paper’s contributions are:

1. **Arabic Mental Health Knowledge Base:** A structured knowledge base of 14,624 words built from four authoritative clinical sources (DSM-5-TR, MSD Manual, WebTeb, WHO) using a three-phase curation methodology that produces retrieval-ready Markdown documents.
2. **Arabic Embedding Benchmark:** The first systematic evaluation of Arabic sentence embedding models for mental health information retrieval, based on a purpose-built 33-query benchmark dataset and six candidate models, identifying Arabic-Triplet-Matryoshka-V2 as the best-performing model and exposing the failure of a strong multilingual mixture-of-experts embedder on this domain.
3. **Symptom Detection Module:** An LLM-based classifier using GPT-4o-mini with structured keyword-guided prompting and a clinically validated 217-expression Arabic lexicon that monitors user messages in real time and maps them to the nine PHQ-9 symptom criteria, achieving 98% accuracy and an F1-score of 0.98.
4. **PHQ-9 Integration Module:** A three-phase protocol that proposes, administers, and scores the PHQ-9 questionnaire within the conversation once sufficient depressive symptoms are detected.
5. **RAG-Based Psychoeducation Module:** A semantic retrieval pipeline that augments LLM responses with passages from the knowledge base, evaluated with RAGAS and shown to make GPT-4o-mini preferable to GPT-4o.

6. **Clinical Validation:** An evaluation by five licensed psychologists across five conversation scenarios confirming the system’s appropriateness, safety, and clinical accuracy.

The three system modules are orchestrated by a LangGraph workflow, exposed through a FastAPI backend, and delivered via a React Native mobile application.

From a healthcare informatics perspective, Rafiq represents a complete deployment of a conversational AI system for a clinically relevant application domain, spanning the design, development, testing, and deployment of digital mental health technology for early depression detection. Its architecture combines a transparent, audit-friendly symbolic layer — a clinically validated Arabic keyword lexicon and a structured knowledge base — with neural components (embedding models and text generators), yielding a hybrid design in which symbolic resources and neural systems are combined to improve transparency and trustworthiness [10, 11]. Every decision is made verifiable: the symptom detector relies on a clinically validated keyword lexicon that makes its decision basis explicit and auditable; the RAG pipeline grounds every generated response in retrievable, authoritative passages, providing transparent provenance for the information conveyed; and the overall behaviour is evaluated by human clinical experts, ensuring that the technology remains human-centered rather than autonomous.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 presents the system architecture, its LangGraph-based orchestration workflow, and the deployed application. Section 4 describes the construction of the Arabic mental health knowledge base. Section 5 reports the Arabic embedding model benchmark. Section 6 presents the symptom detection module and its evaluation. Section 7 describes the PHQ-9 integration module. Section 8 details the RAG pipeline and its quality evaluation. Section 9 reports the clinical evaluation. Section 10 discusses findings and limitations, and Section 11 concludes.

2. Related Work

This section reviews set of research related to the development of Arabic conversational systems for mental health support. The review focuses on five areas which form the basis of the proposed Rafiq system: mental health chatbots, Arabic mental health technology, Arabic natural language processing (NLP) for the detection of depression, Arabic sentence embeddings, and retrieval-augmented generation (RAG).

2.1. Mental Health Chatbots

Conversational AI systems for mental health have gained significant research attention in recent years. Early rule-based systems have given way to machine-learning and LLM-powered approaches capable of nuanced, open-ended dialogue. Wysa [4] and Woebot demonstrated early feasibility for cognitive-behavioural therapy delivery via chatbots. More recent systems leverage LLMs for empathetic response generation and RAG for evidence-grounded information [12, 13, 14].

Several systems have integrated structured depression screening within chatbot interactions. Marcus [15] and PERLA [16] administer the PHQ-9 or similar instruments via conversational interfaces, while HopeBot [17] evaluates an LLM-based chatbot specifically for interactive PHQ-9 screening. SouLLMate [18] combines RAG, adaptive prompting, and LLMs for diverse mental health support. A systematic review by Otero-González et al. [6] identifies symptom

assessment, psychoeducation, and emotional support as the three principal functions of depression-focused conversational agents, all of which Rafiq integrates.

2.2. Arabic Mental Health Technology

Work specifically addressing Arabic-language mental health AI is notably limited. A systematic review by Alnaghaimshi et al. [2] of Arabic mental health mobile applications found significant gaps in clinical grounding and language quality. Ouali and Garouani [3] surveyed Arabic chatbots and identified low NLP quality and absence of clinical protocols as core limitations. OlloBot [7] is an early Arabic health chatbot focused on healthy living promotion, without depression-specific screening. Recent work on the AraHealthQA 2025 shared task [19, 20] has advanced Arabic mental health question answering using RAG-based approaches, though these focus on question answering rather than conversational screening.

2.3. Arabic NLP for Depression

Arabic NLP for depression detection has developed primarily around text classification on social media corpora. ArDep [21] introduced an Arabic depression lexicon. AraDepSu [22] provides a dataset for depression and suicidal ideation detection. Subsequent work applied AraBERT and MARBERT for multi-label symptom classification [23] and binary depression detection from Twitter [24, 25]. Maghraby and Ali [26] constructed a dataset of 1,228 tweets labelled across the nine PHQ-9 symptom categories, which is used as the evaluation corpus for the symptom detection component of this work. Alasmari [27] recently evaluated LLMs on Arabic mental health datasets. These classification studies inform the symptom detection component of Rafiq but do not address the conversational and clinical integration challenges the present work targets.

2.4. Arabic Sentence Embeddings

Sentence embedding models encode text as fixed-dimensional dense vectors such that semantically similar texts are spatially proximate. Early Arabic NLP relied on static word embeddings such as AraVec [28] and fastText [29], which lack contextual sensitivity. The introduction of AraBERT [30]—a BERT-based model pre-trained on large Arabic corpora—marked a significant advance. Subsequent work adapted the Sentence-BERT (SBERT) framework [31] to Arabic, producing models fine-tuned on contrastive objectives, including Arabic-Triplet-Matryoshka-V2 [32], silma-embedding-matryoshka-v0.1 [33], and Arabic-SBERT-100K [34]. Multilingual alternatives include multilingual-e5-base [35], paraphrase-multilingual-MiniLM-L12-v2 [36], and, more recently, state-of-the-art mixture-of-experts (MoE) embedders such as nomic-embed-text-v2-moe [37]. Existing evaluations, however, target general Arabic NLP benchmarks and do not assess performance in specialized domains such as mental health. Studies such as BEIR [38] and domain-specific biomedical work [39] demonstrate that embedding models transfer unevenly across domains, motivating the domain-specific benchmark constructed in this work.

2.5. Retrieval-Augmented Generation

RAG [40] has become a standard technique for grounding LLM responses in domain-specific evidence, reducing hallucination and improving factual accuracy. RAG-based mental health support systems have been proposed by Han et al. [41] and Haider et al. [12], demonstrating

improved response quality on mental health benchmarks. HopeBot [17] and SouLLMate [18] apply RAG for conversational screening and support, Ilapaka and Ghosh [13] built a RAG system fine-tuned on counselling conversations, and Dutta et al. [42] developed a RAG-based framework for mental health support evaluation. For Arabic, AbdelAziz et al. [20] applied multi-task RAG for mental health question answering. The present work distinguishes itself by constructing a dedicated Arabic mental health knowledge base, systematically evaluating Arabic embedding models for this domain, and embedding the resulting pipeline within a fully integrated conversational system with clinical validation.

3. System Architecture and Orchestration

Rafiq is a multi-component conversational AI system designed around three layers: (i) a knowledge and intelligence layer comprising the knowledge base, embedding model, and LLMs; (ii) an orchestration layer implemented with LangGraph [43]; and (iii) an application layer comprising a FastAPI backend [44] and React Native mobile frontend [45].

Figure 1 illustrates the overall system architecture. Each user message enters the FastAPI conversation processing endpoint and is routed through the LangGraph orchestration workflow. Depending on the current session state, the workflow activates the appropriate pipeline: symptom detection, PHQ-9 administration, or RAG-based response generation. The session state, stored in a `GraphState` object, preserves the conversation history, detected symptom counts, and PHQ-9 progress across turns.

