

Article

Empathy-Driven Arabic Conversational Chatbot Using a Pre-Trained Transformer Model

Sarah Masoud Alyami *, Nasser A. Alsadhan and Mohamed Maher Ben Ismail

Computer Science Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia; nsadhan@ksu.edu.sa (N.A.A.); mbenismail@ksu.edu.sa (M.M.B.I.)

* Correspondence: 443203223@student.ksu.edu.sa

Abstract

Recent advancements in sequence generation models have transformed the development of conversational chatbots, enabling more dynamic and emotionally aware interactions. While English-language chatbots have achieved notable progress through large language models (LLMs), Arabic-language systems continue to face significant challenges, particularly in handling dialectal variation, morphological complexity, and generating emotionally aligned responses. This paper introduces two innovative approaches to enhance empathetic response generation in Arabic conversational AI. The first, Emotion-Driven Response Generation (EDRG), employs a two-stage pipeline: it first classifies user emotions using marBERT and then routes inputs to the most suitable Arabic LLM (AraBERT, AraELECTRA, AraGPT-2, or MT5) for contextually appropriate response generation. The second, EmoLlama, is a Retrieval-Augmented Generation (RAG)-based framework that integrates a curated knowledge base with the LLaMA model to retrieve relevant conversational contexts before generating semantically rich and empathetic responses. To support these approaches, a large-scale open-domain Arabic dataset was curated, containing over 600,000 dialogue entries spanning empathetic and neutral responses across seven Ekman-based emotion categories. Experimental evaluations using BLEU, Perplexity (PPL), and Cosine Similarity metrics validated the effectiveness of our models. EDRG achieved strong BLEU scores across multiple emotions, reflecting high lexical alignment, while also attaining a Cosine Similarity of 0.51. In contrast, EmoLlama significantly outperformed in semantic similarity, achieving a Cosine Similarity of 0.91, demonstrating its superior ability to generate contextually and semantically rich responses. These results highlight the complementarity of lexical and semantic metrics in evaluating emotionally intelligent Arabic dialogue systems.

Keywords: Arabic natural language processing; empathy-driven chatbots; emotion classification; retrieval-augmented generation; large language models; response generation; conversational AI



Academic Editor: Chilukuri K. Mohan

Received: 3 June 2026

Revised: 16 June 2026

Accepted: 20 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

The rapid advancement of digital technologies has transformed human interaction with information, services, and communication systems. Among these innovations, conversational agents, commonly known as chatbots, have emerged as powerful tools that enable seamless user engagement, real-time assistance, and streamlined access to digital platforms [1]. Leading technology companies such as Google, Facebook, and Microsoft have invested heavily in the development of chatbot technologies, recognizing their potential to augment or even replace traditional interfaces by providing intuitive, conversational access to services [2,3].

Modern chatbots leverage Natural Language Processing (NLP), a fundamental branch of Artificial Intelligence (AI), combined with Machine Learning (ML) and Deep Learning (DL) techniques to interpret and respond to human language. Early systems such as ELIZA (1966) [4] and PARRY (1972) [5] pioneered rule-based approaches to simulate human conversation. However, contemporary chatbots have evolved into sophisticated neural-based systems capable of handling multi-turn dialogues, performing context-aware reasoning, and generating human-like responses. Applications now span diverse domains, including customer support [6], healthcare [7], education [8], and tourism. One example is Saudi Arabia's Rahhal chatbot, which was designed to enhance visitor experiences [9].

Despite these advances, the development of Arabic-language chatbots remains in its infancy. Arabic poses unique challenges for NLP owing to its rich morphology, orthographic ambiguity, and diglossic nature across Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA) [10–12]. These linguistic complexities, coupled with the scarcity of high-quality annotated datasets, hinder the creation of robust Arabic NLP systems and limit the adaptability of existing chatbot frameworks to Arabic-speaking contexts.

Emotion recognition and empathy are emerging as essential components for conversational agents, especially in sensitive domains such as mental health, education, and customer engagement. Studies such as [13] demonstrate that systems capable of detecting and responding to user emotions foster more natural and engaging interactions. However, many current systems employ superficial sentiment detection, which fails to capture nuanced emotional expressions, especially in Arabic, where cultural and linguistic subtleties profoundly shape how emotions are conveyed. Furthermore, existing Arabic chatbots often overlook the importance of handling neutral or ambiguous user inputs effectively, a critical aspect of maintaining conversational flow.

Advancements in Large Language Models (LLMs) and transformer-based architectures offer promising solutions to these challenges. Pre-trained models such as BERT, GPT, and T5 have revolutionized NLP by enabling context-aware understanding and generation across diverse languages. Arabic-specific models, including AraBERT, AraELECTRA, and AraGPT-2, have further improved performance in tasks such as sentiment analysis and text generation. Yet, these models still face limitations in generating empathetic responses that account for users' emotional states and cultural contexts.

This paper introduces two novel approaches designed to advance empathetic Arabic conversational AI:

- **Emotion-Driven Response Generation (EDRG):** A two-stage pipeline that first classifies user emotions using a dedicated Arabic emotion classifier (marBERT) and then routes the input to the most suitable Arabic LLM for response generation (AraBERT, AraELECTRA, AraGPT-2, or MT5). This design aims to ensure high emotional alignment and contextual relevance.
- **EmoLlama:** A Retrieval-Augmented Generation (RAG) framework that integrates a knowledge base with the LLaMA model. This method retrieves contextually relevant information before generating responses, enabling richer semantic content and enhanced conversational depth.

To support these approaches, a comprehensive open-domain Arabic dataset was curated from multiple sources, combining existing datasets with newly generated dialogues, and featuring both human- and AI-generated empathetic responses. This dataset addresses data imbalance across emotions and dialects, providing a solid foundation for training and evaluating empathy-driven systems.

The models were rigorously evaluated using standard metrics, including BLEU, Perplexity, and Cosine Similarity. Results indicate that EDRG achieves strong lexical alignment, while EmoLlama demonstrates superior semantic understanding and contextual coherence.

Together, these contributions mark a significant step toward overcoming the cultural, linguistic, and emotional challenges unique to Arabic NLP [14,15] and advancing the field of empathetic conversational AI [16,17].

2. Literature Review

This section reviews existing literature across four key themes relevant to this study: English-language chatbots, Arabic-language chatbots, emotion classification using large language models (LLMs), and Retrieval-Augmented Generation (RAG). Each subsection highlights significant research contributions, methodologies, findings, and gaps that motivate our proposed approaches.

2.1. English Chatbots

Early chatbot systems such as ELIZA [4] and PARRY demonstrated rule-based pattern-matching capabilities but lacked true natural language understanding (NLU) and emotional intelligence. Conceptual frameworks like Schank's Conceptual Dependency Theory [18] paved the way for cognitive models, yet remained limited in adaptability.

The advent of deep learning marked a paradigm shift. Meena [19] leveraged a 2.6 billion parameter transformer model trained on 40 billion words, achieving a Sensibleness and Specificity Average (SSA) score of 79% in human evaluations. BlenderBot [20] improved upon this by integrating retrieval-based mechanisms with a generative model to reduce hallucinations and increase factual grounding, reporting a 67% BLEU score and higher human ratings for conversational consistency.

Healthcare applications have benefited from empathetic chatbot designs. This study [21] developed a GPT-2-based generative chatbot for detecting early signs of depression through therapeutic dialogues, achieving a BLEU score of 0.79 and a 30% improvement in perceived empathy and response relevance compared to traditional Seq2Seq models. Similarly, another study [22] developed a Transformer-based empathetic chatbot that achieved a BLEU score of 0.81 and an emotion recognition accuracy of 85%, leading to a 25% increase in user satisfaction compared to traditional chatbots.

Despite these achievements, most research remains English-centric, leaving challenges in multilingual and morphologically rich languages unaddressed. This limitation motivates the exploration of Arabic conversational systems.

