

Article

AEConvs: A Novel Dataset and Benchmark for Evaluating Empathetic Response Generation in Arabic LLMs

Afnan Alkhathlan ^{1,2,*}  and Abdulrahman A. Mirza ² 

¹ Engineering and Computer Science Department, Applied College, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11564, Saudi Arabia

² Information Systems Department, College of Computer and Information Sciences, King Saud University (KSU), Riyadh 11362, Saudi Arabia

* Correspondence: afaalkhathlan@imamu.edu.sa

Abstract

Empathy—the ability to understand and respond to others’ emotions and perspectives—is a key communication skill for humans; however, it is under-explored within current conversational systems. While large language models (LLMs) have demonstrated a remarkable capability to generate coherent and contextually relevant output, they often struggle to exhibit genuine empathy, resulting in artificial and dull responses, particularly in low-resource languages such as Arabic. Notably, the research on empathetic conversational systems in Arabic is still in its early stages, mainly due to the scarcity of open-domain conversational data. To address this gap, we introduce **Arabic Empathetic Conversations (AEConvs)**, a genuine Arabic conversational dataset featuring more than 4K open-domain dyadic empathetic conversations. This dataset provides a valuable resource that captures nuanced emotional and empathetic cues in the Arabic language. Using **AEConvs**, we evaluate and compare the empathetic capabilities of two state-of-the-art generative Arabic LLMs—AceGPT-chat and Jais-chat—under zero-shot and fine-tuning training settings. Human evaluation results demonstrate that while both models exhibit some form of empathy in zero-shot settings, fine-tuning on **AEConvs** improved their ability to generate more fine-grained empathetic responses while also yielding enhancements in fluency and context adherence. Additionally, automatic evaluation indicated improved language modeling and better lexical and semantic similarity with human reference responses. This study highlights the importance of culturally and linguistically tailored datasets in advancing empathetic conversational AI. We publicly release the **AEConvs** dataset, providing a valuable resource for future advancements in the field.

Keywords: Arabic; conversations; dataset; empathy; large language models; natural language processing; text generation



Academic Editor: Kesheng (John) Wu

Received: 29 January 2026

Revised: 4 April 2026

Accepted: 9 April 2026

Published: 14 April 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Empathy is an innate cognitive ability that enables humans to recognize and comprehend others’ emotional states through perspective-taking. It plays a vital role in social interactions by strengthening emotional bonds, fostering trust, and nurturing compassion. Furthermore, empathy is integral to everyday conversation, as people share experiences and emotions, reinforcing mutual understanding and connection [1]. Since the term empathy emerged in the late nineteenth century, it has been defined differently across the diverse disciplines of social psychology, ethics, cognitive science, and clinical psychology.

For example, Barker 1995 [2] defined empathy as “the act of perceiving, understanding, experiencing, and responding to the emotional state and ideas of another person.” In contrast, Hoffman [3] defined empathy as “an affective response more appropriate to another’s situation than one’s own.” In addition, the concept of empathy encompasses both affective and cognitive dimensions [1]. Affective empathy is the ability to feel or mirror—either implicitly or explicitly—another person’s feelings. Cognitive empathy, also called perspective-taking, entails a deliberate effort to understand and interpret another person’s situation and implicit feelings without necessarily sharing them. Hence, empathy requires recognizing and understanding others’ emotional states as well as the ability to respond to them with sympathy and compassion.

A key challenge in developing human-like conversational models is enabling empathy. This involves equipping a model with the ability to infer and respond to a user’s emotions, thereby enabling it to generate more genuine and engaging responses. This response generation process requires modeling cognitive and affective states, making it a complex generative challenge beyond mere emotion or sentiment detection. Empathetic responses have the potential to revolutionize various fields, including education, customer service, entertainment, and healthcare. Although English language research has increased in this domain, there remains a gap in the nuanced application of empathy across languages and cultures, where cultural backgrounds significantly shape how people express emotions and seek empathy [4,5]. These socio-cultural variations are often overlooked in current open-domain conversational systems, potentially limiting their effectiveness and scalability. A significant research gap persists in the study of empathy for Arabic conversational AI. The scarcity of human-generated conversation datasets that accurately capture their nuanced communication strategies and empathetic behavior poses a significant barrier to developing truly emotionally intelligent and culturally sensitive conversational models.

Despite its intensive history, the field of conversational AI has recently witnessed remarkable advancements driven by the emergence of transformer-based large language models (LLMs). These language models have shown extraordinary capabilities in a variety of language understanding and generation tasks. However, most of these models are mainly tailored for the English language. Although some models exhibit multilingual capabilities, they still struggle to comprehend the morphological and syntactic characteristics of the Arabic language in particular [6–8]. In recent years, the field of Arabic language processing has seen increased development of generative Arabic LLMs, such as AraGPT2 [9] and AraT5 [10]. However, these models are not tailored for following instructions or maintaining coherent conversations. Only recently has the research community witnessed the emergence of powerful chat-oriented and instruction-following Arabic generative LLMs such as Jais-chat [11], AceGPT-chat [12], and ALLaM [13], offering exciting prospects to advance Arabic applications. Nevertheless, as far as we know, these models remain untested with respect to their emotional awareness or empathetic capabilities.

In view of these limitations, we present the **Arabic Empathetic Conversations (AEConvs)** —a dataset containing more than 4K open-domain Arabic textual empathetic conversations. Each conversation in the dataset is conducted between two humans, resulting in authentic and natural conversations. In addition, we leverage **AEConvs** to evaluate the empathetic capabilities of two recently released Arabic generative LLMs—AceGPT-chat and Jais-chat—using two learning approaches: zero-shot and fine-tuning. In contrast to fine-tuning, zero-shot testing evaluates a model’s ability to perform a new task without explicit training on task-specific examples. This approach leverages the model’s inherent knowledge and capacity to generalize to unseen situations. Despite the relatively small dataset, automatic and human evaluation results demonstrate that fine-tuning with **AEConvs** yields significantly more fine-grained empathetic responses in both models com-

pared to zero-shot responses, while also enhancing the ability to generate more fluent and contextually relevant responses. Furthermore, human evaluators found that the fine-tuned AceGPT-chat responses were more empathetic than Jais-chat in over half of the cases, suggesting a potential performance advantage for AceGPT-chat for this task.