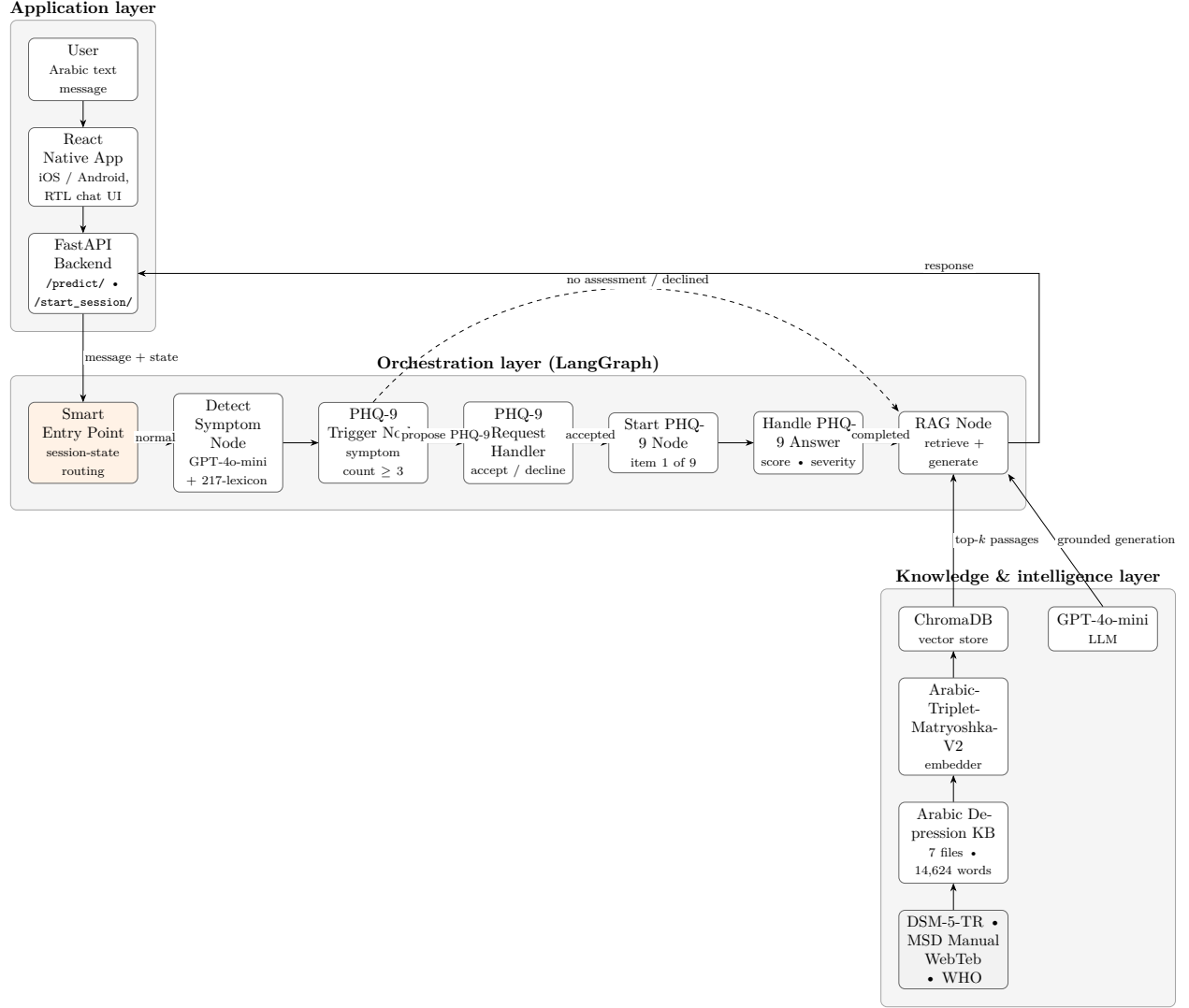


Figure 1: System architecture of Rafiq. The **application layer** forwards each user message together with the session’s **GraphState** to the **orchestration layer**, a LangGraph workflow whose Smart Entry Point routes the turn based on the current PHQ-9 status. In normal flow the message passes through the Detect Symptom Node (GPT-4o-mini with the 217-expression keyword lexicon) and the PHQ-9 Trigger Node; when the accumulated symptom count reaches three, the PHQ-9 Request Handler proposes the assessment, which then proceeds through the Start PHQ-9 and Handle PHQ-9 Answer nodes until completion, at which point the workflow resumes the RAG Node. An ongoing assessment is routed directly to the Handle PHQ-9 Answer Node, and a completed assessment bypasses detection and triggering to prevent re-proposing the PHQ-9. The **knowledge and intelligence layer** supplies the RAG Node with the top- k passages retrieved from the ChromaDB vector store, populated from the Arabic depression knowledge base (DSM-5-TR, MSD Manual, WebTeb, WHO) via the Arabic-Triplet-Matryoshka-V2 embedder, and grounds response generation with GPT-4o-mini.

3.1. LangGraph Workflow

The three modules are integrated into a coherent conversational workflow using LangGraph [43], a framework for building stateful, graph-structured LLM applications. The workflow is represented as a directed graph in which nodes correspond to processing operations and edges define routing logic.

3.2. GraphState

A **GraphState** object serves as the shared session context, carrying: the conversation history (list of human and AI messages); the set of cumulative detected PHQ-9 symptom indices; a flag indicating whether a PHQ-9 session is active; a flag indicating whether a PHQ-9 session has been completed; the current PHQ-9 question index; and the accumulated PHQ-9 scores.

3.3. Nodes

The workflow includes six nodes:

1. **Detect Symptom Node:** Receives the user message and current symptom state. Calls GPT-4o-mini with the keyword-guided symptom detection prompt. Updates the session state with any newly detected symptoms.
2. **PHQ-9 Trigger Node:** Checks whether the accumulated symptom count reaches the threshold (≥ 3 unique symptoms). If so, generates a proposal message inviting the user to complete the PHQ-9 assessment.
3. **PHQ-9 Request Handler Node:** Processes the user's response to the PHQ-9 proposal. Detects acceptance or declination and updates the session state accordingly.
4. **Start PHQ-9 Node:** Loads the first PHQ-9 question from the structured JSON file and delivers it to the user together with its four fixed response options.
5. **Handle PHQ-9 Answer Node:** Processes each PHQ-9 response, stores the score, advances the question index, delivers the next question, or computes and presents the final score and severity interpretation when all nine items are completed.
6. **RAG Node:** Retrieves relevant passages from the ChromaDB knowledge base and calls GPT-4o-mini to generate a contextually appropriate, empathetic response.

3.4. Routing Logic

The workflow employs a Smart Entry Point that evaluates the current session state at the beginning of each conversational turn and routes accordingly:

- **PHQ-9 ongoing:** Routes directly to the Handle PHQ-9 Answer Node, bypassing symptom detection and RAG to ensure questionnaire continuity.
- **PHQ-9 completed:** Routes to the RAG Node, skipping symptom detection and trigger nodes to prevent re-proposing the assessment.
- **Normal flow:** Routes to the Detect Symptom Node, which then proceeds deterministically through the PHQ-9 Trigger Node and, if no assessment is active, terminates at the RAG Node.

This routing logic ensures that structured assessments proceed without interruption while regular conversation remains fluid whenever no assessment is active.

3.5. Application Layer and Deployment

The complete system is exposed through a FastAPI [44] backend and a React Native [45] mobile application. The backend wraps the LangGraph workflow behind two endpoints: `/start_session/` creates a session identifier, and `/predict/` processes each user message against the stored conversation history. The mobile application, built for iOS and Android from a single codebase, provides a minimal, accessible Arabic interface with right-to-left (RTL) layout and switches to structured selectable buttons for PHQ-9 items during assessment, disabling free-text input to enforce structured response collection. User interaction data — the message, the response, and the user’s feedback (like, dislike, or copy) — is logged to Firebase Cloud Firestore [46], providing a foundation for future analysis of user engagement and response quality.

4. Arabic Mental Health Knowledge Base

The knowledge base (KB) provides the factual foundation for the RAG pipeline. A key challenge is the absence of a curated, authoritative Arabic knowledge base suited to retrieval-augmented generation. We develop it through a three-phase methodology: data collection, Markdown conversion with manual review, and data processing.

4.1. Source Selection and Data Collection

We identify six core clinical topics required for comprehensive depression coverage: (1) definition of depression and depressive disorders; (2) types and their symptoms; (3) causes and risk factors; (4) treatment approaches and interventions; (5) depression in special populations (elderly, children, adolescents, postpartum); and (6) patient care management and support strategies.

Four authoritative sources are selected for clinical credibility, Arabic availability, and complementary coverage: the *DSM-5-TR* (American Psychiatric Association, Arabic translation) [47], the *MSD Manual* (Arabic edition) [48], the *World Health Organization* Arabic mental health pages [1], and *WebTeb* [49], a peer-reviewed Arabic medical information portal. Table 1 summarises the selected sources and their contribution.