2.2. Arabic Chatbots

Arabic chatbots face unique challenges due to morphological complexity, diglossia, and diverse dialects. This study [23] addressed these challenges through personality and emotion estimation from Arabic social media posts, offering foundational insights into how dialect and affective features can be modeled in conversational AI.

Rule-based systems like Botta [6] leveraged AIML to facilitate Egyptian dialect conversations, but were constrained by their static knowledge bases. Rahhal [9] supported tourism in Saudi Arabia through predefined templates, achieving a 70% user satisfaction rate during pilot tests. However, these systems lacked adaptability and struggled with unstructured inputs.

Generative-based chatbots have also emerged. DZchatbot [24], built using Seq2Seq architectures with GRU and LSTM encoders, handled Algerian dialect healthcare queries with an F1-score of 0.85. Another study [25] fine-tuned AraBERT within a BERT2BERT architecture for empathetic dialogue generation, achieving a BLEU score of 0.55 and a human-rated empathy score of 4.3/5.

Hybrid systems are rare but promising. Nabiha [26], a Saudi dialect chatbot for IT students, combined AIML-based templates with dynamic retrieval, achieving faster

response times and higher user satisfaction (85%). Similarly, ArRASA [27], built on the RASA framework, combined intent detection and entity extraction for domain-specific tasks, achieving 96% accuracy in intent classification.

While these studies highlight progress, they also underscore several limitations. In particular, there remains a lack of emotionally adaptive systems capable of understanding nuanced emotions in Arabic dialogue. Our work addresses this gap by integrating emotion classification and retrieval-augmented generation for Arabic.

2.3. Emotion Classification Using LLMs

Emotion classification is critical for empathetic chatbot design. Early approaches relied on lexicon-based methods, but lacked contextual understanding. Transformer-based models like BERT [28] and ELECTRA [29] revolutionized emotion detection in English, achieving over 90% accuracy on datasets like GoEmotions.

For Arabic, fine-tuned models such as AraELECTRA [30] and AraGPT-2 [31] have shown promising results. MARBERT [32], pre-trained on Arabic Twitter data, achieved 95.2% accuracy in Egyptian dialect emotion classification tasks, outperforming traditional methods by over 20%. Studies applying MARBERT to sarcasm detection and political discourse demonstrated its advanced contextual understanding in highly nuanced texts.

Recent multilingual LLMs, including DeepSeek-R1 [33], have extended these capabilities to over 20 languages, showing comparable zero-shot emotion classification performance to GPT-3.5 in Arabic-specific benchmarks.

While progress is evident, existing models still struggle with low-resource dialectal variations and multi-turn emotional contexts. Our proposed EDRG approach builds on MARBERT's strengths to enhance emotion-driven response generation.

2.4. Retrieval-Augmented Generation (RAG)

RAG frameworks combine dense retrievers with generative models, providing context-aware responses by accessing external knowledge during generation [34]. They reported a 25% improvement in factual grounding for open-domain QA systems. LongRAG [35] extended this by handling documents up to 4000 tokens, achieving an F1-score of 88% on HotpotQA.

In conversational AI, Dialogue-RAG [36] introduced a dialogue-aware retrieval module, improving multi-turn coherence by 15%. Domain-specific RAG systems for healthcare and legal chatbots demonstrated 85–90% accuracy in providing contextually relevant answers [37,38].

For Arabic, progress is limited. JAIS-30B Chat [39], developed with bilingual training data, improved logical reasoning tasks by 19% in Arabic. Arabic Copilot [40] tailored RAG for enterprise applications, achieving a 14% increase in document relevance. MedRAG-Ar [41] addressed medical queries in Arabic with 90% accuracy.

These studies confirm RAG's potential for enhancing chatbot capabilities. Our EmoLlama approach extends this by integrating RAG with LLAMA to generate emotionally aligned responses in Arabic. To the best of our knowledge, this represents an early attempt to apply RAG with LLAMA-based models for Arabic empathetic response generation.

3. Proposed Approach

In this study, we propose two complementary approaches for developing an empathy-driven Arabic conversational chatbot. These approaches are designed to address the linguistic complexity and cultural nuances of the Arabic language while enhancing emotional awareness in dialogue systems. To provide a solid foundation for these approaches, we first present an overview of the key language models employed in this study.

This section is organized as follows. We begin with an overview of the core language models and technologies that underpin both approaches, including MARBERT, AraBERT, AraELECTRA, AraGPT-2, MT5, LLAMA, and Retrieval-Augmented Generation (RAG). We then describe the first approach, Emotion-Driven Response Generation (EDRG), in Section 3.2, including its emotion classification and response generation components. Finally, we present EmoLlama in Section 3.3, a retrieval-augmented framework that integrates LLAMA with a curated knowledge base to produce semantically grounded and empathetic responses.

3.1. Overview of Adopted Models

3.1.1. MARBERT

MARBERT (Morphologically Aware BERT for Arabic) is a transformer-based model fine-tuned on over one billion Arabic tweets. It excels in handling Arabic dialects and informal text, making it particularly effective for emotion classification tasks. MARBERT leverages the BERT architecture with 12 transformer encoder layers and 12 self-attention heads (Figure 1).

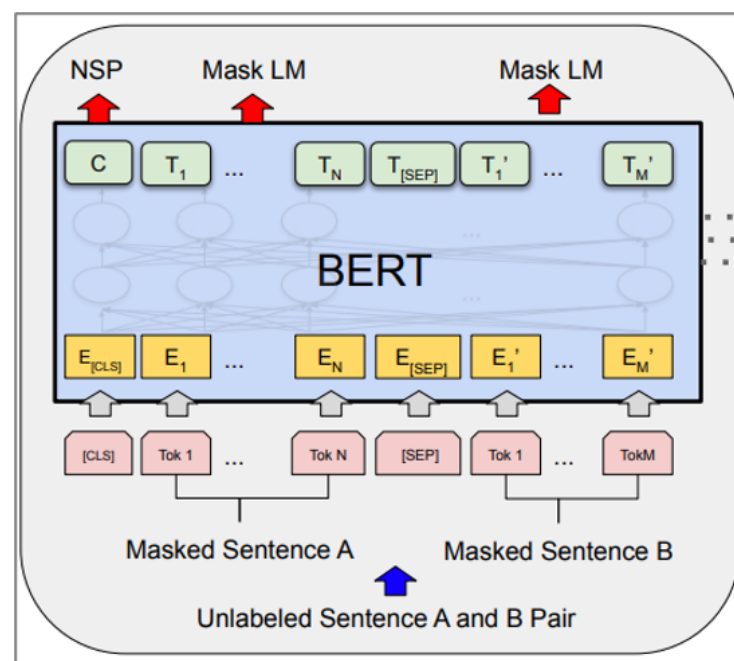


Figure 1. BERT Architecture, the basis for MARBERT [28].

3.1.2. AraBERT

AraBERT, based on Google's BERT, is pre-trained on 77 GB of Arabic text including news articles, books, and Wikipedia. It uses bidirectional encoding to understand context, making it ideal for generating responses aligned with user sentiment. AraBERT adopts the same architecture as BERT-base, consisting of 12 encoder layers, 12 attention heads, and 768 hidden units, as illustrated in Figure 1.

While both MARBERT and AraBERT share the BERT-base architecture, MARBERT is trained on informal Arabic tweets and excels in handling dialects, making it ideal for emotion classification. In contrast, AraBERT is pre-trained on formal Arabic sources, making it more suitable for generating coherent responses in structured conversations.

3.1.3. AraELECTRA

AraELECTRA adopts a replaced token detection mechanism, enabling efficient pre-training. Its architecture (Figure 2) consists of 12 encoder layers and is optimized for token-level understanding.