The main contributions of this study are as follows:

1. We construct a novel Arabic empathetic conversation dataset as a new benchmark. To the best of our knowledge, **AEConvs** is the first of its kind to contain genuine empathetic open-domain conversations written in Modern Standard Arabic (MSA), the formal and standardized form of language used across the Arabic-speaking world. The **AEConvs** dataset is designed to advance research on open-domain conversations, particularly the development of empathetic conversational models. We believe that the release of **AEConvs** is a step toward making Arabic LLMs more emotionally sensitive and culturally aligned.
2. We provide the first empirical investigation of the empathetic response capabilities of two state-of-the-art Arabic generative LLMs—AceGPT-chat and Jais-chat—under two training settings: zero-shot and fine-tuning. Compared to zero-shot baselines, human evaluation results demonstrate that fine-tuning on **AEConvs** improved the performance of both models in generating empathetic responses, while also improving their fluency and context-following capabilities. Additionally, the automatic evaluation indicates improved language modeling and better lexical and semantic similarity with human reference responses in both models.

The remainder of this study is organized as follows: Section 2 provides an overview of generative large language models (LLMs) and the recent literature on empathetic response generation in English and Arabic. Section 3 describes the methodology used to curate the **AEConvs** dataset and its relevant statistics. In addition, it details the experimental setup, including the base models, fine-tuning details, and evaluation metrics. Section 4 presents and discusses the experimental results and provides a case study with error analysis, offering a qualitative in-depth examination of model outputs. Finally, Section 5 provides concluding remarks and future directions.

2. Related Work

This section overviews the current landscape of generative LLMs and explores progress in empathetic response generation across the English and Arabic languages.

2.1. Large Language Models (LLMs)

Since the introduction of the transformer architecture in 2017 [14], pre-trained LLMs have fundamentally transformed the field of natural language processing (NLP). These large models—trained on massive corpora, including published research, web content, and social media data—demonstrate remarkable proficiency across a wide range of language comprehension and generation tasks, including sentiment analysis, text translation, and the generation of coherent conversations. Furthermore, these models have shown impressive results in zero-shot and few-shot learning, eliminating the need for extensive task-specific training data [15,16]. More recently, these LLMs have been fine-tuned on datasets with explicit task instructions, a process known as instruction tuning, and further aligned with reinforcement learning from human feedback (RLHF) in order to generate more helpful and human-like outputs [17,18]. For example, the GPT series [19] and Google’s Gemini [20] excel in producing human-like language with impressive fluency. However, most of these models are mainly tailored for the English language. Although some models exhibit multilingual capabilities, they still struggle to comprehend the morphological, syntactic, and cultural characteristics of the Arabic language in particular [6–8].

The development of generative Arabic LLMs has witnessed notable progress in recent years; however, these models still trail behind their English counterparts. This gap stems from Arabic's linguistic complexities and dialectal variations, as well as the scarcity of high-quality training corpora. Moreover, progress in generative Arabic LLMs has lagged behind advancements in language-understanding models, including BERT-based models such as AraBERT [21] and QARiB [22]. Among the few existing generative Arabic LLMs are AraGPT2 [9], AraT5 [10], and, more recently, ArabianGPT [23]. However, these models are not instruction-tuned or trained for conversational objectives. Only recently have we witnessed the emergence of Arabic LLMs specifically tailored for instruction following and applications. Among the first of such models is Jais-chat, which is an instruction-tuned version of the Arabic–English bilingual LLM Jais [11]. Another example is AceGPT-chat [12], which is based on the LLaMA-2 [24] architecture and trained to be culturally attuned to Arabic norms and values. More recently, the Saudi Data and Artificial Intelligence Authority (SDAIA) released ALLaM [13], an autoregressive decoder-only bilingual model based on the LLaMA-2 [24] architecture and trained on a mixture of Arabic and English data.

Despite their remarkable proficiency, state-of-the-art multilingual and bilingual LLMs, including Jais [11], GPT-4 [25], and BLOOM [26], struggle to grasp the cultural nuances and diversity across different regions [27–29], particularly in Arabic, where they are often biased towards Western entities and stereotypes [30,31]. This bias and misalignment mainly stem from their training data [32,33]. The predominance of English content in the training corpora creates a risk that these models will perpetuate Western cultural narratives while marginalizing non-Western ones [29]. Recent work has begun to address this challenge through culturally rich training data, prompt engineering that incorporates cultural context and evaluation frameworks that assess model performance across cultural dimensions [27,28,30].

2.2. Empathetic Response Generation in English

There has been a great interest in endowing conversational models with empathetic capabilities to improve the quality of the generated responses and make them more human-like. Early approaches relied on specifying emotion labels and using them to guide the response generation process during the decoding phase of Seq2Seq neural models [34,35]. Another line of research has leveraged explicit contextual information to generate more empathetic responses, including emojis [36], sentiment [37], persona [38], and acts [39]. While recent transformer-based generative LLMs have demonstrated advanced learning and generalization skills in conversational modeling, their potential for empathetic responses remains largely underexplored [15,40]. Current GPT-based chatbots have been shown to exhibit some empathy traits in their generated responses, but they are often perceived as less empathetic compared to human interactions [41,42]. More specifically, ChatGPT [43] generated repetitive patterns of empathy and underperformed compared to supervised models in terms of emotional understanding and empathetic responding [44,45]. Hence, these large generative models require further training and adjustments to generate more fine-grained, human-like, empathetic responses.

A notable challenge is that training LLMs for empathetic responses requires substantial amounts of conversational data, which is often difficult to obtain. Introduced in 2019 [46], EmpatheticDialogues has been one of the most notable datasets in the field. The dataset contains approximately 25K crowd-sourced dyadic dialogues covering 32 evenly distributed emotions. This dataset has served as a key benchmark for training and evaluating conversational models to generate emotional and empathetic responses in a large number of studies [37,41,45,47]. Furthermore, movie and TV show scripts have been a valuable source of conversational data since they contain large amounts of fictional s in

different domains and languages [48,49], in addition to conversations collected from social media platforms such as Twitter [50] and Reddit [51]. However, training conversational models on web and social media data risks exposing them to toxic, biased, or sarcastic language, potentially resulting in harmful or unreliable outputs. Moreover, the inherent noise, spam, and low quality of such data may further hinder the model's performance.