Table 1: Selected Sources for the Arabic Depression Knowledge Base.

Source	Fmt.	Contribution to KB
DSM-5-TR [47]	PDF	Official diagnostic criteria; classification of depressive disorders
MSD Manual [48]	HTML	Definitions, symptoms, causes, treatment; children/postpartum sections
WebTeb [49]	HTML	Patient-friendly Arabic explanations; coping strategies; lifestyle advice
WHO [1]	HTML	Risk factors, prevention, self-care guidance, help-seeking resources

4.2. Conversion to Markdown and Manual Review

All materials are converted to standardised Markdown to ensure consistency, readability, and compatibility with RAG processing.

HTML sources (MSD Manual, WebTeb, WHO) are converted using the Aspose HTML-to-Markdown tool, which preserves heading hierarchy, lists, and inline formatting without summarisation. Each file undergoes manual review to remove navigation elements, footers, advertisements, and off-topic content, with word-for-word verification against the original pages.

PDF source (DSM-5-TR Arabic edition, 29 pages): automated batch converters produce unsatisfactory results for Arabic text and complex diagnostic tables. We adopt a four-step page-by-page workflow: (1) split the PDF using iLovePDF; (2) convert each page individually using ChatGPT-4 [50] with a structure-preserving extraction prompt that explicitly prohibits summarisation or content modification; (3) supply pages as screenshots when PDF text extraction fails (improving preservation of diagnostic tables and structured lists); (4) concatenate all per-page outputs into a single document. A comprehensive manual review verifies diagnostic table structure, criteria ordering, and Arabic text flow throughout.

4.3. Data Processing and KB Statistics

All converted Markdown files are loaded using LangChain’s `UnstructuredMarkdownLoader` with automatic directory discovery. A lightweight normalisation step standardises line breaks introduced during conversion, preventing chunking artefacts while preserving all content. The final KB consists of 7 Markdown files comprising 14,624 words in total (Table 2).

Table 2: Knowledge Base Content Statistics.

Source	Files	Words
DSM-5-TR	1	5,757
MSD Manual	3	6,293
WebTeb	2	1,278
WHO	1	1,296
Total	7	14,624

The DSM-5-TR provides the most densely structured clinical content (5,757 words); MSD Manual contributes the largest volume across three thematic files (6,293 words); WebTeb and WHO add patient-facing perspectives and public health content. The multi-source design ensures the RAG system can address queries spanning clinical-to-lay user needs.

4.4. Document Chunking

Given that all KB documents underwent comprehensive manual review to ensure proper hierarchical Markdown structure, we adopt a document-structured chunking strategy using LangChain’s `MarkdownHeaderTextSplitter` [43]. The splitter divides documents at three heading levels (`#`, `##`, `###`), annotating each resulting chunk with its full header hierarchy as metadata. This structure-aware segmentation produces variable-length chunks aligned with topical organisation, preserving semantic coherence within each clinical concept. In mental

health contexts, maintaining the integrity of diagnostic criteria, treatment descriptions, and safety guidelines within individual chunks ensures that retrieved passages contain complete, clinically accurate information.

4.5. Embedding Model Selection

The choice of embedding model is critical for Arabic semantic retrieval. Through the systematic benchmark study presented in Section 5, *Arabic-Triplet-Matryoshka-V2* [32] was identified as the best-performing model, achieving MAP = 80%, NDCG = 86%, and Precision@1 = 93%. This model is therefore selected to encode all knowledge base documents and to embed user queries at retrieval time.

5. Benchmarking Arabic Sentence Embeddings for Mental Health Information Retrieval

Selecting an appropriate sentence embedding model is a critical yet underexplored step in building Arabic NLP systems for mental health applications. In a RAG pipeline, the embedding model converts both user queries and knowledge base documents into dense vectors, and retrieval quality is directly determined by this choice [38]. To the best of our knowledge, no prior work has systematically evaluated Arabic embedding models in the mental health domain. This section presents the first benchmark of its kind.

5.1. Benchmark Dataset Construction

A key challenge is the absence of a suitable Arabic evaluation resource. We construct a purpose-built Arabic Mental Health Benchmark Dataset through a three-phase process: prompt design and data generation, manual review and quality assurance, and dataset formatting.

Prompt Design and Data Generation.. We adopt a controlled prompt engineering approach using ChatGPT-4 [50] to generate annotated query–document triplets in Arabic. Each run of the prompt produces one Arabic query representing a realistic mental health information need, followed by three candidate response documents reflecting varying degrees of relevance: *highly relevant* (score 3, directly and specifically answers the query), *moderately relevant* (score 2, related but does not fully address the information need), and *irrelevant* (score 0, thematically unrelated). The generation prompt instructs GPT-4 to produce output in a structured JSON format, ensuring consistency across all examples. Query topics span the core mental health domains relevant to depression, including symptoms, causes, treatment, risk factors, special populations, and coping strategies.

Manual Review and Quality Assurance.. All generated examples undergo manual review to ensure semantic correctness, linguistic clarity, and domain alignment. Relevance labels are verified for consistency with each query; queries are checked to reflect realistic user information needs; documents are reviewed for factual accuracy and clinical appropriateness; and mislabelled, ambiguous, or off-topic items are corrected or replaced.

Dataset Statistics. The final dataset comprises **33 Arabic mental health queries**, each annotated with three candidate documents, yielding **99 query–document pairs** in total. Table 3 summarises the dataset statistics.

Table 3: Arabic Mental Health Benchmark Dataset Statistics.

Metric	Value
Total Queries	33
Responses per Query	3
Total Query–Document Pairs	99
Language	Arabic
Domain	Mental Health
Generation Method	GPT-4 + Manual Validation
Format	JSON
Average Query Length (words)	6.67
Avg. Highly Relevant Doc. Length (words)	27.09
Avg. Moderately Relevant Doc. Length (words)	19.42
Avg. Irrelevant Doc. Length (words)	14.30

5.2. Candidate Embedding Models

We evaluate six Arabic-capable sentence embedding models, selected to represent a range of architectures, training objectives, and language coverage: two general multilingual models, three Arabic-specific models, and one recent multilingual mixture-of-experts (MoE) model. All models are available through the HuggingFace library [51]. Table 4 summarises their key characteristics.

1. **intfloat/multilingual-e5-base** [35]: 278M-parameter, 768-dimensional model pre-trained using contrastive learning on over one billion multilingual text pairs covering approximately 100 languages.
2. **Arabic-Triplet-Matryoshka-V2** [32]: 135M-parameter model fine-tuned from AraBERT v2 [30] using triplet loss on Arabic sentence pairs. Employs Matryoshka Representation Learning [52], enabling dimensionality truncation with minimal performance degradation.
3. **paraphrase-multilingual-MiniLM-L12-v2** [36]: 118M-parameter, 384-dimensional model supporting more than 50 languages, trained on paraphrastic text pairs.
4. **silma-ai/silma-embedding-matryoshka-v0.1** [33]: 135M-parameter Arabic/English bilingual model fine-tuned from AraBERT v2 on over 2.25 million triplet samples.
5. **akhooli/Arabic-SBERT-100K** [34]: 135M-parameter Arabic-specific model fine-tuned on 100,000 Arabic sentence pairs using MatryoshkaLoss.
6. **nomic-ai/nomic-embed-text-v2-moe** [37]: 475M-parameter (305M active) multilingual mixture-of-experts (MoE) model with 8 experts and top-2 routing, supporting approximately 100 languages and trained on 1.6 billion contrastive pairs. It is included as a representative of the current state of the art among general-purpose multilingual embedders.

Table 4: Summary of Candidate Arabic-Processing Embedding Models.

Model	Dimension	Parameters	Training Data	Domain
multilingual-e5-base	768	278M	1B+ multilingual pairs (100+ langs)	General multilingual
Arabic-Triplet-Matryoshka-V2	768 (scalable 128–768)	135M	Arabic sentence pairs	Arabic-specific
paraphrase-multilingual-MiniLM-L12-v2	384	118M	50+ languages paraphrase data	General multilingual
silma-embedding-matryoshka-v0.1	768 (scalable 8–768)	135M	2.25M Arabic/English triplets	Arabic/English bilingual
Arabic-SBERT-100K	768	135M	100K Arabic sentence pairs	Arabic-specific
nomic-embed-text-v2-moe	768 (scalable 256–768)	475M (305M active)	1.6B multilingual pairs (~100 langs)	Multilingual MoE

5.3. Evaluation Methodology

For each candidate model, we encode all 99 documents and all 33 queries into dense vectors. The encoded documents are indexed in ChromaDB [53], accessed via the LangChain framework [43]. For each query, we retrieve the top- k most similar documents using `similarity_search_with_relevance_scores`, which computes cosine similarity between query and document vectors.