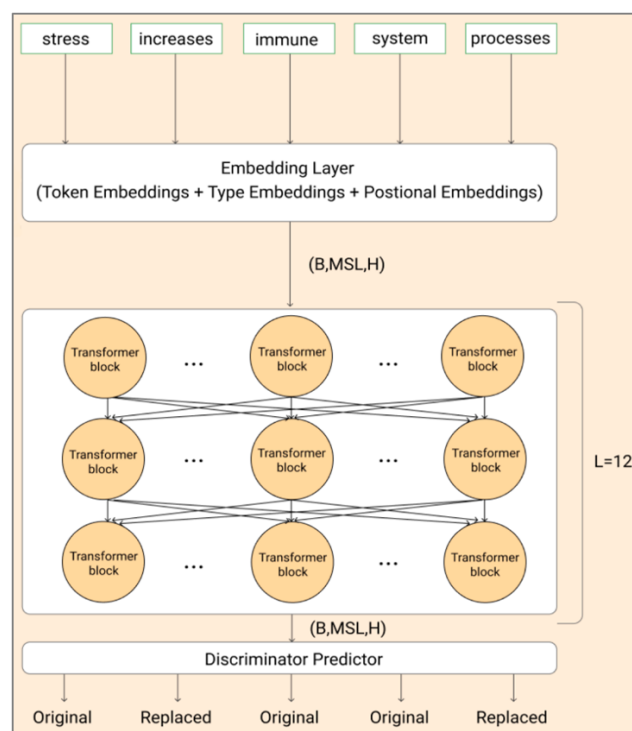


Figure 2. ELECTRA Architecture, foundation for AraELECTRA [29].

3.1.4. AraGPT-2

AraGPT-2 is a transformer-decoder model designed for Arabic text generation. With its autoregressive architecture, it predicts each subsequent token, producing fluent, contextually rich text. The model is illustrated in Figure 3.

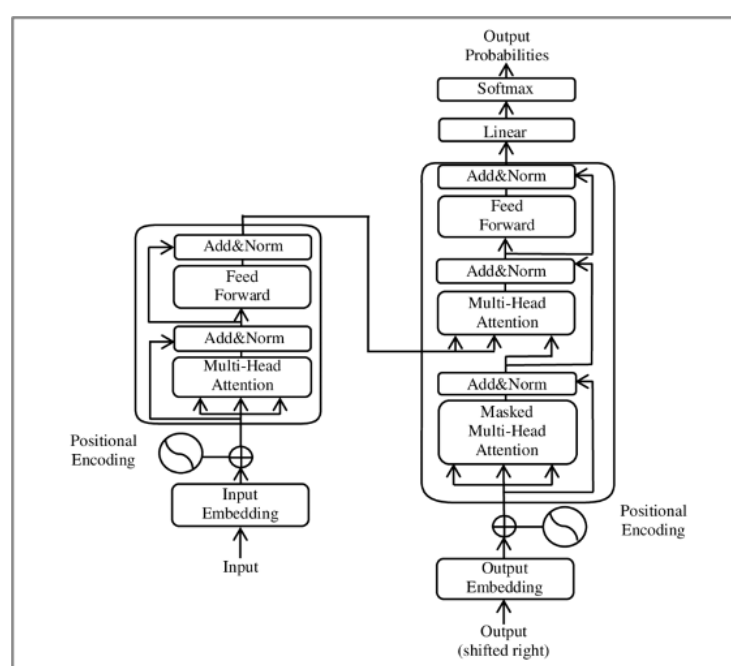


Figure 3. Transformer Architecture, utilized in AraGPT-2 [42].

3.1.5. MT5

MT5 extends Google's T5 architecture to multilingual contexts, supporting Arabic alongside over 100 other languages. Its encoder–decoder design frames all NLP tasks as text-to-text, enabling versatile response generation. The encoder processes the input while the decoder generates the corresponding output sequence, following the architecture described by Raffel et al. [43], as illustrated in Figure 4. While MT5 is not exclusively trained on Arabic, its exposure to multilingual corpora enables it to generalize well across various dialects and forms of Modern Standard Arabic, making it a strong candidate for generating empathetic and coherent responses in Arabic dialogue systems.

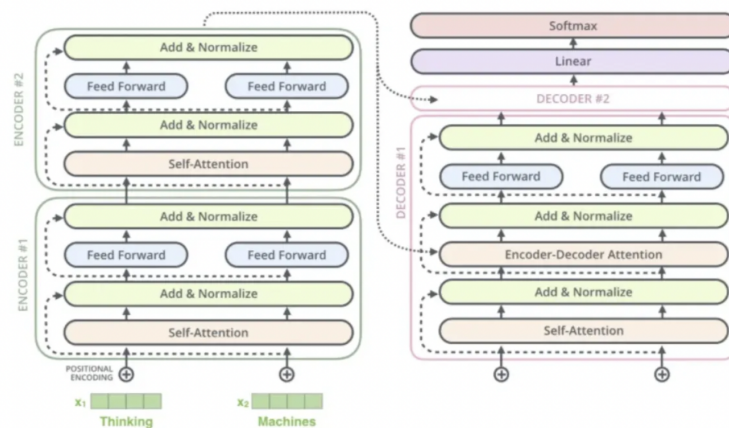


Figure 4. Transformer Encoder–Decoder Architecture.

3.1.6. LLAMA

LLAMA (Large Language Model Meta AI) is a decoder-only transformer architecture developed by Meta AI, optimized for text generation tasks. It utilizes components such as Rotary Positional Encoding, Grouped Multi-Query Attention, and SwiGLU activation in the feed-forward layers, allowing efficient and scalable generation across long sequences. Figure 5 illustrates the core structure of LLAMA's decoder block, as proposed in the official LLAMA architecture [44].

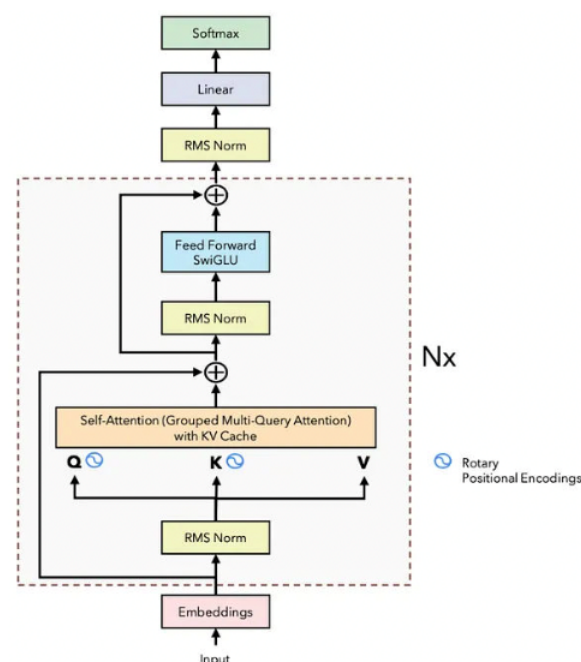


Figure 5. Decoder-only architecture used in LLAMA models [44].

3.1.7. Retrieval-Augmented Generation (RAG)

RAG enhances generative models by retrieving relevant documents from external knowledge bases. Its architecture (Figure 6) integrates a retriever and a generator, enabling context-rich responses.

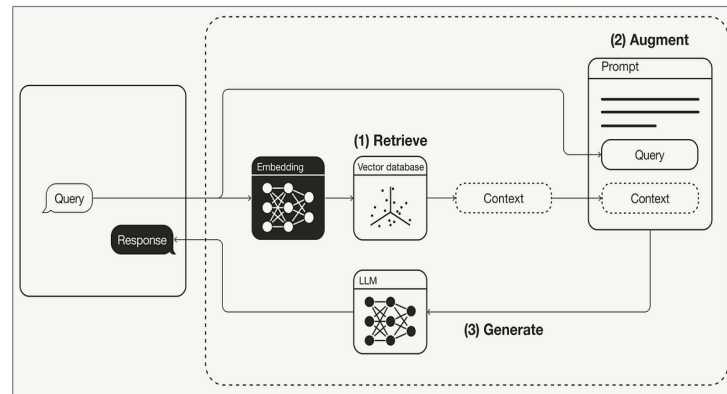


Figure 6. RAG Architecture Model [34].