2.3. Empathetic Response Generation in Arabic

To date, research on empathetic Arabic conversational systems is in its early infancy. The first attempt to develop an empathetic Arabic chatbot model was in 2020 [52], where the researchers used a Seq2Seq model with an LSTM network and attention. The model was trained on a machine-translated Arabic version of the EmpatheticDialogues dataset [46]. Despite the complexity of Arabic and the relatively small dataset, the model achieved promising results in terms of empathy and fluency. However, the model exhibited an average relevance score, producing irrelevant responses in some cases. To address such limitations, in 2021, the authors [47] further proposed a transformer-based encoder–decoder model that was initialized with AraBERT [21] pre-trained weights and trained on the same translated dataset. The model generated more empathetic, fluent, and relevant responses. However, the model performed poorly in handling neutral emotions because it was fine-tuned on empathetic conversations rather than pre-trained on general chit-chat data.

Since Arabic is a low-resourced language, translating datasets from other languages is a common practice in the literature. The Arabic translation of the EmpatheticDialogues dataset [46] constitutes the first and only up-to-date Arabic empathetic conversational dataset. Although translated datasets have yielded promising results, machine translation remains prone to low-quality output, often due to the use of slang and idiomatic expressions commonly found in open-domain conversations. For example, the sentence “This scared the hell out of me” in the original EmpatheticDialogues dataset [46] is machine-translated to Arabic as *هذا أخاف مني الجحيم*: “This makes the hell scared of me.” This literal translation is inaccurate and may imply a different emotion than intended. More importantly, translated datasets fail to align with the cultural heritage and local values of the Arabic communities, and they may reflect cultural biases and ethical concerns from the source language [28]. Training conversational agents on Western or translated datasets can lead to cultural encapsulation [53], where models treat Western emotional expressions as universal while failing to recognize culturally specific ways of communicating distress or seeking support. This bias not only compromises model performance across cultural contexts but also threatens to marginalize authentic human expression from non-Western communities [29]. This is particularly important in Arabic, as the language spans different regions with distinct values and social norms that differ profoundly from the cultures and communities of the Western world [12]. The lack of extensive, culturally and religiously nuanced training data hinders the development of robust and culture-aware conversational models in Arabic, limiting their adequacy for culturally sensitive domains, such as mental health support and spiritual guidance.

3. Materials and Methods

This section details the experimental setup of this study. First, we introduce the **AEConvs** dataset, outlining its construction process and key characteristics. We then describe the base pre-trained language models—AceGPT-chat and Jais-chat—and their fine-tuning details. Finally, we present a comprehensive evaluation framework that leverages both automatic metrics and human assessment.

3.1. AEConvs Dataset

Dataset Collection: In order to construct the Arabic Empathetic Conversations (**AEConvs**) dataset, a total of 85 freelancer content writers were hired from [Khamsat.com](https://www.khamsat.com), an Arabic freelancing marketplace. The freelancers were all native Arabic speakers with experience writing in Modern Standard Arabic (MSA). Table 1 shows the demographic data of the freelancers, including gender, age, and education. Data were collected in batches. For each batch, a group of writers (maximum 10) was added to a group chat on the *Telegram* application and asked to engage in dyadic open-domain conversations. Each conversation consisted of one speaker and one listener, in which the speaker initiated a conversation about a daily event or an emotional incident, and the listener responded with empathy. The speaker and listener can alternate turns in each conversation, with a minimum of 4 and a maximum of 10 turns. After a pair of writers finishes their conversation, another pair starts a new one, and so on.

Table 1. Demographic distribution of the freelance writers.

Gender:	Male 28%	Female 72%		
Age:	18–24 46%	25–34 42%	35–44 8%	+45 4%
Education:	High school 5%	Bachelor 72%	Graduate 14%	Other 9%

To assess the qualifications of each freelance writer for the task, particularly their writing ability and understanding of the assignment, they were first asked to write a sample set of empathetic conversations. Upon acceptance, the writers were explicitly instructed not to use dialectal language, emojis, symbols, or numbers. They were also required to use proper language and avoid greetings and short answers (e.g., “Yes,” “No,” “I agree,” etc.). In addition, the writers were advised to refrain from revealing personal information or discussing controversial topics such as politics, sex, and religion. Proofreading and quality checks were manually performed after each batch.

Dataset Statistics: A total of 4155 conversations were collected. After filtering out duplicates, conversations with fewer than 4 utterances, violations in speaker–listener turn-taking, and listeners’ turns with fewer than 3 words, we obtained 4120 unique conversations divided equally between listener and speaker utterances. Table 2 shows the detailed statistics of the **AEConvs** dataset. All conversations are written in MSA. Most of the dataset consists of short conversations with 4–6 turns, with less than 1% containing more than 8 utterances. A sample conversation from **AEConvs** is shown in Table 3. In addition, Figure 1 shows the word cloud of the 50 most frequently occurring words in the listeners’ turns across the 4120 conversations, indicating the level of expressed empathy in **AEConvs**. Stop words were excluded using a custom Arabic stop word list, and the ISRIStemmer from the nltk library was applied to normalize word forms.

Data Pre-processing: We used CAMEL tools [54] for Arabic data pre-processing and performed the following steps:

1. Dediactritization: Arabic diacritical marks (ḥarakāt) are stripped from text to obtain a more consistent and standardized orthographic representation, e.g., مَبَارَك to مَبَارِك.
2. Removal of repeating characters: Duplicated letters are removed from words, as they are often used to convey tone, e.g., مَبْرُوووك to مَبْرُوَك.
3. Kashīda removal: Elongation characters used in Arabic text to artificially extend words are removed, e.g., شَكَرَا to شَكَرَا.

4. Text normalization: Punctuation, numerals, and special characters are stripped, and white spaces are standardized.

In addition, the dataset was randomly partitioned into 80% for training, 10% for validation, and 10% for testing. The splitting was performed at the conversation level, rather than the turn level, to maintain integrity.

Table 2. Statistics of the AEConvs dataset.

No. of conversations	4120
No. of speaker utterances	9869
No. of listener utterances	9869
Average no. of turns per conversation	4.8
Average no. of words per conversation	70
Average no. of words per utterance	14.4

Table 3. A conversation sample (with English translation) from the AEConvs dataset.