We evaluate retrieval quality using five standard information retrieval metrics [54], implemented via `pytrec_eval` [55], a Python interface to the standard TREC evaluation framework. Evaluation is conducted at two cutoff points: $k = 2$ (matching the number of relevant documents per query in the dataset) and $k = 4$ (a broader retrieval window). The metrics are: **Precision@K**, the proportion of relevant documents among the top K results; **Recall@K**, the proportion of all relevant documents retrieved within the top K ; **NDCG@K**, which rewards highly relevant documents appearing earlier in the ranked list; **MAP@K**, the average of precision values at ranks where relevant documents appear, averaged over all queries; and **MRR**, the mean reciprocal rank of the first relevant document.

5.4. Results

Table 5 reports performance of all six models at $k = 2$. **Arabic-Triplet-Matryoshka-V2** achieves the highest scores on MAP (80%), NDCG (86%), Recall@2 (80%), Precision@2 (80%), and Precision@1 (93%). `intfloat/multilingual-e5-base` achieves a perfect MRR of 100, indicating it always ranks at least one relevant document first, but its lower MAP (63%) and Precision@1 (75%) show weaker overall retrieval quality. As an additional experiment, we also evaluated `nomic-ai/nomic-embed-text-v2-moe`; despite its state-of-the-art performance on general English and multilingual retrieval benchmarks, it achieves near-zero scores on our Arabic mental health benchmark (MRR = 9%, MAP = 4%, and Precision@1 = 9% at $k = 2$), indicating that its embedding space does not transfer to domain-specific Arabic text.

Table 5: Performance of Arabic-Processing Embedding Models at $k = 2$ (values in %).

Model	MRR	MAP	NDCG	Recall@2	Precision@2	Precision@1
multilingual-e5-base	100	63	77	63	63	75
Arabic-Triplet-Matryoshka-V2	96	80	86	80	80	93
Arabic-SBERT-100K	93	74	79	74	74	90
paraphrase-multilingual-MiniLM-L12-v2	90	66	74	66	66	81
silma-embedding-matryoshka-v0.1	90	66	73	66	66	75
nomic-embed-text-v2-moe	9	4	4	4	4	9

Table 6 reports results at $k = 4$ for the five models that yielded usable retrieval behaviour in the $k = 2$ evaluation. **Arabic-Triplet-Matryoshka-V2** continues to lead on MAP

(90%), NDCG (93%), and Recall@4 (90%), now sharing the top MRR score (100%) with **multilingual-e5-base**. The broader retrieval window reveals a clearer separation between Arabic-specific models and general multilingual models on MAP and Recall.

Table 6: Performance of Arabic-Processing Embedding Models at $k = 4$ (values in %).

Model	MRR	MAP	NDCG	Recall@4
multilingual-e5-base	100	89	92	89
Arabic-Triplet-Matryoshka-V2	100	90	93	90
Arabic-SBERT-100K	96	80	85	80
paraphrase-multilingual-MiniLM-L12-v2	93	83	86	83
silma-embedding-matryoshka-v0.1	96	83	86	83

Figure 2 provides a visual comparison of the five models with non-degenerate performance across the primary metrics at $k = 2$ (the MoE model is omitted because its near-zero scores would compress the axis). **Arabic-Triplet-Matryoshka-V2** dominates on the three most clinically important metrics—MAP, NDCG, and Precision@1—by a consistent margin of 6–18 percentage points over the next best model.

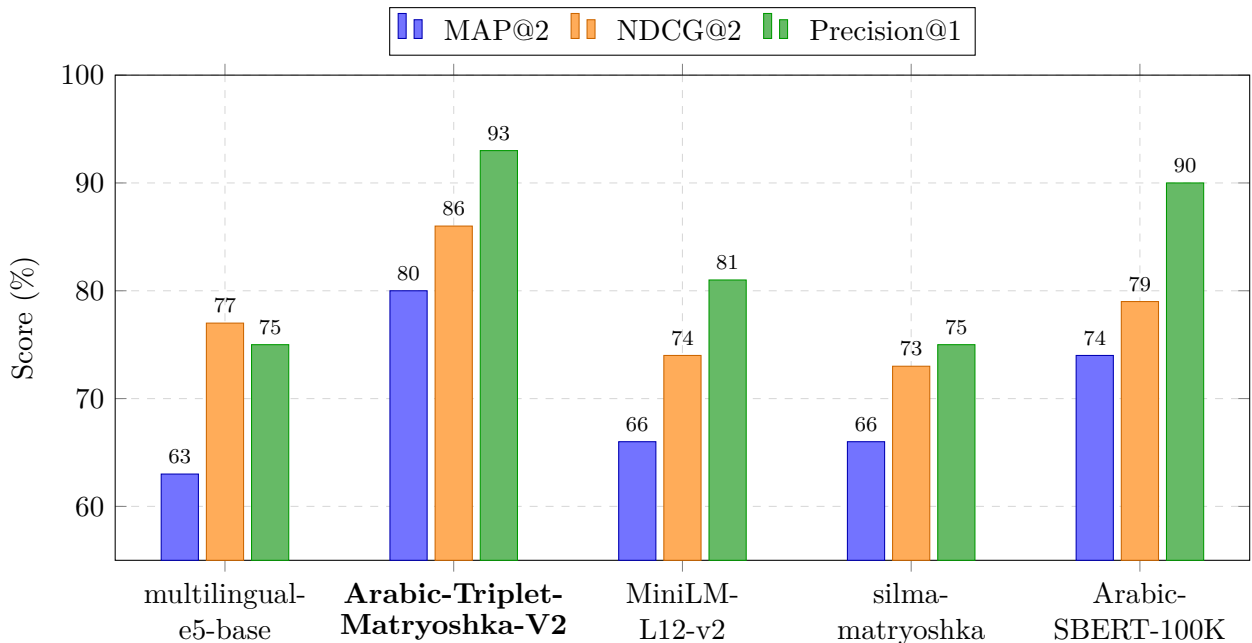


Figure 2: Performance comparison of the five evaluated Arabic-processing embedding models on MAP@2, NDCG@2, and Precision@1 at $k = 2$ (values in %). **Arabic-Triplet-Matryoshka-V2** (second group) achieves the highest scores on all three metrics, outperforming the next best model by 6–18 percentage points. The multilingual MoE model is omitted because its near-zero scores would compress the vertical axis.

The results reveal a consistent pattern: Arabic-specific models trained on Arabic sentence pairs outperform general multilingual models on mental health queries. Mental health vocabulary in Arabic presents particular challenges for multilingual models, involving Arabic-specific morphological forms, dialectal variations, and culturally loaded expressions. Arabic-

Triplet-Matryoshka-V2, fine-tuned specifically on Arabic sentence pairs with triplet contrastive learning, learns a representation space better suited to capturing these distinctions. The exception is `multilingual-e5-base`, which achieves the highest MRR at $k = 2$ (100%)—it is effective at ranking at least one relevant document very highly but less consistent in distinguishing relevance levels across the full ranked list.

The failure of `nomic-embed-text-v2-moe` is particularly instructive. This model is competitive with or superior to the other multilingual models on general English and multilingual retrieval benchmarks, yet on our benchmark it ranks relevant documents near the bottom of the retrieved list ($\text{MAP} = 4\%$). The result underscores that strong general-purpose multilingual performance does not imply adequate transfer to a specialised domain and language pair, and that models must be validated on the target domain before deployment in clinical applications.

On this benchmark, the difference between the worst-performing model (`nomic-embed-text-v2-moe`, $\text{MAP} = 4\%$) and the best (Arabic-Triplet-Matryoshka-V2, $\text{MAP} = 80\%$) represents a 76-percentage-point gap in retrieval quality—a clinically significant difference in a mental health context where incorrect or irrelevant information retrieval can have direct consequences for users. The Matryoshka architecture additionally offers practical benefits: embedding dimensions can be reduced at deployment time (e.g., from 768 to 256 dimensions) with minimal performance degradation, reducing storage and inference costs.

6. Symptom Detection Module

The symptom detection module monitors each user message in real time, classifying it against the nine PHQ-9 depression symptom criteria. Rather than treating symptom detection as a standalone text classification task, the module is framed as a structured prompting task delegated to GPT-4o-mini [56], which offers strong instruction-following capability and cost-efficient inference, and requires no task-specific fine-tuning.