3.2. Emotion-Driven Response Generation Approach

The EDRG approach is structured as a two-stage pipeline. The first stage classifies user emotions using MARBERT, while the second stage generates responses using a dynamically selected Arabic LLM, such as AraBERT, AraELECTRA, AraGPT-2, or MT5. Each of these models is carefully selected for its unique capabilities in handling Arabic linguistic features and emotional expressions.

3.2.1. Emotion Classification with MARBERT

MARBERT, a variant of BERT optimized for Arabic, serves as the foundation for emotion classification due to its robustness in understanding both Modern Standard Arabic (MSA) and dialectal variations. MARBERT was fine-tuned on our custom empathy-focused dataset, which includes diverse emotional expressions across Arabic dialects. The pre-processing pipeline employed Farasa Segmenter for tokenization, handling clitics and diacritics effectively to ensure linguistic consistency.

3.2.2. Response Generation with Arabic LLMs

Upon detecting the user's emotion, the system dynamically assigns one of four fine-tuned Arabic LLMs for response generation.

The selection of the optimal model for each emotion is informed by BLEU scores obtained during the validation phase. For instance, AraGPT-2 may be selected for generating responses to joyful emotions, while AraELECTRA may be more effective in addressing nuanced negative emotions. The overall workflow of the proposed EDRG approach is illustrated in Figure 7.

3.3. RAG-Based Empathetic Responses Approach Using LLAMA

To complement EDRG, the second approach, EmoLlama, integrates Retrieval-Augmented Generation (RAG) with LLAMA. This design enables the chatbot to retrieve relevant contextual information from an external knowledge base, enhancing its ability to provide emotionally aligned and contextually appropriate responses.

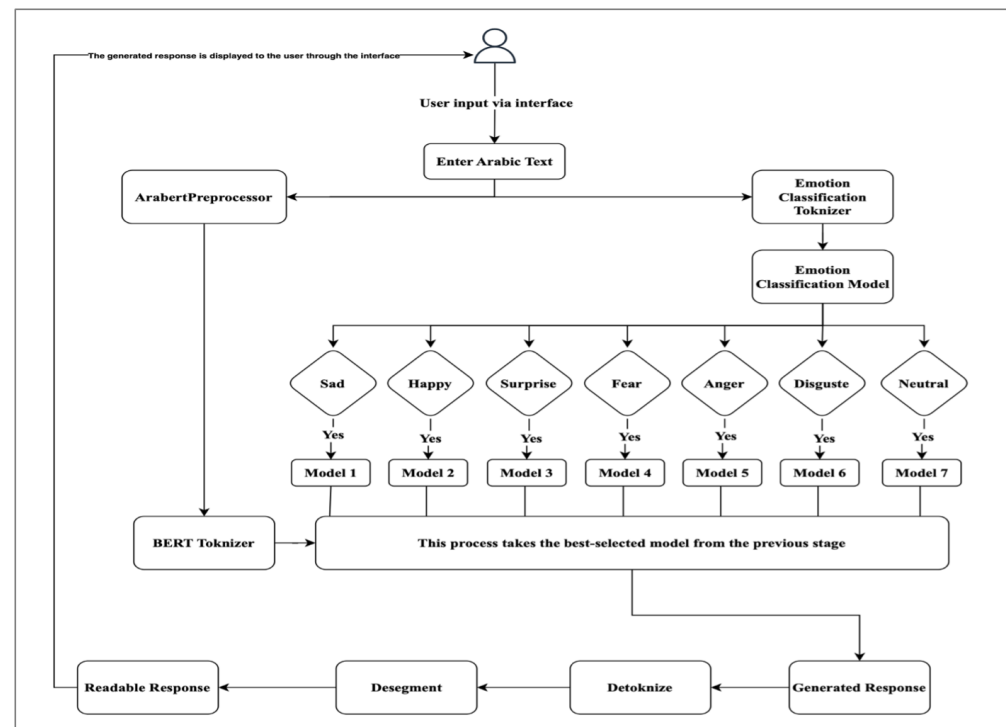


Figure 7. Architecture of the proposed EDRG approach.

3.3.1. Integration of Retrieval-Augmented Generation (RAG)

RAG architecture consists of two primary components: (i) a retriever that searches for contextually relevant documents and (ii) a generator that produces responses conditioned on both user input and retrieved content. In EmoLlama, markdown-based datasets containing user inputs, responses, and emotions are preprocessed and chunked using RecursiveCharacterTextSplitter. Ollama Embeddings embed these chunks, which are stored in a Chroma vector database for efficient retrieval.

When a user query is received, it is embedded and matched against the Chroma database to retrieve top-ranked relevant documents. This retrieval step mitigates knowledge cut-off issues in LLAMA and provides domain-specific context to the generator.

3.3.2. Response Generation with LLAMA

In EmoLlama, LLAMA processes a prompt constructed from both the retrieved documents and user query using a customized Arabic prompt template. To ensure conversational coherence over multiple turns, ConversationBufferMemory is employed, preserving dialogue history. This facilitates emotionally consistent and context-aware interactions, as illustrated in Figure 8.

By combining LLAMA's generative capabilities with RAG's retrieval mechanism, EmoLlama addresses limitations in EDRG, such as reliance on pre-classified emotions and static knowledge. It dynamically adapts responses based on real-time contextual retrieval, ensuring empathy and relevance even in multi-turn conversations.

Together, these two approaches provide complementary strategies for enhancing empathy and contextual understanding in Arabic conversational AI. EDRG offers precise emotion alignment, while EmoLlama delivers contextual richness and flexibility through retrieval-based augmentation.

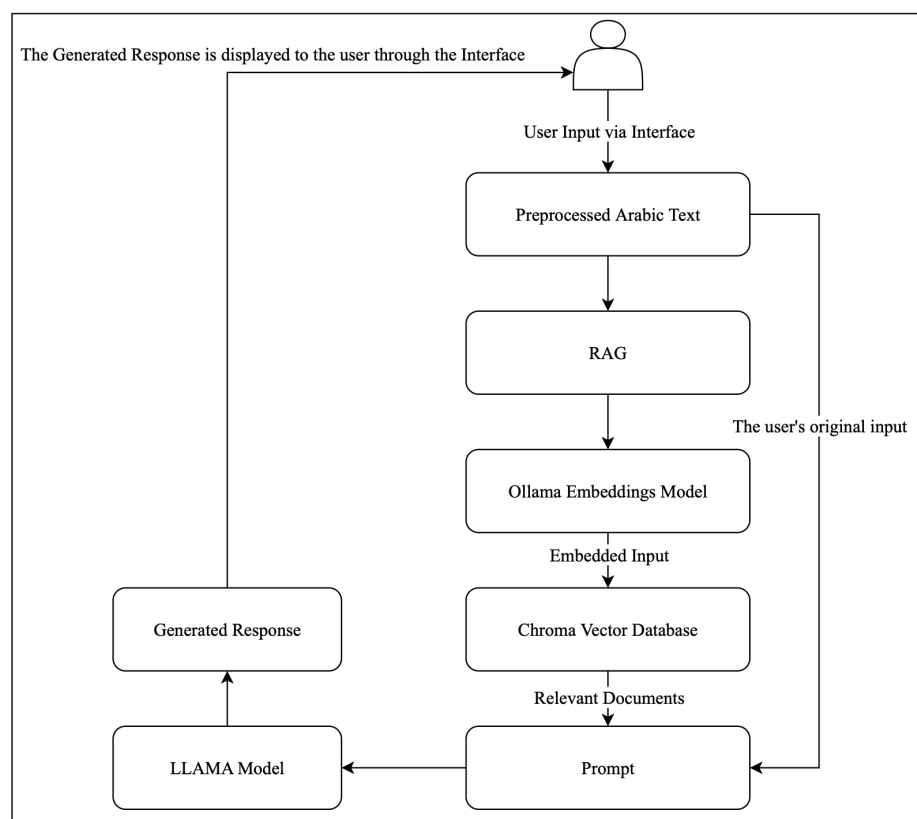


Figure 8. EmoLlama architecture.