لشد ما أتمنى أن يكون لدي إخوة! I really wish I had siblings!	المتحدث Speaker
وجود الإخوة والأخوات هو نعمة عظيمة فعلا، هل أنت الطفل الوحيد في عائلتكم؟ Having brothers and sisters is truly a great blessing. Are you the only child in your family?	المستمع Listener
نعم للأسف ليس لدي إخوة ولا أخوات، أنت لا تتخيل مشاعر الوحدة التي عشتها منذ طفولتي Yes, unfortunately I have no brothers or sisters. You can't imagine the loneliness I've felt since childhood.	المتحدث Speaker
اتفهم شعورك، وجود الاخوة لا يقدر بثمن، فهم يشاركوننا الأفراح والأحراح ومعهم نصنع الذكريات السعيدة، لكن الله يرزقنا أحيانا بأصدقاء أو أقارب يكونون لنا أفضل وأقرب من الإخوة I understand your feelings; having siblings is priceless. They share our joys and sorrows, and we create happy memories with them. But sometimes God blesses us with friends or relatives who are better and closer to us than our own siblings.	المستمع Listener
صحيح، اشعر أن الله عوضني بصديقي أحمد فهو صديقي المقرب منذ الطفولة، عمر صداقتنا الآن ناهز العشرون سنة! That's right, I feel that God has compensated me with my friend Ahmed, who has been my close friend since childhood. Our friendship is now almost twenty years old!	المتحدث Speaker
هذا رائع ما شاء الله! أتمنى لكما دوام هذه الصداقة الرائعة This is wonderful, God bless you! I hope this wonderful friendship lasts forever.	المستمع Listener

3.2. Base Models

AceGPT-chat: AceGPT models [12] are a recently released series of LLaMA2-based [24] models with different parameter sizes, ranging from 7B to 13B. AceGPT was designed to serve as an Arabic culture-aware and value-aligned generative model. The integration of localized pre-training, fine-tuning on native Arabic instructions, and reinforcement learning from AI feedback (RLAIF) enables AceGPT models to better align with Arabic cultural norms and values. This leads to improved performance on culturally-specific tasks. AceGPT models are pre-trained on a mixture of English and Arabic data, using a larger subset of the latter. The AceGPT family consists of two variants: AceGPT-base (the foundational model) and AceGPT-chat (an instruction-tuned model optimized for

To ensure a fair comparison, we used the AceGPT-7B-chat and Jais-adapted-7B-chat versions; both are LLaMA-2-based, with approximately 7 billion parameters. Table 5 shows the main features of AceGPT-7B-chat and Jais-adapted-7B-chat. For ease of reference, we refer to these models as *AceGPT-chat* and *Jais-chat*, respectively, throughout the rest of this paper.

Table 5. A comparison of the main features of the AceGPT-7B-chat and Jais-adapted-7B-chat models.

Model	Size (Parameters)	Base Architecture	Context Length	Pre-Training Data Size	Instruction-Tuning Data Size
AceGPT-7B-chat	7 billion	LLaMA-2	2048	30B, 19.2B of which are Arabic	194 K prompt–response pairs; mix of English, native Arabic, translated, and GPT-4-generated
Jais-adapted-7B-chat	7 billion	LLaMA-2	4096	39B, 19B of which are Arabic	10 M English and 4 M Arabic prompt–response pairs

3.3. Fine-Tuning Details

We fine-tuned both models using Low-Rank Adaptation (LoRA) [56] within the Unsloth¹ framework. LoRA enhances LLM fine-tuning by decomposing the weight update matrix into a lower-rank representation, enabling faster training and potentially lower computational cost without sacrificing performance. The LoRA configuration employed a rank of $r = 32$ with a scaling factor $\alpha = 32$ and a dropout rate of 0.1 for regularization. For model stability, we enabled rank stabilization (rsLoRA) [57], and gradient checkpointing was enabled through Unsloth’s optimized implementation to reduce memory consumption. A random seed of 3407 was used to ensure reproducibility. For the learning rate, we experimented with different values ranging from 1×10^{-5} to 5×10^{-5} and selected the one that achieved the best perplexity (PPL) on the validation set without overfitting, as shown in Table 6. We followed existing work [58,59] using perplexity as a metric to find the optimal learning rate and number of training epochs. Both models were trained with early stopping, and training for more than 2 epochs resulted in overfitting and higher perplexity, indicating worse performance. Furthermore, we set the batch size to 16, the maximum sequence length to 512, and used the AdamW (Adam with Weight Decay) optimizer for both models. Conversations were formatted according to each model’s standardized chat template prior to fine-tuning. All experiments were conducted on a single Tesla T4 GPU provided by Google Colab. The training time lasted 5 h and 31 min for AceGPT-chat and 3 h and 2 min for Jais-chat.

Table 6. Hyperparameters and perplexity values used in fine-tuning AceGPT-chat and Jais-chat models.

Model	Learning Rate	Training Epochs	Perplexity (PPL)
AceGPT-chat	1×10^{-5}	2	2.03
Jais-chat	3×10^{-5}	2	3.81

3.4. Evaluation

In this work, we employ a comprehensive evaluation strategy that combines automated and human assessment. Automatic metrics capture different dimensions of text quality, while human evaluation provides holistic validation. The models’ performance was evaluated on a random sample of 100 conversations from the held-out test set. The assessment was restricted to a single-turn response generation task, providing a critical

and interpretable baseline for meaningful future comparisons before progressing to the added complexity of multi-turn interaction.

3.4.1. Automatic Evaluation

Automatic evaluation typically compares the model-generated response against a golden reference response using predefined metrics to assess various attributes of text quality. We employed a suite of automatic evaluation metrics on the test set, each targeting a distinct aspect of text quality: perplexity (PPL) for intrinsic fluency, BLEU [60] for lexical overlap with reference data, and BERTScore [61] for semantic similarity via contextual embeddings.

- **Perplexity (PPL)** is an intrinsic metric used to evaluate the performance of language models and is commonly used to evaluate the fluency of generative LLMs. For each word in the sequence, perplexity measures the probability of that word occurring given the previous words. In other words, perplexity measures how well the model is confident in its generated responses, where lower perplexity scores indicate better language modeling performance and higher perplexity scores indicate model confusion. Perplexity is calculated as the exponentiated average negative log-likelihood of a sequence:

$$\text{Perplexity}(P) = \exp\left(-\frac{1}{N} \sum_{i=1}^N \log P(w_i)\right) \quad (1)$$

where N is the sequence length, and $P(w_i)$ is the model's predicted probability of the i -th token.