6.1. PHQ-9 Symptom Taxonomy

The classification target consists of nine symptom dimensions derived directly from the PHQ-9 [8] and aligned with the DSM-5 criteria for major depressive disorder [47]:

- S1. Low mood or feelings of sadness
- S2. Markedly diminished interest or pleasure in almost all activities
- S3. Difficulty falling or staying asleep, or sleeping excessively
- S4. Fatigue or loss of energy
- S5. Significant weight loss or gain, or changes in appetite
- S6. Feelings of worthlessness or excessive or inappropriate guilt
- S7. Diminished ability to think or concentrate, or indecisiveness
- S8. Psychomotor agitation or retardation

S9. Recurrent thoughts of death, suicidal ideation, or suicide attempt

A tenth class, *non-depression*, is used to evaluate the system’s ability to correctly identify utterances that express no depressive symptoms.

6.2. Arabic PHQ-9 Keyword Lexicon

To improve detection accuracy in dialectal Arabic, a comprehensive keyword lexicon was constructed in three steps. First, keywords were collected from three prior Arabic depression resources [25, 21, 26] and assigned to the corresponding PHQ-9 symptom dimension. Second, because Arabic morphology requires gender agreement in verbs, adjectives, and participles, the lexicon was expanded to include both masculine and feminine forms of each expression, as well as Levantine (Shami) dialectal variants and gender-neutral expressions. Third, each keyword was manually reviewed to ensure unambiguous assignment to a single symptom category, and the complete lexicon was reviewed and validated by a licensed clinical psychologist.

The resulting lexicon comprises **217 Arabic expressions** in total. Table 7 summarises the keyword count per symptom dimension, together with illustrative examples.

Table 7: Summary of the Arabic PHQ-9 keyword lexicon: keyword count and sample expressions per symptom dimension.

#	Symptom (English)	Keywords	Sample Expressions (Romanised)
1	Low mood / sadness	43	<i>hazin, ta’ban, ya’is, ba’is</i>
2	Anhedonia	32	<i>mush muhtamm, mush musta’mitm</i>
3	Sleep disorder	16	<i>mush qadir anam, nommi maqtu’</i>
4	Fatigue / loss of energy	20	<i>mut’ab, munhak, mush qadir a’mal</i>
5	Weight / appetite change	17	<i>mush mushtahi akl, shehiti izdadat</i>
6	Worthlessness / guilt	24	<i>mush radi ’an nafsi, ana fashil</i>
7	Concentration / decisions	13	<i>mush qadir arkiz, tafkiri mushatat</i>
8	Psychomotor changes	14	<i>hasis bkhomul, a’ani min al-kasal</i>
9	Suicidality	38	<i>biddi amut, laytani kont mayyit</i>
Total		217	

6.3. Dataset Preparation and Leakage-Aware Train–Test Split

Base dataset. We use the Arabic PHQ-9 tweet dataset introduced by Maghraby and Ali [26], which contains 1,228 tweets labelled with one of nine symptom categories corresponding to

PHQ-9 dimensions. The dataset was collected using PHQ-9 symptom keywords as search queries, meaning that a significant overlap exists between the dataset’s keywords and our lexicon. Evaluating the classifier on tweets collected with these same keywords would risk inflating performance metrics through data leakage.

Non-depressive control set.. To evaluate the model’s ability to correctly classify utterances containing no depressive symptom expressions, 50 non-depressive Arabic tweets were generated using GPT-4 [57] with a prompt designed to elicit casual, colloquial Arabic content unrelated to depression, matching the style, tone, and length distribution of the existing dataset and explicitly avoiding any PHQ-9-related vocabulary.

Train–test split with keyword withheld.. The full dataset (1,228 tweets + 50 non-depressive controls) was divided using stratified sampling to maintain class proportions: approximately 70% for training and 30% for testing. This yields a test subset of 365 labelled tweets plus the 50 non-depressive controls, for a total test set of 415 instances.

To prevent leakage, we manually reviewed the test subset and identified all keywords from our lexicon that appear in test-set tweets. These *test keywords* were removed from the keyword dictionary provided to the LLM in the classification prompt. The remaining *train keywords* were retained in the prompt. This split ensures that the classifier cannot rely on exact keyword matches for test instances and must instead apply semantic generalisation to identify symptom expressions phrased differently from the provided examples. Table 8 presents the test set composition.

Table 8: Test set composition and class distribution (415 instances total).

Class	Count
Diminished ability to think or concentrate	81
Weight / appetite disorder	67
Psychomotor agitation or retardation	36
Sleep disorder	35
Low mood	33
Loss of energy	31
Feelings of worthlessness	30
Suicidality	29
Losing interest or pleasure in activities	23
Non-depression	50
Total	415

6.4. Keyword-Guided Prompting

The symptom detection prompt instructs the model to: (1) analyse the user message against the keyword lexicon; (2) return a binary classification for each of the nine PHQ-9 symptoms; (3) output a structured response listing only the newly detected symptoms in the current turn.

The prompt template instructs the model to identify symptom expressions, including synonyms and paraphrases that are semantically equivalent to the provided keywords but not

listed verbatim, and to return *only* a JSON object with the detected symptom labels—an empty symptoms list if no depressive symptom is detected. The {KEYWORDS} placeholder is replaced at runtime with the structured JSON keyword dictionary containing train keywords only, mapping each symptom label to its array of Arabic example expressions. This machine-readable format provides structured guidance while keeping the prompt concise. In evaluation, a prediction is counted as correct if the ground-truth label appears among the predicted symptoms, regardless of any additional labels returned. When the symptoms list is empty, a default “none” label is assigned, treated as a false negative for any non-empty ground-truth class.

6.5. Trigger Logic

Detected symptoms are accumulated across conversational turns in the session state. When three or more *unique* symptoms have been cumulatively identified, the system sets a flag that triggers the PHQ-9 proposal in the subsequent turn. The threshold of three is deliberately set lower than the DSM-5 diagnostic criterion of five symptoms, as reaching three indicates substantial likelihood that additional symptoms may be present but not yet expressed. This threshold balances sensitivity (detecting genuine distress early) against specificity (avoiding premature questionnaire proposals), and the symptom accumulator is reset at the beginning of each new session.

6.6. Evaluation

The symptom detection module was evaluated on the held-out test set of 415 instances described above. Temperature was set to 0 for deterministic output. Table 9 reports the per-class and aggregate classification results, computed with scikit-learn [58].

Table 9: Per-class and aggregate classification results (test set, $n = 415$).

Class	Precision	Recall	F1	Support
Diminished concentration	1.00	1.00	1.00	81
Feelings of worthlessness	0.97	0.97	0.97	30
Losing interest / pleasure	1.00	1.00	1.00	23
Loss of energy	0.97	1.00	0.98	31
Low mood	0.97	0.97	0.97	33
Psychomotor changes	0.97	0.97	0.97	36
Sleep disorder	0.95	1.00	0.97	35
Suicidality	1.00	1.00	1.00	29
Weight / appetite disorder	0.99	1.00	0.99	67
Non-depression	1.00	0.92	0.96	50
Accuracy			0.98	415
Macro avg	0.98	0.98	0.98	415
Weighted avg	0.98	0.98	0.98	415

The classifier achieves an overall accuracy of 0.98 with macro-averaged and weighted-averaged precision, recall, and F1-score all equal to 0.98. Per-class F1-scores range from 0.96

(non-depression) to 1.00 (diminished concentration, anhedonia, and suicidality), indicating consistently strong performance across all symptom dimensions. The lowest F1 score (0.96) is observed for the non-depression class, where recall of 0.92 indicates that 4 of the 50 non-depressive controls were incorrectly assigned a symptom label. This conservative error mode (false positives on non-depressive text) is preferable in a clinical screening context to the reverse (false negatives on symptomatic text).

The leakage-aware evaluation design confirms that this performance reflects genuine semantic generalisation rather than exact keyword matching: the model correctly classifies instances whose specific phrasings were not present in the prompt dictionary.

7. PHQ-9 Integration Module

7.1. Clinical Foundation

The PHQ-9 is a validated, nine-item self-report questionnaire derived from the DSM criteria for major depressive disorder [8]. Each item asks about the frequency of a symptom over the past two weeks on a four-point Likert scale: *not at all* (0), *several days* (1), *more than half the days* (2), and *nearly every day* (3). Total scores range from 0 to 27 and map to five severity categories: minimal (0–4), mild (5–9), moderate (10–14), moderately severe (15–19), and severe (20–27).

An Arabic version of the PHQ-9 has been validated for use in primary care settings [9]. The Rafiq implementation of the questionnaire was reviewed and approved by a licensed clinical psychologist prior to deployment.