4. Experiments and Results

This section presents the experiments and results for the proposed approaches: Emotion-Driven Response Generation (EDRG) and EmoLlama. It covers dataset collection, preparation, training procedures, and a comparative analysis of model performance.

4.1. Dataset and Experimental Setup

A custom Arabic conversational dataset was developed for this study, comprising 622,903 user-input, response, and emotion triples. The data was collected and generated from multiple sources, combining both human-authored and AI-generated dialogues to enrich diversity and coverage. The dataset spans seven Ekman-based emotions: happiness, sadness, anger, fear, disgust, surprise, and neutral. This distribution supports the development of balanced and empathetic dialogue models. Table 1 illustrates the post-balancing distribution across emotion categories.

Table 1. Emotion distribution after class balancing.

Emotions	Human-Generated Records	AI-Generated Records	Total Records	Percentage
Happiness	88,536	4587	93,123	14.95
Sadness	68,274	22,598	90,872	14.59
Anger	30,320	58,405	88,725	14.24
Fear	27,365	61,318	88,683	14.24
Neutral	12,397	76,010	88,407	14.19
Surprise	11,787	76,324	88,111	14.15
Disgust	7451	77,531	84,982	13.64
Total	246,130	376,773	622,903	100.00

4.2. Experimental Results

The marBERT model, fine-tuned for emotion classification, demonstrated a significant improvement after balancing the dataset. Table 2 summarizes its performance, showing accuracy increasing from 77.52% to 85.19% and F1-score from 77.50% to 85.59%. These results highlight the importance of dataset balancing in improving classification accuracy and generalization.

Table 2. marBERT Classification Performance Before and After Balancing.

Setting	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
Imbalanced Data	77.52	77.50	78.10	77.00
Balanced Data	85.19	85.59	86.35	85.19

To provide a more detailed class-wise evaluation of the marBERT classifier, Table 3 and Figure 9 present the row-normalized confusion matrix computed on the human-generated test set. The results show that the model achieved the highest recognition rates for Neutral (98.75%), Disgust (94.04%), and Surprise (90.71%), while maintaining strong performance for Anger (81.39%) and Fear (81.98%). The lowest recognition rates were observed for Sadness (73.91%) and Happiness (75.01%). Furthermore, the confusion matrix reveals that the largest proportion of misclassifications occurred between these two emotion categories, with 17.90% of Sadness samples classified as Happiness and 15.20% of Happiness samples classified as Sadness. Overall, these results provide additional insight into the classifier's behavior across individual emotion categories and complement the overall performance metrics reported in Table 2.

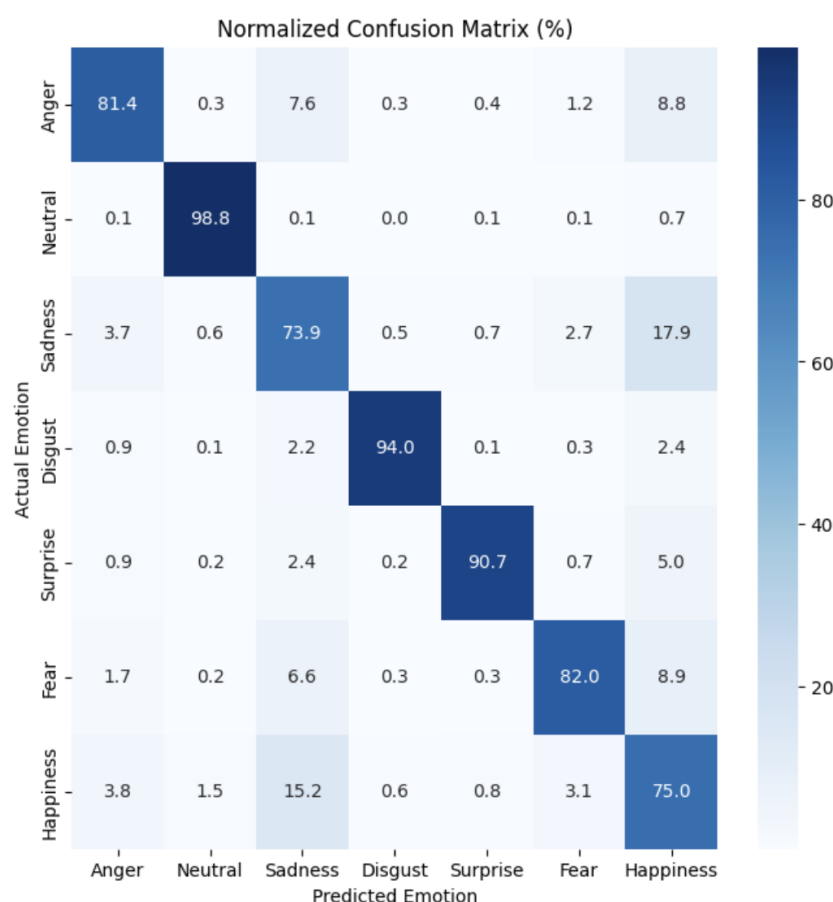


Figure 9. Heatmap Visualization of the Row-Normalized Confusion Matrix for the marBERT Emotion Classification Model.

Table 3. Row-Normalized Confusion Matrix (%) for the marBERT Emotion Classification Model on the Human-Generated Test Set.

Actual	Anger	Neutral	Sadness	Disgust	Surprise	Fear	Happiness
Anger	81.39	0.32	7.65	0.32	0.35	1.21	8.76
Neutral	0.14	98.75	0.13	0.02	0.11	0.13	0.73
Sadness	3.67	0.57	73.91	0.53	0.71	2.70	17.90
Disgust	0.86	0.09	2.19	94.04	0.11	0.28	2.43
Surprise	0.87	0.16	2.41	0.15	90.71	0.66	5.05
Fear	1.72	0.20	6.59	0.29	0.29	81.98	8.94
Happiness	3.77	1.52	15.20	0.59	0.82	3.08	75.01

The dataset was used to train both the emotion classification and response generation models. Preprocessing included emoji removal, normalization, and segmentation, with tokenization performed using model-specific tokenizers. For EDRG, the balanced dataset was split into training (70%), validation (15%), and testing (15%) sets using stratified sampling.

During fine-tuning, we performed a grid search over hyperparameters such as batch size, learning rate, and epochs to optimize classification accuracy. Early stopping with a patience parameter of two epochs was applied to prevent overfitting. The model outputs logits for seven emotion classes: happiness, sadness, anger, fear, disgust, surprise, and neutral. It then selects the class with the highest probability.

Decoding configurations, including beam search with num_beams = 6, top-k sampling (top_k = 50), and enabling do_sample = True, were applied to balance coherence and diversity in the generated responses.

It is important to note that model evaluation was conducted on a human-generated test subset to assess generalization to real-world user inputs.

Five models were evaluated for response generation. These included four Arabic LLMs (AraBERT, AraELECTRA, AraGPT-2, and MT5) and the EmoLlama retrieval-augmented model, which was based on qwen2:7b-text embeddings. Table 4 presents the BLEU and Cosine Similarity scores for each model. AraBERT and AraELECTRA demonstrated the strongest performance among the LLMs, achieving BLEU scores of 0.57 and 0.56 and Cosine Similarity of 0.51 each. MT5 showed the weakest performance, with a BLEU score of 0.28 and Cosine Similarity of 0.29. While EmoLlama had a lower BLEU score (0.34), it significantly outperformed all other models in terms of semantic similarity, achieving a Cosine Similarity score of 0.91. This result underscores the effectiveness of retrieval-augmented generation in capturing user intent and producing semantically rich responses.