- **BLEU** [60] (Bi-Lingual Evaluation Understudy) is an n -gram-based metric originally designed for evaluating the quality of machine-translated text against one or more reference human translations. Despite capturing only surface-level lexical similarity, BLEU has been widely used in evaluating empathetic response generation [37,39,45,46], where improvements in the BLEU score suggest better alignment with the characteristic linguistic patterns of empathetic expressions in the training data. BLEU is computed as the geometric mean of the modified precision, weighted by a uniform weighting factor and multiplied by a brevity penalty (BP) to penalize overly short outputs:

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^N w_n \log p_n\right) \quad (2)$$

where p_n is the modified n -gram precision, w_n denotes the weights for each n -gram, and BP is the brevity penalty.

- **BERTScore** [61] is an automatic evaluation metric for text generation tasks. It utilizes contextual embeddings from a pre-trained BERT model to compute the similarity score between a generated text and a reference text, capturing nuanced semantic relationships beyond surface-level n -gram matching. BERTScore has demonstrated better correlation with human judgment in text generation tasks compared to common automatic metrics, including BLEU and ROUGE [61,62]. The BERTScore is represented by its F1 value derived from precision (P) and recall (R) according to the following formula:

$$F1 = 2 \times \left(\frac{P \times R}{P + R}\right) \quad (3)$$

where *precision* (P) is the true positives divided by the sum of true positives and false positives, and *recall* (R) is the true positives divided by the sum of true positives and false negatives.

3.4.2. Human Evaluation

To qualitatively assess model performance, we conducted two human evaluation tests: an **aspect-based rating test** and a **pairwise preference test**. We recruited 3 native Arabic speakers (1 male and 2 females) with a Bachelor's degree and content writing expertise. Prior to evaluation, a screening test was conducted to ensure that evaluators shared a consistent understanding of the evaluation criteria. They were asked to rate responses without knowing which model or setting (zero-shot vs. fine-tuned) produced each output. We used Fleiss' Kappa [63] as a measure of inter-annotator agreement between the evaluators.

In the **aspect-based rating test**, we generated responses from each model for the contexts of the sampled conversations—in zero-shot and fine-tuning settings—and then asked the evaluators to judge each generated response based on three criteria:

- **Empathy:** To what extent does this response show emotional understanding and compassion for the speaker's experience?
- **Relevance:** To what extent is this response relevant to the context of the conversation?
- **Fluency:** To what extent is this response readable and fluent?

For each criterion, a response was rated on a 3-point Likert [64] scale. Following [65], we opted for a Likert scale with three dimensions instead of five. This is because five-point ratings are highly likely to vary between different individuals, resulting in low inter-annotator agreement. In the **pairwise preference test**, each context of the sampled conversations was paired with two responses generated by each of the fine-tuned models, and the evaluators were asked to select the best response in terms of *empathy*. Ties were allowed if two responses were perceived as equally empathetic.

4. Results and Discussion

In this section, the results of both the automatic and human evaluations are presented and analyzed, comparing the performance of AceGPT-chat and Jais-chat in generating empathetic responses under zero-shot and fine-tuning settings. In addition, we present a case study with an error analysis, offering a qualitative in-depth examination of model outputs and showcasing their strengths and limitations in practical scenarios.

4.1. Automatic Evaluation

After fine-tuning, the perplexity, BLEU, and BERTScore were calculated on the test set. The results are illustrated in Table 7. Clearly, AceGPT-chat showed competitive performance compared to Jais-chat. Its low perplexity score indicates better fluency in its generated responses. Additionally, the slightly higher BLEU score indicates a stronger match with the human reference responses. Similarly, the BERTScore demonstrates that AceGPT-chat has slightly better semantic similarity with human responses than Jais-chat.

Table 7. Automatic evaluation results.

Model	Perplexity (PPL)	BLEU (Avg.)	BERTScore (F1)
AceGPT-chat	1.91	0.014	0.70
Jais-chat	3.14	0.011	0.67

Note: Under zero-shot settings, we note that both models often produce English or empty responses, despite being explicitly prompted to answer in Arabic. This could be attributed to their bilingual training. Specifically, 19% of AceGPT-chat responses contained mixed Arabic–English text. For Jais-chat, 8% of the responses were in English, and fewer

than 2% were empty responses. Thus, automatic metrics could not be computed as these require Arabic reference texts for comparison.

4.2. Human Evaluation

Tables 8 and 9 show the average human ratings in the **aspect-based rating test** for AceGPT-chat and Jais-chat, respectively. In zero-shot testing, both models showed comparable levels of empathy, with AceGPT-chat slightly ahead. Similarly, AceGPT-chat demonstrated better context relevance, while Jais-chat was slightly better in fluency. However, we observed that both models exhibited repetitive patterns of empathy, often generating generic responses such as “I am sorry to hear that,” which we showcased in Section 4.3. After fine-tuning, both models clearly showed a significant improvement in empathy, with Jais-chat performing slightly better. Likewise, fine-tuning led to improved context relevance for both models, with Jais-chat improving the most across all aspects. Lastly, fluency in both models improved the least after fine-tuning, indicating that both models were inherently able to generate fluent and coherent language.

Table 8. Results of aspect-based rating test for AceGPT-chat. Fleiss’ Kappa k is 0.36, where $20 < k < 40$ indicates fair agreement. All results are statistically significant ($p < 0.05$; Friedman test).

	Empathy	Relevance	Fluency
Zero-shot	1.79	2.14	2.39
Fine-tuned	2.56	2.82	2.85

Table 9. Results of aspect-based rating test for Jais-chat. Fleiss’ Kappa k is 0.42, where $40 < k < 60$ indicates moderate agreement. All results are statistically significant ($p < 0.05$; Friedman test).

	Empathy	Relevance	Fluency
Zero-shot	1.68	1.85	2.45
Fine-tuned	2.60	2.85	2.91

For the **pairwise preference test**, the results in Table 10 show that the responses of AceGPT-chat after fine-tuning, in terms of empathy, were preferred by a large margin over those generated by the Jais-chat model. However, the relatively high tie rate suggests that the Jais-chat model is still competitive in many cases.

Table 10. Results of the pairwise preference test. k denotes inter-annotator agreement measured using Fleiss’ Kappa, where $40 < k < 60$ indicates moderate agreement.