7.2. Three-Phase Protocol

The PHQ-9 module operates in three sequential phases:

Phase 1 – Proposal and Consent.. When the symptom accumulation threshold is reached, the PHQ-9 Trigger Node generates a supportive proposal message that contextualises the questionnaire within the conversation, explains its purpose in accessible language, and invites the user to participate voluntarily. The user’s response is checked for acceptance phrases (e.g., *nibda’* [Let’s start], *tamam* [alright], *okei* [okay]). If accepted, the assessment begins; if declined, the conversation resumes normally.

Phase 2 – Questionnaire Administration.. The Start PHQ-9 Node loads the first item from a structured JSON file containing all nine questions with their four fixed response options in Arabic. Each question is delivered as a structured message with selectable buttons in the mobile interface, preventing free-text input that could introduce ambiguity. The Handle PHQ-9 Answer Node processes each response, stores the score, updates the progress index, and delivers the next question until all nine items are completed.

Phase 3 – Scoring and Feedback.. Upon completion, the Handle PHQ-9 Answer Node computes the total score, determines the severity category, and generates a supportive closing message that communicates the result in accessible language without providing a clinical diagnosis, and recommends consulting a mental health professional when appropriate.

7.3. Questionnaire Design Validation

The Arabic PHQ-9 items, response options, and scoring thresholds were reviewed by a licensed clinical psychologist to confirm clinical correctness and culturally appropriate Arabic language use. No changes were required following this review.

8. RAG-Based Information Retrieval

8.1. Architecture

The RAG pipeline follows the standard retrieve-then-generate paradigm [40]. At each conversational turn, the user message is encoded using Arabic-Triplet-Matryoshka-V2 [32] and used to retrieve the most semantically similar document chunks from the ChromaDB [53] vector store. Retrieved passages are injected into the LLM prompt as context, grounding the generated response in authoritative clinical content while maintaining a conversational, empathetic tone.

The retrieval process proceeds in three steps: (1) the query is embedded using the same embedding model as the stored chunks; (2) the query embedding is compared against all stored chunk embeddings using cosine similarity; and (3) the top $k = 4$ most relevant chunks are retrieved. Retrieved chunks are concatenated and inserted into the `{context}` variable of the prompt template.

8.2. LLM Selection

Two OpenAI models were evaluated for response generation: GPT-4o and GPT-4o-mini [57, 56]. RAGAS-based evaluation [59] on a sample of mental health queries demonstrated comparable performance across all metrics, while GPT-4o-mini exhibited notably lower latency and cost. Based on these findings, GPT-4o-mini was selected as the production LLM.

8.3. Prompt Engineering

A carefully engineered system prompt instructs the RAG LLM to act as a supportive digital psychological companion providing non-diagnostic emotional support in Arabic. Key constraints encoded in the prompt include: (i) never use diagnostic terms such as *ikti'ab* (depression) or *idtirab nafsi* (mental disorder); (ii) never give medical advice or medication suggestions; (iii) limit responses to 3–4 short sentences; (iv) maintain a warm, culturally appropriate, conversational Arabic register; (v) always validate the user’s emotional state before offering information; and (vi) refer to professional help whenever the situation warrants.

The prompt also specifies a structured response format: Empathy statement + Simplified information from retrieved context + Open-ended question to maintain dialogue continuity.

8.4. Crisis Management

The system prompt includes explicit crisis management instructions: when a user expresses thoughts of self-harm or suicide, the system shows deep empathy, gently but clearly guides the user toward professional help and crisis resources, and remains supportive and available. This protocol was reviewed and confirmed as clinically appropriate by the evaluating psychologists.

8.5. RAG Evaluation

The RAG pipeline was evaluated using RAGAS [59], a framework for automated evaluation of retrieval-augmented generation systems. RAGAS generates a test set from the curated KB: the KB is represented as a knowledge graph using GPT-4o-mini for processing and Arabic-Triplet-Matryoshka-V2 for embedding, with a custom user persona defined to ensure questions reflect realistic interactions of an Arabic-speaking individual seeking mental health information. A total of 100 Arabic questions with reference answers were generated, ranging from single-hop (38 questions requiring one chunk) to multi-hop reasoning. Each question was processed through the deployed RAG pipeline and answered by both GPT-4o and GPT-4o-mini using an identical minimal prompt.

RAGAS computes five metrics, each normalised to $[0, 1]$: **Context Precision** (proportion of retrieved chunks relevant to the question, weighted by rank); **Context Recall** (proportion of ground-truth context chunks successfully retrieved); **Faithfulness** (proportion of claims in the generated response supported by the retrieved context); **Answer Relevancy** (semantic alignment between the response and the query); and **Answer Correctness** (combined lexical F1 and semantic similarity against the reference answer). Table 10 presents the results.

Table 10: RAGAS performance comparison: GPT-4o-mini vs. GPT-4o.

Model	Faith.	Relev.	Ctx-P	Ctx-R	Corr.
GPT-4o-mini	0.88	0.84	0.91	0.85	0.70
GPT-4o	0.90	0.79	0.91	0.85	0.68

Both models achieve identical retrieval scores (context precision = 0.91, context recall = 0.85), confirming that generation differences are not confounded by retrieval quality. The strong context precision (0.91) validates the structure-aware chunking strategy; high context recall (0.85) indicates that top-4 retrieval recovers 85% of the ground-truth information needed to answer each question, with 88% of questions requiring no more than 2 ground-truth chunks.

On generation metrics, GPT-4o-mini achieves markedly higher answer relevancy (0.84 vs. 0.79) and marginally higher answer correctness (0.70 vs. 0.68), while GPT-4o achieves slightly higher faithfulness (0.90 vs. 0.88). The faithfulness gap (0.02) is small and within typical evaluation variance. The moderate answer correctness of 0.70 reflects the inherent difficulty of exact-match scoring in Arabic, where rich morphology and dialectal variation produce many lexically different but semantically equivalent formulations. Based on these results, **GPT-4o-mini** is selected as the final language model: it operates nearly twice as fast as GPT-4o with substantially lower computational cost while achieving comparable or superior generation quality.

9. Clinical Evaluation

9.1. Methodology

A comprehensive human evaluation study was conducted with five licensed clinical psychologists to assess the clinical appropriateness, safety, and overall quality of Rafiq’s responses. Evaluators were recruited from clinical practice in Gaza, the West Bank, and Lebanon, ensuring diverse professional perspectives across Arabic-speaking communities.

Five distinct conversation scenarios were developed to represent the range of interactions users might experience:

1. Symptom detection followed by PHQ-9 assessment proposal
2. Distinguishing insomnia from depression symptoms
3. Postpartum depression inquiry
4. Suicidal ideation response
5. Adolescent depression inquiry

Each scenario was presented as a complete conversation thread, including user messages and Rafiq’s responses across all relevant pipeline stages. Each clinical evaluator independently reviewed all five scenarios and provided: (i) numerical ratings (1–5 scale) across five evaluation dimensions; (ii) identification of any critical safety concerns; (iii) written qualitative comments on strengths and areas for improvement.

The five evaluation dimensions are defined as follows:

- **Clinical Accuracy:** Alignment of information with established clinical knowledge from DSM-5-TR and PHQ-9 criteria.
- **Clinical Appropriateness:** Suitability of responses for the user’s mental state, including appropriate referrals and avoidance of diagnostic claims.
- **Empathy and Support:** Effectiveness of empathetic communication, emotional validation, and balance of warmth and professionalism.
- **Safety and Risk Management:** Appropriateness of crisis identification and response, including referrals to professional help.
- **Overall Quality:** Holistic assessment across all dimensions.

The evaluation was designed to yield 25 total assessments (5 evaluators $\times$ 5 scenarios).

9.2. Results

Table 11 presents the mean scores across all dimensions and evaluators.

Table 11: Clinical evaluation results (5 psychologists, 5 scenarios, $n = 25$).

Dimension	Mean Score (/ 5.0)
Clinical Accuracy	4.75
Clinical Appropriateness	4.40
Empathy and Support	4.40
Safety and Risk Management	4.65
Overall Quality	4.50
Overall Average	4.54

Clinical Accuracy achieved the highest score (4.75/5.0), reflecting strong alignment with DSM-5-TR and PHQ-9 criteria. Safety and Risk Management scored 4.65/5.0, demonstrating appropriate crisis identification and response protocols across all scenarios including suicidal ideation. Overall Quality scored 4.50/5.0, while Clinical Appropriateness and Empathy and Support both scored 4.40/5.0. Critically, **no evaluators identified critical safety concerns** across any of the five conversation scenarios.