Table 4. BLEU and Cosine Similarity Scores for Response Generation Models.

Model	BLEU (Average)	Cosine Similarity
AraBERT	0.57	0.51
AraELECTRA	0.56	0.51
AraGPT-2	0.38	0.48
MT5	0.28	0.29
EmoLlama	0.34	0.91

To further understand the effectiveness of each model across specific emotions, we analyzed the BLEU scores per emotion category. Table 5 presents the results, highlighting how AraBERT and AraELECTRA models consistently outperformed others across most emotion types.

Table 5. BLEU Scores per Emotion Category for Response Generation Models.

Model	Anger	Disgust	Fear	Happy	Neutral	Sadness	Surprise
AraBERT	0.63	0.57	0.63	0.55	0.53	0.57	0.51
AraELECTRA	0.64	0.55	0.62	0.53	0.52	0.57	0.50
AraGPT-2	0.39	0.27	0.44	0.39	0.35	0.39	0.27
MT5	0.30	0.25	0.32	0.28	0.29	0.32	0.28

The results indicate that the highest BLEU score was observed for the anger emotion using the AraELECTRA model (0.64), followed closely by AraBERT (0.63). AraBERT outperformed other models in fear, happiness, and surprise, making it particularly effective for a broader range of emotions. AraELECTRA achieved similar results, especially in sadness and neutral, matching AraBERT with a BLEU score of 0.57 for sadness. In contrast, AraGPT-2 and MT5 showed significantly lower scores across all emotions, highlighting the superiority of the fine-tuned encoder-based models.

These emotion-specific BLEU scores support the core design of the Emotion-Driven Response Generation (EDRG) approach, where the system dynamically selects the optimal response model based on the detected emotion in the user input. For instance, if the detected emotion is fear, the system will prioritize AraBERT as the generator due to its superior performance for that emotion. In cases where two models exhibit similar performance (e.g., AraBERT and AraELECTRA for sadness), the system resorts to additional metrics such as the percentage of sentiment alignment (introduced in the next paragraph) between the user input and the generated response to resolve tie-breaking. This mechanism ensures both emotional relevance and contextual coherence in real-time response generation.

To enhance the emotion-alignment evaluation, we introduce the sentiment match percentage, which measures how well the sentiment of the generated response aligns with the user's original input. This metric is particularly useful when two models yield similar BLEU scores for the same emotion category.

As shown in Table 6, although both models perform similarly on BLEU scores, the sentiment match percentage provides further insights. For instance, while both AraBERT and AraELECTRA achieved a BLEU score of 0.57 for sadness, AraELECTRA showed a higher sentiment match (47.46%) compared to AraBERT (38.21%). Such comparisons allow the system to make informed decisions when choosing between models with otherwise similar performance.

Table 6. Sentiment Match Percentage for araBERT and araELECTRA Models.

Model	Anger %	Disgust %	Fear %	Happy %	Neutral %	Sadness %	Surprise %
araBERT	68.40	67.94	58.15	39.23	69.71	38.21	59.89
araELECTRA	68.73	67.04	53.80	36.56	69.34	47.46	68.88

In Table 7, the EmoLlama model demonstrates strong sentiment alignment for Neutral and Surprise, supporting its effectiveness in retrieval-augmented response generation. However, the model shows moderate alignment for other emotions, suggesting that while semantic relevance is high, emotional matching can still be further optimized.

Table 7. Sentiment Match Percentage for EmoLlama Approach.

Model	Anger %	Disgust %	Fear %	Happy %	Neutral %	Sadness %	Surprise %
EmoLlama	65.32	66.21	50.45	35.28	75.89	38.62	70.12

In addition to evaluating BLEU scores and sentiment match percentages, a deeper analysis was conducted to examine how accurately each model matched the target emotion

of the user input. This was done by computing the percentage of correctly matched emotions, as well as identifying the most frequent mismatches for each emotion category.

Tables 8, 9, and 10 present the percentages of correctly matched and mismatched emotions for the araBERT, araELECTRA, and EmoLlama models, respectively.

This analysis reveals common confusion patterns among models. For instance, the emotion Happiness was frequently mismatched with Surprise and Neutral, particularly in the MT5 and araGPT-2 models, highlighting the difficulty of distinguishing between closely related emotional tones. Conversely, Neutral and Anger were consistently recognized with higher accuracy across most models.

These insights are valuable for understanding the limitations of each model and guiding future improvements in emotion recognition and generation. Importantly, high BLEU and Cosine scores do not necessarily imply that the predicted responses align with the user's intended emotion. Measuring this alignment provides a complementary perspective, as demonstrated in Tables 8–10, and highlights a critical avenue for future work.

Table 8. Correct and Mismatched Emotion Percentages—**araBERT** Model.

Emotion	Correct %	Mismatched %
Anger	68.4	Sadness: 12.3, Disgust: 10.4, Fear: 3.1, Surprise: 2.7, Neutral: 2.2, Happy: 1.3
Disgust	67.9	Anger: 13.2, Sadness: 9.7, Fear: 3.8, Surprise: 2.4, Neutral: 1.9, Happy: 1.1
Fear	58.2	Sadness: 15.2, Surprise: 12.4, Disgust: 5.7, Neutral: 4.2, Anger: 2.9, Happy: 1.4
Happy	39.2	Surprise: 19.7, Neutral: 15.3, Sadness: 9.5, Fear: 7.9, Anger: 5.1, Disgust: 3.0
Neutral	69.7	Happy: 12.2, Surprise: 8.3, Sadness: 4.8, Fear: 2.3, Disgust: 2.1, Anger: 1.6
Sadness	38.2	Fear: 19.6, Anger: 15.2, Neutral: 9.8, Surprise: 8.2, Disgust: 5.7, Happy: 2.3
Surprise	59.9	Happy: 17.9, Fear: 7.8, Neutral: 7.2, Sadness: 4.3, Anger: 2.1, Disgust: 0.9

Table 9. Correct and Mismatched Emotion Percentages—**araELECTRA** Model.

Emotion	Correct %	Mismatched %
Anger	68.7	Sadness: 12.6, Disgust: 10.8, Fear: 2.6, Surprise: 2.2, Neutral: 1.9, Happy: 1.2
Disgust	67.0	Anger: 12.7, Sadness: 11.2, Fear: 3.7, Surprise: 2.5, Neutral: 1.9, Happy: 1.0
Fear	53.8	Sadness: 16.5, Surprise: 12.6, Disgust: 6.3, Neutral: 4.9, Anger: 3.4, Happy: 2.0
Happy	36.6	Surprise: 22.3, Neutral: 15.0, Sadness: 10.8, Fear: 8.2, Anger: 4.1, Disgust: 3.1
Neutral	69.3	Happy: 12.7, Surprise: 7.9, Sadness: 5.1, Fear: 2.7, Disgust: 2.0, Anger: 1.3
Sadness	47.5	Fear: 17.8, Anger: 13.5, Neutral: 9.5, Surprise: 7.3, Disgust: 3.8, Happy: 1.6
Surprise	68.9	Happy: 11.7, Fear: 8.5, Neutral: 5.2, Sadness: 3.2, Anger: 2.1, Disgust: 1.1

Table 10. Correct and Mismatched Emotion Percentages—**EmoLlama** Model.