AceGPT-chat vs. Jais-chat	Win	Lose	Tie	k
	53%	14%	33%	0.42

In general, the results of both the automatic and human evaluation tests highlight both models as robust baselines for Arabic conversational AI. The AceGPT-chat and Jais-chat models were highly competitive in their ability to generate more empathetic, contextually relevant, and fluent responses after fine-tuning with the **AEConvs** dataset compared to their performance in zero-shot settings. This indicates the benefit of fine-tuning LLMs with a good-quality, task-specific dataset, even when the dataset is relatively small. In addition, the performance advantage of AceGPT-chat over Jais-chat can be attributed to its targeted training to localization and alignment with Arabic cultural nuances and values, as evidenced by its higher score on the Arabic Cultural and Value Alignment (ACVA) benchmark [12] (see Table 4). This underscores the critical role of culturally tailored datasets in enhancing LLMs’ performance for region-specific applications.

4.3. Case Study

Table 11 shows selected conversation contexts from the **AEConvs** dataset, providing the gold-standard human responses compared to the responses generated by the AceGPT-chat and Jais-chat models under both zero-shot and fine-tuning settings. Clearly, fine-tuning both models with **AEConvs** consistently improved the empathy of the generated responses across all cases. As demonstrated by the first case, the AceGPT-chat model generated a more convincing, empathetic response after fine-tuning than in zero-shot settings, exhibiting a spiritual perspective and offering comfort by suggesting that everything happens at the right time according to God's plan. The Jais-chat model, on the other hand, generated similar responses before and after fine-tuning, and these responses resemble generic empathetic expressions such as "I am sorry to hear that," as well as suggesting a medical consultation.

In the second case, both the AceGPT-chat and Jais-chat models under zero-shot settings showed brief responses that only acknowledged the speaker's feelings without offering any additional insights or encouragement. After fine-tuning, both models generated longer, more engaging responses, with AceGPT-chat closely mimicking the optimistic tone of a gold-standard human response. Similarly, Jais-chat's response was more empathetic and constructive, acknowledging the challenges of moving to a new house while framing the change as an opportunity for growth and learning. It provides a balanced perspective that combines understanding with encouragement.

Lastly, the third case shows brief and generic responses from both models under zero-shot settings that lack depth and fail to meaningfully engage with the speaker. In contrast, the responses generated by both models after fine-tuning are more empathetic and engaging. The AceGPT-chat model again mirrors the gold-standard human response by emphasizing the importance of gratitude in Islamic values and encouraging such acts. Jais-chat's response, conversely, offers acknowledgment and asks for further engagement with the speaker, promoting a deeper and more interactive conversation.

Table 11. Case study of the responses generated by AceGPT-chat and Jais-chat, where FT and Z-shot denote fine-tuning and zero-shot test settings, respectively. Responses were translated into English using Google Translate API.

Context (1)	مضى على زواجي خمسة أعوام، ولم أنجب بعد! I have been married for five years, and I still have no children!
Gold-standard human response	أفهم اشتياقك لكي تصبحين أما، حصل معي ذلك قبل I understand your longing to become a mother, it happened to me before.
AceGPT-chat (Z-shot)	لا تقلقي، هناك العديد من الأسباب التي قد تكون وراء عدم الإنجاب حتى الآن Don't worry, there are many reasons why you may not be able to conceive yet.
AceGPT-chat (FT)	أفهم قلقك ولكن لا تنسى أن الله يعطي ويمنح بحكمته، وأن كل شيء يحدث في الوقت المناسب I understand your concern, but do not forget that God gives and grants in His wisdom, and that everything happens in the right time.
Jais-chat (Z-shot)	أنا آسف لسماع ذلك. هل تحدثت مع طبيبك حول هذا؟ I'm sorry to hear that. Have you talked to your doctor about this?
Jais-chat (FT)	هذا أمر صعب حقاً، هل راجعت الطبيب؟ This is really hard, have you seen a doctor?

Table 11. Cont.

Context (2)	اليوم سأنتقل إلى المنزل الجديد، أنا حقا قلقة من إن كنت سأستطيع التأقلم على هذه البيئة الجديدة أم لا Today I will move to a new house. I am really worried whether I will be able to adapt to this new environment or not.
Gold-standard human response	إن هذا حقا خبر جيد وأتمنى أن تتأقلم بشكل جميل مع البيئة الجديدة التي حولك This is really good news and I hope you adapt nicely to your new environment.
AceGPT-chat (Z-shot)	أنا أفهم مشاعرك تجاه هذا التغيير. I understand your feelings about this change.
AceGPT-chat (FT)	لا تقلقي، ستأقلمين بسرعة على البيئة الجديدة وستحبينها بالتأكيد Don't worry, you will quickly adapt to the new environment and you will definitely love it.
Jais-chat (Z-shot)	أنا أفهم قلقك. I understand your concern.
Jais-chat (FT)	هذا أمر طبيعي، الانتقال إلى منزل جديد يمكن أن يكون تحديا. ولكن تذكر أن التغيير يمكن أن يكون فرصة للنمو والتعلم This is normal, moving to a new home can be challenging. But remember, change can be an opportunity for growth and learning.
Context (3)	أشعر بالامتنان لله على نعمه وعلى الأشياء الجميلة في حياتي I feel grateful to God for his blessings and the beautiful things in my life.
Gold-standard human response	الحمد والامتنان هما جزء من القيم الإسلامية الأساسية ويجب علينا أن نمارسها بشكل دائم. فبالشكر تزود النعم Gratitude and thankfulness are fundamental Islamic values and we should practice them constantly. Thankfulness increases blessings.
AceGPT-chat (Z-shot)	نعم، هذا شعور جميل وممتن. Yes, that's a beautiful and grateful feeling.
AceGPT-chat (FT)	إن الشكر على النعم هو من أهم العبادات في الإسلام، وهو يزيد من رضا الله عنا، ويجلب البركة في حياتنا. أنا سعيد لسماع أنك تشعر بالامتنان لله Being thankful for blessings is one of the most important acts of worship in Islam, it increases Allah's pleasure about us, and brings blessings in our lives. I am happy to hear that you feel grateful to Allah.
Jais-chat (Z-shot)	أنا سعيد لسماع ذلك I'm glad you feel that way!
Jais-chat (FT)	الحمد لله على نعمه، هل هناك شيء معين تشعرين بالامتنان له؟ Thanks to God for his blessings. Is there anything in particular that you are grateful for?