9.3. Qualitative Findings

Evaluators’ written feedback highlighted several notable strengths: Rafiq’s factual accuracy, empathetic conversational tone, methodical PHQ-9 questioning approach, reliance on scientific sources, and avoidance of diagnostic claims. One evaluator explicitly noted that the system “truly feels like a companion (*Rafiq*)”, directly affirming the conceptual framing of the system. Evaluators also commended the non-judgmental acceptance of users and the clear referrals to professional help when appropriate.

Key improvement areas identified included: the need for gathering more contextual information before offering advice; providing more space for user self-expression through open-ended questioning; keeping responses shorter to avoid overwhelming users; ensuring cultural appropriateness of suggested activities; using age-appropriate language for adolescent scenarios; and strengthening crisis protocols with specific local emergency contacts and support resources.

All five evaluators confirmed that Rafiq serves as a useful tool for early depression detection and for helping individuals understand their symptoms and identify appropriate next steps. Three evaluators indicated they would “definitely” or “certainly” recommend the system to patients, with the qualifier that it should complement rather than replace professional consultation.

10. Discussion

10.1. Contributions

Rafiq represents the first reported Arabic conversational AI system to integrate three clinically grounded capabilities – conversational symptom monitoring, structured PHQ-9 administration, and RAG-based psychoeducation – in a unified platform. This integration addresses a documented gap in Arabic mental health technology [2, 3] and advances the state of the art in several respects.

The keyword-guided prompting strategy for Arabic symptom detection, which achieves 98% accuracy without requiring labelled training data for fine-tuning, demonstrates an effective approach for adapting LLM-based classifiers to colloquial dialectal Arabic in resource-constrained domains. The three-phase PHQ-9 integration within a natural conversation is, to our knowledge, the first implementation of its kind in Arabic, extending prior English-language work such as Marcus [15] and HopeBot [17] to the Arabic language context. The clinical validation by licensed psychologists provides a level of evidence absent from most comparable systems.

The embedding benchmark provides evidence-based guidance for model selection that is directly actionable for Arabic health NLP. On the constructed benchmark, the choice

of embedder alone accounts for up to 76 percentage points of MAP, and the identification of Arabic-Triplet-Matryoshka-V2 as the best-performing model directly informs the deployed RAG pipeline. The near-zero performance of the multilingual MoE embedder further demonstrates that advertised multilingual support does not guarantee domain-specific Arabic retrieval quality and that embedding models must be validated on the target domain before deployment in clinical applications. Similarly, the RAGAS evaluation demonstrates that a smaller, faster model (GPT-4o-mini) can match a much larger model on factual grounding, an important finding for real-time deployment.

10.2. Limitations

Several limitations should be acknowledged. First, the symptom detection module was evaluated on a Twitter dataset, which may not fully represent the linguistic characteristics of private chatbot conversations. Second, the clinical evaluation involved five evaluators, which, while sufficient for qualitative validity, is a relatively small sample for quantitative generalisation. Third, the system currently targets Palestinian colloquial Arabic and may exhibit reduced performance on markedly different Arabic dialects such as Moroccan or Gulf Arabic. Fourth, no user study with individuals experiencing depressive symptoms has been conducted, which is necessary to assess real-world usability and clinical impact.

The embedding benchmark is modest in size (33 queries) relative to large-scale benchmarks such as BEIR, and its candidate documents were generated by GPT-4 rather than extracted from real clinical texts. The additional MoE model was evaluated without its recommended task-specific prefixes (e.g., “search_query:”), which may partially account for its near-zero scores; a follow-up study controlling for prompt formatting would clarify this interaction. The KB covers only depression and does not include anxiety, bipolar disorder, or other common mental health conditions. RAGAS evaluation relies on automated metrics, which may not fully capture clinical appropriateness or cultural sensitivity. Finally, the symptom detection framework does not yet account for negation handling (e.g., “I am *not* sad”), which could generate false positives in adversarial or nuanced inputs.

10.3. Ethical Considerations

Rafiq is explicitly designed as a supportive companion and screening tool, not as a diagnostic or therapeutic system. The system does not provide clinical diagnoses or treatment recommendations, and all conversation responses emphasise the importance of professional mental health consultation. Crisis management protocols follow clinical best practices and were validated by licensed psychologists. Interaction data collected via Firebase is used solely for research analysis and system improvement.

10.4. Future Work

Based on clinical evaluator feedback and the identified limitations, several directions for future development are prioritised. On the model side: expanding dialect coverage through dialect-adaptive prompting; implementing more comprehensive crisis protocols including direct links to regional mental health emergency contacts; providing more individualised, contextually sensitive responses by gathering richer user background before generating advice; and incorporating explicit negation handling in the symptom detection prompt. On the evaluation side: conducting user studies with individuals experiencing depressive symptoms

to assess real-world impact on help-seeking behaviour; expanding clinical validation to include psychologists from a broader range of Arabic regions; extending the embedding benchmark to additional mental health conditions and real clinical documents; evaluating emerging Arabic LLM embeddings; re-testing multilingual MoE models under their recommended usage conditions; and implementing robust safety guardrails for crisis scenarios.

11. Conclusion

This paper has presented Rafiq, an end-to-end Arabic conversational AI system for early depression detection and mental health support, together with the complete research pipeline underlying it. We constructed a 14,624-word Arabic depression knowledge base from four authoritative sources using a three-phase curation methodology; we presented the first systematic benchmark of Arabic sentence embedding models for mental health retrieval, identifying Arabic-Triplet-Matryoshka-V2 as the best-performing embedder (MAP = 80%, Precision@1 = 93% at $k = 2$) and demonstrating that even a state-of-the-art multilingual MoE model fails to transfer to this domain (MAP = 4%); and we developed a keyword-guided LLM symptom detection framework achieving 98% accuracy and an F1-score of 0.98 under leakage-aware evaluation.

By integrating LLM-based symptom detection, automated PHQ-9 administration, and retrieval-augmented generation within a LangGraph-orchestrated workflow, Rafiq provides a clinically grounded and culturally appropriate conversational experience for Arabic-speaking users. The RAG pipeline achieved faithfulness of 0.88, answer relevancy of 0.84, and context precision of 0.91 on RAGAS evaluation, and the selection of GPT-4o-mini demonstrates that strong factual grounding can be delivered at substantially lower latency and cost. Clinical validation by five licensed psychologists yielded a mean overall score of 4.54/5.0 with no safety concerns identified, and all evaluators confirmed the system’s utility for early depression detection.

Rafiq’s performance demonstrates the feasibility of conversational AI for mental health support in Arabic-speaking communities, and establishes a baseline and methodology for future work in this clinically important and underserved area. We hope this work encourages further research and development of culturally appropriate Arabic-language mental health technologies that support early intervention while maintaining appropriate clinical boundaries.

Statements and Declarations

Funding. The authors did not receive support from any organization for the submitted work. No funds, grants, or other support was received.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethics approval and consent to participate. The system was evaluated by five licensed psychologists on scripted conversation scenarios in their professional capacity; no patients, personal data, or human subjects research were involved, and no informed consent was required.

Consent for publication. Not applicable.

Data availability. In line with the journal’s research data policy, the Arabic mental health knowledge base and the evaluation datasets will be made freely available in a public repository upon acceptance of this manuscript. Reviewer access can be granted on request.

Code availability. Implementation details are described in the manuscript. The source code will be made publicly available in a public repository upon acceptance, consistent with the journal’s open and FAIR software sharing practices; reviewer access can be granted on request.

Authors’ contributions. Alaa Alshurafa, Basem Alijla, and Motaz K H SAAD conceptualised the study. Alaa Alshurafa developed the system, conducted the experiments, and wrote the original draft. Basem Alijla and Motaz K H SAAD supervised the research and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Plagiarism statement. The authors confirm that this manuscript is an original work and has not been published previously, nor is it under consideration for publication elsewhere. All sources used in the preparation of this manuscript have been appropriately cited and referenced. The manuscript does not contain plagiarized content from previously published articles.

Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] World Health Organization, Depressive disorder (depression), <https://www.who.int/news-room/fact-sheets/detail/depression>, accessed: 2025 (2025).
- [2] N. I. S. Alnaghaimshi, et al., A systematic review of features and content quality of Arabic mental health mHealth apps, *Frontiers in Digital Health* 6 (2024) 1472251.
- [3] S. Ouali, S.-E. Garouani, Arabic chatbots challenges and solutions: A systematic literature review, *Iraqi Journal for Computer Science and Mathematics* 5 (3) (2024) 8.
- [4] B. Inkster, S. Sarda, V. Subramanian, An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study, *JMIR mHealth and uHealth* 6 (11) (2018) e12106.
- [5] M. Casu, S. Triscari, S. Battiato, et al., AI chatbots for mental health: A scoping review of effectiveness, feasibility, and applications, *Applied Sciences* 14 (13) (2024) 5889.
- [6] I. Otero-González, M. R. Pacheco-Lorenzo, M. J. Fernández-Iglesias, L. E. Anido-Rifón, Conversational agents for depression screening: A systematic review, *International Journal of Medical Informatics* 181 (2024) 105272.
- [7] A. Fadhil, OlloBot – towards a text-based Arabic health conversational agent: Evaluation and results, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)* (2019).