Emotion	Correct %	Mismatched %
Anger	65.32	Sadness: 12.14, Disgust: 10.40, Fear: 5.20, Surprise: 3.47, Neutral: 2.08, Happy: 1.39
Disgust	66.21	Anger: 10.81, Sadness: 9.46, Fear: 5.41, Surprise: 3.38, Neutral: 2.70, Happy: 2.03
Fear	50.45	Sadness: 16.35, Surprise: 14.86, Disgust: 6.94, Anger: 4.96, Neutral: 3.96, Happy: 2.48
Happy	35.28	Surprise: 22.0, Neutral: 16.83, Sadness: 10.36, Fear: 6.47, Anger: 5.18, Disgust: 3.88
Neutral	75.89	Happy: 8.20, Surprise: 6.27, Sadness: 3.86, Fear: 2.41, Disgust: 1.93, Anger: 1.45
Sadness	38.62	Fear: 17.82, Anger: 15.06, Neutral: 9.04, Surprise: 5.03, Disgust: 3.42, Happy: 1.83
Surprise	70.12	Happy: 10.77, Fear: 7.47, Neutral: 4.98, Sadness: 3.11, Anger: 2.49, Disgust: 1.06

Overall, AraBERT and AraELECTRA demonstrated superior performance for emotion-aligned response generation. Meanwhile, EmoLlama excelled in maintaining semantic consistency and adapting to multi-turn conversations, highlighting its strength in real-world conversational AI applications where contextual grounding is critical. This comparison underscores the complementary nature of both approaches in addressing the challenges of empathetic dialogue systems for Arabic.

4.3. Ethical Considerations

The proposed chatbot frameworks raise several ethical considerations that should be acknowledged. While the Emotion-Driven Response Generation (EDRG) approach

and EmoLlama employ different response generation mechanisms, both systems interact with users in emotionally sensitive contexts and may influence user perceptions and decision-making.

First, the EDRG approach relies on automatic emotion classification to guide response generation. Although the emotion classifier achieved strong performance, misclassification remains possible and may lead to responses that do not appropriately reflect the user's actual emotional state. This limitation is particularly important in sensitive domains where emotional understanding is critical.

Second, emotional alignment does not necessarily imply that the generated response should mirror the user's detected emotion. For example, a response to sadness may be more effective when providing support and encouragement rather than simply reproducing the same emotional tone. Therefore, future work should further investigate the relationship between detected emotions and the most contextually appropriate response strategies.

Third, both EDRG and EmoLlama may be affected by biases present in the training and evaluation data. The datasets used may not fully represent all Arabic dialects, cultural backgrounds, and communication styles. As a result, system performance may vary across different user groups. Expanding emotion datasets and improving dialectal coverage are important steps toward reducing potential biases and improving fairness.

Finally, the proposed systems are intended as conversational support tools and should not be considered substitutes for professional medical, psychological, or counseling services. Human oversight remains essential in high-risk applications, particularly those involving mental health or sensitive personal situations. In addition, appropriate user awareness and transparency measures should be maintained, as AI-generated responses may occasionally contain inaccuracies or inappropriate interpretations.

5. Conclusions and Future Work

This study introduced and evaluated two complementary approaches for Arabic emotion-aware chatbot response generation. The first approach, Emotion-Driven Response Generation (EDRG), employs several fine-tuned Arabic large language models (LLMs), including marBERT, AraBERT, AraELECTRA, AraGPT-2, and MT5, to classify user emotions and generate emotionally aligned responses. EDRG demonstrated high accuracy in emotion classification using marBERT. Additionally, AraBERT, AraELECTRA, AraGPT-2, and MT5 showed strong structural consistency in their generated responses.

The second approach, EmoLlama, utilizes Retrieval-Augmented Generation (RAG) with LLaMA to retrieve semantically relevant conversational examples prior to generation. This retrieval mechanism significantly enhanced contextual coherence and adaptability, especially in multi-turn and ambiguous queries. EmoLlama also proved effective for domain-specific applications such as mental health, customer support, and education, due to its ability to draw on external knowledge sources.

Future work will focus on further improving the empathy and adaptability of Arabic conversational AI systems.

One important direction is the expansion into Arabic dialects to support more natural and colloquial interactions, which remains an underexplored area in Arabic NLP research.

Another promising avenue involves developing personality-aware chatbots. This includes integrating psycholinguistic models to detect user personality traits (such as extroversion or introversion) and adjusting responses accordingly. Such approaches provide a valuable foundation for building more personalized Arabic conversational systems.

A third direction is the exploration of hybrid architectures that combine EDRG's emotion-sensitive response generation with EmoLlama's retrieval-based contextualization. Such integration can leverage both emotional accuracy and context relevance.

Additionally, extending emotion recognition beyond text is vital. This includes integrating multimodal signals such as voice tone, pitch, and facial expressions through audio-text fusion techniques, which could significantly enhance the emotional intelligence of the chatbot.

Future development also requires enriching existing Arabic emotional datasets, particularly for underrepresented dialects and specialized domains such as mental health and education. Broader data coverage will help build more robust and generalizable models.

Finally, ensuring effective alignment between emotions and responses remains a critical challenge. Not all responses should mirror the user's exact emotional state; for example, expressions of happiness may be better complemented by encouragement or empathy rather than repetition of the same sentiment. This nuanced perspective on sentiment matching, discussed in the results section, highlights the importance of designing future response strategies that balance emotional appropriateness with conversational diversity.

In conclusion, this research demonstrated that emotion-aware and retrieval-augmented approaches can significantly improve the quality of Arabic conversational AI systems. The results showed that large language models originally developed for other languages can be effectively adapted to Arabic through fine-tuning and retrieval-augmented techniques. Among the evaluated models, marBERT achieved strong emotion classification performance, while AraBERT and AraELECTRA demonstrated superior capabilities in generating emotionally aligned responses. In addition, the EmoLlama approach provided substantial improvements in semantic relevance and contextual understanding through retrieval-augmented generation. These findings highlight the value of combining emotional awareness with contextual retrieval to develop more empathetic, coherent, and adaptable Arabic dialogue systems.

Author Contributions: Conceptualization, S.M.A. and N.A.A.; Methodology, S.M.A. and N.A.A.; Software, S.M.A.; Validation, S.M.A., N.A.A. and M.M.B.I.; Formal analysis, S.M.A., N.A.A. and M.M.B.I.; Investigation, S.M.A.; Resources, S.M.A.; Data curation, S.M.A.; Writing—original draft, S.M.A.; Writing—review & editing, S.M.A., N.A.A. and M.M.B.I.; Visualization, S.M.A.; Supervision, N.A.A. and M.M.B.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research and the APC were funded by King Saud University, Riyadh, Saudi Arabia, through the Ongoing Research Funding Program (ORF-2026-846).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors would like to thank King Saud University, Riyadh, Saudi Arabia, for supporting this work through the Ongoing Research Funding Program (ORF-2026-846).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Brandtzaeg, P.B.; Følstad, A. Why people use chatbots. In *Proceedings of the 4th International Conference, INSCI 2017, Thessaloniki, Greece, 22–24 November 2017*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017; pp. 377–392. [[CrossRef](#)]
2. Følstad, A.; Brandtzaeg, P.B. Chatbots and the new world of HCI. *Interactions* **2017**, *24*, 38–42.
3. Xu, A.; Liu, Z.; Guo, Y.; Sinha, V.; Akkiraju, R. A new chatbot for customer service on social media. In *CHI '17 Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*; Association for Computing Machinery: New York, NY, USA, 2017; pp. 3506–3510. [[CrossRef](#)]