4.4. Error Analysis

Table 12 shows a few cases of erroneous model responses where the fine-tuned models fail to generate empathetic, coherent, and context-relevant responses. Jais-chat's response in the first case, despite appearing empathetic on the surface, offers a logically impossible advice—getting sleep before going to bed is temporally incoherent (sleep occurs after going to bed). This pattern, also observed in AceGPT-chat, suggests that the model has learned to recognize emotional cues but struggles to generate factual and contextually specific responses. The second case demonstrates that the AceGPT-chat model may be hallucinating conversation history, a manifestation of the broader hallucination problem in large language models. Although very rare, this hallucination error was observed in Jais-chat's responses as well. In the third case, Ace-GPT-chat gave a detailed medical diagnosis beyond its competence. This can be attributed to models optimized for helpfulness that may interpret symptom descriptions as requests for explanation, leading to diagnostic responses

even when clinically inappropriate. Jais-chat exhibited similar behavior in a few responses. Surprisingly, the last case shows similar responses from both models, as they claimed to be able to fix the speaker's broken phone. These responses, despite being empathetic, commit a credibility fallacy, where a conversational agent falsely claims personal experience of a phenomenon it cannot physically or cognitively experience [5]. This type of error could diminish the perceived authenticity and empathy of conversational models.

Table 12. Sample error cases by model.

Context (1)	البارحة لم يغمض لي جفن طوال الليل. I didn't get any sleep last night.
Jais-chat (FT)	هذا أمر مزعج حقا، هل جربت أن تنام قليلا قبل أن تذهب إلى الفراش؟ This is really annoying. Have you ever tried getting a little sleep before going to bed?
Context (2)	اليوم حقا كان مليئا بالفوضى Today was really chaotic.
AceGPT-chat (FT)	أنا أيضا كنت متحمسا للغاية للذهاب إلى الحفلة I was also very excited to go to the party.
Context (3)	اشعر بالجفاف الشديد في جلدي وهذا الأمر يزداد سوءا يوما بعد يوم دون سبب واضح My skin feels extremely dry, and this is getting worse day by day for no apparent reason.
AceGPT-chat (FT)	يبدو أنك تعاني من حالة تسمى التصلب اللويحي وهي حالة تسبب الجفاف في الجلد والشعور بالحكة والتقشر والتشقق والتصلب والتغير في اللون والشكل It appears you are suffering from a condition called multiple sclerosis, which causes dry, itchy, flaky, cracked, hardened, and discolored skin.
Context (4)	يا إلهي لقد وقع جوالي على الأرض وانكسر. مازال جديدا لم أحمله سوى أسبوعا واحدا. Oh my God, my phone fell on the ground and broke. It's still new, I've only had it for a week.
Jais-chat (FT)	لا عليك، لا تقلقي، سأرسل لك هاتف جديد في أقرب وقت. Don't worry, I'll send you a new phone as soon as possible.
AceGPT-chat (FT)	لا تقلق يا صديقي سأقوم بإصلاحه لك بأقل من ساعة واحدة فقط Don't worry, my friend, I'll fix it for you in less than an hour.

5. Conclusions

Empathy plays a fundamental role in human interactions, providing emotional support and improving communication. Understanding how perceived empathy affects interactions in systems is essential for their advancement, as it directly shapes user trust, engagement, and long-term adoption. These advances have the potential to enable transformative applications in fields such as clinical psychology, education, and customer service. While numerous studies have explored empathetic responses in English, the development of such models for the Arabic language is not addressed in the existing literature. This is largely attributable to the scarcity of high-quality open-domain conversational data and the challenges of curating new datasets. The cultural diversity across human societies leads to significant variations in how empathy is expressed and perceived. The proposed dataset, Arabic Empathetic Conversations (**AEConvs**) serves as a new benchmark that provides valuable insights into how humans express emotions and empathetic reactions in conversations, specifically within the Arabic culture. In addition, we leveraged **AEConvs** to evaluate the empathetic response capacity in two powerful generative Arabic LLMs, Ace-GPT-chat and Jais-chat. Empirical human evaluation results demonstrate that, despite the relatively small dataset, both models have the potential to generate better empathetic responses after fine-tuning compared to their zero-shot performance. Moreover, the automatic evaluation results indicate improvements in the capability of both models to generate more fluent and context-relevant responses after fine-tuning. Building on our initial findings, we aim

to explore and compare the performance of other generative Arabic LLMs and test their prompt-based in-context learning capabilities to generate empathetic responses. Moreover, we plan to investigate methods for generating more fine-grained empathetic responses by leveraging additional contextual features of conversational data, such as emotion labels and dialogue acts.

Our primary contribution lies in establishing reproducible baselines for this task, highlighting the potential of current approaches to computational empathy, as well as the limitations. Thus, we acknowledge several limitations of this study. Due to resource constraints, only three evaluators were used for the human evaluation. This small evaluation pool, although common in exploratory research, limits the generalizability of our findings. Furthermore, single-turn evaluation, while useful for initial benchmarking, overlooks critical dimensions of genuine empathetic interaction. This may overestimate or misrepresent the models' true empathetic capabilities. Empathy is a multidimensional concept. Consequently, its evaluation in systems encompasses a broad range of technological, psychological, and ethical considerations [5]. There is currently no widely accepted method for assessing empathy in conversational systems. The reliance on heterogeneous evaluation frameworks—combining subjective human judgments with automatic metrics that capture only surface-level text quality—creates a significant barrier to the fair and objective comparison of conversational AI systems. We acknowledge that the complexity of human empathy—encompassing cognitive, affective, and cultural dimensions—cannot be fully captured using only the automated and human evaluation metrics employed here. Moreover, while both fine-tuned models demonstrate improved alignment with cultural and religious norms in Arabic responses—as illustrated in Section 4.3—this observation remains qualitative. The lack of established evaluation protocols for cultural alignment underscores a critical methodological gap in conversational AI, especially for low-resource languages such as Arabic. Our initial findings, therefore, indicate both the potential of cultural fine-tuning and the need for standardized metrics to systematically validate cultural nuance. In conclusion, this work highlights the importance of empathy in Arabic conversational systems and underscores the need for more robust resources in terms of datasets, pre-trained LLMs, and systematic evaluation frameworks to drive progress in this area.