- [8] K. Kroenke, R. L. Spitzer, J. B. W. Williams, The PHQ-9: Validity of a brief depression severity measure, *Journal of General Internal Medicine* 16 (9) (2001) 606–613.
- [9] S. M. Al-Hadi, et al., Validity of the Arabic version of the Patient Health Questionnaire-9 (PHQ-9) for the detection of depression in primary care, *Annals of General Psychiatry* 16 (2017) 32.
- [10] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion* 58 (2020) 82–115.
- [11] M. Aljohani, J. Hou, S. Kommu, X. Wang, A comprehensive survey on the trustworthiness of large language models in healthcare, in: *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 6720–6748.
- [12] T. Haider, N. Shafi, N. U. Haq, AI-driven mental health chatbot: Empowering well-being with conversational AI and retrieval-augmented generation, in: *2025 IEEE Conference on Artificial Intelligence*, 2025.
- [13] A. Ilapaka, R. Ghosh, A comprehensive RAG-based LLM for AI-driven mental health chatbot, in: *2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA)*, 2025.
- [14] G. Soman, M. Judy, A. M. Abou, Human guided empathetic AI agent for mental health support leveraging reinforcement learning-enhanced retrieval-augmented generation, *Cognitive Systems Research* 90 (2025) 101337.
- [15] P. Toulme, J. Nanaw, P. Apostolellis, Marcus: A chatbot for depression screening based on the PHQ-9 assessment, *Technical Report* (2023).
- [16] R. Arrabales, Perla: A conversational agent for depression screening in digital ecosystems, *Design, Implementation and Validation* (2020).
- [17] Z. Guo, A. Lai, J. Ive, et al., Development and evaluation of HopeBot: An LLM-based chatbot for structured and interactive PHQ-9 depression screening, *arXiv preprint arXiv:2507.05984* (2025).
- [18] Q. Guo, J. Tang, W. Sun, et al., SouLLMate: An application enhancing diverse mental health support with adaptive LLMs, prompt engineering, and RAG techniques, *arXiv preprint arXiv:2410.16322* (2024).
- [19] H. Alhuzali, D. Alasmari, H. Alsaleh, et al., AraHealthQA 2025: The first shared task on Arabic health question answering, in: *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks*, 2025.
- [20] R. AbdelAziz, M. A. Youssef, M. M. Koritam, et al., Arabic mental health question answering: A multi-task approach with advanced retrieval-augmented generation, in: *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks*, 2025.

- [21] B. Bataineh, R. Duwairi, M. Abdullah, ArDep: An Arabic lexicon for detecting depression, in: Proceedings of the 3rd International Conference on Advances in Artificial Intelligence, 2019.
- [22] M. Hassib, N. Hossam, J. Sameh, M. Torki, AraDepSu: Detecting depression and suicidal ideation in Arabic tweets using transformers, Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP) (2022).
- [23] S. Elmajali, I. Ahmad, Toward early detection of depression: Detecting depression symptoms in Arabic tweets using pretrained transformers, IEEE Access 12 (2024) 88134–88145.
- [24] E. M. Rabie, A. F. Hashem, F. Alsheref, Depressive state detection model in Arabic user-generated content, Data in Brief 41 (2022) 107999.
- [25] A. Alaskar, M. Ykhlef, Depression detection from Arabic tweets using machine learning techniques, Journal of Computer Science, Software Development (2021) 1–10.
- [26] A. Maghraby, M. Ali, Modern standard Arabic mood changing and depression dataset, Data in Brief 41 (2022) 107999.
- [27] D. Alasmari, A scoping review of Arabic natural language processing for mental health, Healthcare (2025).
- [28] A. B. Soliman, K. Eissa, S. R. El-Beltagy, AraVec: A set of Arabic word embedding models for use in Arabic NLP, in: Procedia Computer Science, Vol. 117, 2017, pp. 256–265.
- [29] E. Grave, P. Bojanowski, P. Gupta, et al., Learning word vectors for 157 languages, in: Proceedings of LREC, 2018.
- [30] W. Antoun, F. Baly, H. Hajj, AraBERT: Transformer-based model for Arabic language understanding, arXiv preprint arXiv:2003.00104 (2020).
- [31] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-Networks, in: Proceedings of EMNLP, 2019, pp. 3982–3992.
- [32] O. Nacar, et al., Arabic-Triplet-Matryoshka-V2: General Arabic text embedding for enhanced semantic textual similarity, HuggingFace Model Hub, <https://huggingface.co/0martificial-Intelligence-Space/Arabic-Triplet-Matryoshka-V2> (2025).
- [33] SILMA AI, silma-embedding-matryoshka-v0.1, HuggingFace Model Hub, <https://huggingface.co/silma-ai/silma-embedding-matryoshka-v0.1> (2024).
- [34] A. Akhooli, Arabic-SBERT-100K, HuggingFace Model Hub, <https://huggingface.co/akhooli/Arabic-SBERT-100K> (2024).
- [35] L. Wang, N. Yang, X. Huang, et al., Text embeddings by weakly-supervised contrastive pre-training, arXiv preprint arXiv:2212.03533 (2022).

- [36] Dataloop, paraphrase-multilingual-MiniLM-L12-v2 model card, HuggingFace Model Hub (2024).
- [37] Z. Nussbaum, B. Duderstadt, Training sparse mixture of experts text embedding models, arXiv preprint arXiv:2502.07972, <https://arxiv.org/abs/2502.07972> (2025).
- [38] N. Thakur, N. Reimers, A. Rücklé, et al., BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: Proceedings of NeurIPS Datasets and Benchmarks, 2021.
- [39] Q. Chen, A. Rankine, et al., Domain-specific sentence embedding for biomedical information retrieval, Journal of Biomedical Informatics 141 (2023) 104367.
- [40] P. Lewis, E. Perez, A. Piktus, et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 9459–9474.
- [41] Y. Han, C. Liu, P. Wang, A comprehensive survey on vector database: Storage and retrieval technique, challenge, arXiv preprint arXiv:2310.11703 (2023).
- [42] A. Dutta, S. Mruthyunjaya, J. Saddington, K. S. Islam, Mentalic Net: Development of RAG-based conversational AI and evaluation framework for mental health support, arXiv preprint arXiv:2509.04456 (2025).
- [43] LangChain, LangChain overview, <https://python.langchain.com/docs/> (2025).
- [44] FastAPI, FastAPI: Modern, fast (High-Performance) web framework for building APIs with Python, <https://fastapi.tiangolo.com> (2025).
- [45] Meta, React native, <https://reactnative.dev> (2025).
- [46] Google, Cloud Firestore, <https://firebase.google.com/docs/firestore> (2025).
- [47] American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR), American Psychiatric Association, 2022.
- [48] MSD Manual, MSD manual (Arabic Edition), <https://www.msdmanuals.com/ar> (2025).
- [49] WebTeb, WebTeb: Arabic medical information portal, <https://www.webteb.com> (2025).
- [50] OpenAI, GPT-4 technical report, <https://openai.com/research/gpt-4> (2025).
- [51] T. Wolf, et al., HuggingFace’s Transformers: State-of-the-art natural language processing, arXiv preprint arXiv:1910.03771 (2019).
- [52] A. Kusupati, et al., Matryoshka representation learning, Advances in Neural Information Processing Systems 35 (2022).
- [53] Chroma, ChromaDB: The open-source embedding database, <https://www.trychroma.com/> (2025).

- [54] L. Monigatti, Evaluating RAG systems: Information retrieval metrics, Towards Data Science (2024).
- [55] C. Van Gysel, M. de Rijke, `pytrec_eval`: An extremely fast Python interface to `TREC_eval`, in: Proceedings of SIGIR, 2018, pp. 873–876.
- [56] OpenAI, GPT-4o Mini: Advancing cost-efficient intelligence, <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/> (2024).
- [57] OpenAI, Hello GPT-4o, <https://openai.com/index/hello-gpt-4o/> (2024).
- [58] scikit-learn, Metrics and scoring: Quantifying the quality of predictions, https://scikit-learn.org/stable/modules/model_evaluation.html (2025).
- [59] Ragas, RAGAS: Evaluation framework for retrieval-augmented generation, <https://docs.ragas.io/en/stable/> (2025).