4. Weizenbaum, J. ELIZA—A computer program for the study of natural language communication between man and machine. *Commun. ACM* **1966**, *9*, 36–45. [\[CrossRef\]](#)
5. Colby, K.M. Human-computer conversation in a cognitive therapy program. In *Machine Conversations*; Springer: Boston, MA, USA, 1999.
6. Ali, D.A.; Habash, N. Botta: An arabic dialect chatbot. In *Proceedings of the COLING 2016, System Demonstrations*; The COLING 2016 Organizing Committee: Osaka, Japan, 2016; pp. 208–212.
7. Driss, M.; Almomani, I.; Alahmadi, L.; Alhajjam, L.; Alharbi, R.; Alanazi, S. COVIBOT: A Smart Chatbot for Assistance and E-Awareness during COVID-19 Pandemic. In *Proceedings of the 2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*; IEEE: New York, NY, USA, 2022; pp. 124–129.
8. Al-Madi, N.A.; Maria, K.A.; Al-Madi, M.A.; Alia, M.A.; Maria, E.A. An Intelligent Arabic Chatbot System Proposed Framework. In *Proceedings of the 2021 International Conference on Information Technology (ICIT)*; IEEE: New York, NY, USA, 2021; pp. 592–597. [\[CrossRef\]](#)
9. Alhumoud, S.; Diab, A.; AlDukhai, D.; AlShalhoub, A.; AlAbdullatif, R.; AlQahtany, D.; AlAlyani, M.; Bin-Aqeel, F. Rahhal: A Tourist Arabic Chatbot. In *Proceedings of the 2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*; IEEE: New York, NY, USA, 2022; pp. 66–73.
10. AlHumoud, S.; Wazrah, A.A.; Aldamegh, W. Arabic chatbots: A survey. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **2018**, *9*, 535–541. [\[CrossRef\]](#)
11. Shawar, B.A. A chatbot as a natural web interface to arabic web QA. *iJET* **2011**, *6*, 37–43. [\[CrossRef\]](#)
12. Alsadhan, N.A. A Novel Dialect-Aware Framework for the Classification of Arabic Dialects and Emotions. *arXiv* **2025**, arXiv:2502.09128.
13. Ayanouz, S.; Abdelhakim, B.A.; Benhmed, M. A smart chatbot architecture based NLP and machine learning for health care assistance. In *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 1–6.
14. AlHagbani, E.S.; Khan, M.B. Challenges facing the development of the Arabic chatbot. In *Proceedings of the First International Workshop on Pattern Recognition*; SPIE: Bellingham, WA, USA, 2016; pp. 192–199.
15. Hammo, B.; Abu-Salem, H.; Lytinen, S.; Evens, M. QARAB: A Question answering system to support the Arabic language. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*; Association for Computational Linguistics: San Diego, CA, USA, 2002.
16. Naous, T.; Antoun, W.; Mahmoud, R.A.; Hajj, H. Empathetic BERT2BERT conversational model: Learning Arabic language generation with little data. *arXiv* **2021**, arXiv:2103.04353.
17. Naous, T.; Hokayem, C.; Hajj, H. Empathy-driven Arabic conversational chatbot. In *Proceedings of the Fifth Arabic NLP Workshop*; Association for Computational Linguistics: Barcelona, Spain, 2020; pp. 58–68.
18. Schank, R.C. Conceptual dependency: A theory of natural language understanding. *Cogn. Psychol.* **1972**, *3*, 552–631. [\[CrossRef\]](#)
19. Adiwardana, D.; Luong, M.T.; So, D.R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. Towards a Human-like Open-Domain Chatbot. *arXiv* **2020**, arXiv:2001.09977. [\[CrossRef\]](#)
20. Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Boureau, Y.L.; Weston, J. Recipes for Building an Open-Domain Chatbot. *arXiv* **2021**, arXiv:2004.13637. [\[CrossRef\]](#)
21. Zaranis, E.; Paraskevopoulos, G.; Katsamanis, A.; Potamianos, A. EmpBot: A T5-based Empathetic Chatbot focusing on Sentiments. *arXiv* **2021**, arXiv:2111.00310. [\[CrossRef\]](#)
22. Smith, A.; Brown, L. Enhancing Emotional Connections in Chatbots Using Transformer-Based Architectures. *J. AI Hum. Interact.* **2023**, *15*, 122–138.
23. Alsadhan, N.; Skillicorn, D. Estimating Personality from Social Media Posts. In *Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW)*; IEEE: New York, NY, USA, 2017; pp. 350–356. [\[CrossRef\]](#)
24. Boulesnane, A.; Saidi, Y.; Kamel, O.; Bouhamed, M.M.; Mennour, R. DZchatbot: A Medical Assistant Chatbot in the Algerian Arabic Dialect using Seq2Seq Model. In *Proceedings of the 2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*; IEEE: New York, NY, USA, 2022; pp. 1–8.
25. Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-based model for Arabic language understanding. *arXiv* **2020**, arXiv:2003.00104. [\[CrossRef\]](#)
26. Al-Ghadhban, D.; Al-Twairish, N. Nabiha: An Arabic dialect chatbot. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 452–459. [\[CrossRef\]](#)
27. Alruily, M. ArRASA: Channel Optimization for Deep Learning-Based Arabic NLU Chatbot Framework. *Electronics* **2022**, *11*, 3745. [\[CrossRef\]](#)
28. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805. [\[CrossRef\]](#)
29. Clark, K.; Luong, M.T.; Le, Q.V.; Manning, C.D. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv* **2020**, arXiv:2003.10555. [\[CrossRef\]](#)

30. Antoun, W.; Baly, F.; Hajj, H. AraELECTRA: Pre-training Text Discriminators for Arabic Language Understanding. *arXiv* **2020**, arXiv:2012.15516. [[CrossRef](#)]
31. Antoun, W.; Baly, F.; Hajj, H. AraGPT2: Pre-trained Transformer for Arabic Language Generation. *arXiv* **2020**, arXiv:2012.15520. [[CrossRef](#)]
32. Abdul-Mageed, M.; Elmadany, A.; Nagoudi, E.M.B. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. *arXiv* **2020**, arXiv:2101.01785. [[CrossRef](#)]
33. Huang, D.; Wang, Z. Explainable Sentiment Analysis with DeepSeek-R1: Performance, Efficiency, and Few-Shot Learning. *arXiv* **2025**, arXiv:2503.11655. [[CrossRef](#)]
34. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Yih, W.T.; Rocktäschel, T.; Riedel, S.; Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv* **2020**, arXiv:2005.11401. [[CrossRef](#)]
35. Jiang, Z.; Ma, X.; Chen, W. LongRAG: Enhancing Retrieval-Augmented Generation with Long-context LLMs. *arXiv* **2024**, arXiv:2406.15319. [[CrossRef](#)]
36. Ahmed, F.; Khan, R.; Lee, J. Dialogue-RAG: Retrieval-Augmented Generation for Multi-Turn Conversational Systems. In *Proceedings of the Conversational AI Symposium*; AAAI Press: Washington, DC, USA, 2023; pp. 91–108.
37. Hussein, M.; Zhang, L.; Patel, S. Adaptive RAG Pipelines for Domain-Specific Applications in Healthcare and Legal Assistance. *Domain-Specif. AI Res. Q.* **2024**, *9*, 47–65.
38. Ahmed, B.; Kim, Y.; Al-Shehri, H. Legal Assistance Chatbots with Retrieval-Augmented Generation. In *Proceedings of the Legal AI Innovations Symposium*; AAAI Press: Washington, DC, USA, 2023; pp. 57–71.
39. Group, J.A.D. JAIS 30B Chat: Enhancing Logical Reasoning in Arabic Retrieval-Augmented Generative Models. In *Proceedings of the AI in Multilingual NLP Conference*; AAAI Press: Washington, DC, USA, 2024; pp. 201–214.
40. Solutions, H. Arabic Copilot: A Retrieval-Augmented System for Real-Time Enterprise Applications. *ArabNLP J.* **2023**, *12*, 78–95.
41. Ali, N.; Ibrahim, R.; Al-Mutairi, S. MedRAG-Ar: RAG-Powered Healthcare Chatbots for Arabic Medical Queries. *Arab. Healthc. Inform.* **2023**, *11*, 102–120.
42. Jia, Y. Attention Mechanism in Machine Translation. *J. Phys. Conf. Ser.* **2019**, *1314*, 012186.
43. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
44. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Gallé, M.; Ram, A.; et al. LLaMA: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.