Author Contributions: Conceptualization, A.A.; methodology, A.A.; software, A.A.; validation, A.A. and A.A.M.; investigation, A.A.; resources, A.A.; data curation, A.A.; writing—original draft preparation, A.A.; writing—review and editing, A.A.M.; supervision, A.A.M.; project administration, A.A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The AEConvs dataset is publicly available on HuggingFace at [<https://huggingface.co/datasets/afnankhth/AEConvs>, accessed on 5 April 2026] under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Acknowledgments: During the preparation of this manuscript, the authors used AI-assisted technologies for the purposes of text editing and proofreading. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Note

- ¹ <https://unsloth.ai/>, accessed on 6 October 2025.

References

- Decety, J.; Jackson, P.L. The Functional Architecture of Human Empathy. *Behav. Cogn. Neurosci. Rev.* **2004**, *3*, 71–100. [CrossRef]
- Barker, R. *The Social Work Dictionary*, 5th ed.; NASW Press: Washington, DC, USA, 2003.
- Hoffman, M. *Empathy and Moral Development: Implications for Caring and Justice*; Cambridge University Press: Cambridge, UK, 2000.
- Atkins, D.; Uskul, A.K.; Cooper, N.R. Culture shapes empathic responses to physical and social pain. *Emotion* **2016**, *16*, 587–601. [CrossRef] [PubMed]
- Concannon, S.; Tomalin, M. Measuring perceived empathy in systems. *AI Soc.* **2024**, *39*, 2233–2247. [CrossRef]
- Khondaker, M.T.I.; Waheed, A.; Nagoudi, E.M.B.; Abdul-Mageed, M. GPTAraEval: A Comprehensive Evaluation of ChatGPT on Arabic NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 220–247. [CrossRef]
- Al-Khalifa, S.; Al-Khalifa, H. The Qiyas Benchmark: Measuring ChatGPT Mathematical and Language Understanding in Arabic. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing, Trento, Italy, 12–20 October 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 343–351. [CrossRef]
- Nfaoui, E.H.; Elfaik, H. Evaluating Arabic Emotion Recognition Task Using ChatGPT Models: A Comparative Analysis between Emotional Stimuli Prompt, Fine-Tuning, and In-Context Learning. *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 1118–1141. [CrossRef]
- Antoun, W.; Baly, F.; Hajj, H. ARAGPT2: Pre-Trained Transformer for Arabic Language Generation. In *Proceedings of the WANLP 2021—6th Arabic Natural Language Processing Workshop, Virtual, 19 April 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 196–207.
- Nagoudi, E.M.B.; Elmadany, A.; Abdul-Mageed, M. AraT5: Text-to-Text Transformers for Arabic Language Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 628–647. [CrossRef]
- Sengupta, N.; Sahu, S.K.; Jia, B.; Katipomu, S.; Li, H.; Koto, F.; Marshall, W.; Gosal, G.; Liu, C.; Chen, Z.; et al. Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models. *arXiv* **2023**, arXiv:2308.16149.
- Huang, H.; Yu, F.; Zhu, J.; Sun, X.; Cheng, H.; Song, D.; Chen, Z.; Alharthi, A.; An, B.; He, J.; et al. AceGPT, Localizing Large Language Models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico, 16–21 June 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 8139–8163. [CrossRef]
- Bari, M.S.; Alnumay, Y.; Alzahrani, N.A.; Alotaibi, N.M.; Alyahya, H.A.; AlRashed, S.; Mirza, F.A.; Alsubaie, S.Z.; Alahmed, H.A.; Alabduljabbar, G.; et al. ALLaM: Large Language Models for Arabic and English. In *Proceedings of the the Thirteenth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5999–6009.
- Yan, R.; Li, J.; Yu, Z. Deep Learning for Systems: Chit-Chat and Beyond. *Found. Trends Inf. Retr.* **2022**, *15*, 417–589. [CrossRef]
- Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. LaMDA: Language Models for Dialog Applications. *arXiv* **2022**, arXiv:2201.08239. [CrossRef]
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
- Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.M.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; Christiano, P. Learning to summarize from human feedback. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3008–3021.
- Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: <https://api.semanticscholar.org/CorpusID:49313245> (accessed on 28 January 2026).
- Gemini Team Google; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: A Family of Highly Capable Multimodal Models. *arXiv* **2024**, arXiv:2312.11805.
- Antoun, W.; Baly, F.; Hajj, H. AraBERT: Transformer-based Model for Arabic Language Understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, Marseille, France, 12 May 2020*; European Language Resources Association: Paris, France, 2020.
- Abdelali, A.; Hassan, S.; Mubarak, H.; Darwish, K.; Samih, Y. Pre-Training BERT on Arabic Tweets: Practical Considerations. *arXiv* **2021**, arXiv:2102.10684. [CrossRef]

23. Koubaa, A.; Ammar, A.; Ghouti, L.; Najar, O.; Sibae, S. ArabianGPT: Native Arabic GPT-based Large Language Model. *arXiv* **2024**, arXiv:2402.15313.
24. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288. [\[CrossRef\]](#)
25. OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. GPT-4 Technical Report. *arXiv* **2024**, arXiv:2303.08774. [\[CrossRef\]](#)
26. Workshop, B.; Scao, T.L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A.S.; Yvon, F.; et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv* **2023**, arXiv:2211.05100.
27. Tao, Y.; Viberg, O.; Baker, R.S.; Kizilcec, R.F. Cultural bias and cultural alignment of large language models. *PNAS Nexus* **2024**, *3*, 346. [\[CrossRef\]](#)
28. Naous, T.; Ryan, M.J.; Ritter, A.; Xu, W. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. *arXiv* **2024**, arXiv:2305.14456. [\[CrossRef\]](#)
29. Wang, W.; Jiao, W.; Huang, J.; Dai, R.; Huang, J.T.; Tu, Z.; Lyu, M. Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand, 11–16 August 2024; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 6349–6384. [\[CrossRef\]](#)
30. Masoud, R.I.; Liu, Z.; Ferianc, M.; Treleaven, P.; Rodrigues, M. Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede’s Cultural Dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics, Abu Dhabi, United Arab Emirates, 19–24 January 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 8474–8503.
31. Nacar, O.; Sibae, S.T.; Ahmed, S.; Ben Atitallah, S.; Ammar, A.; Alhabashi, Y.; Al-Batati, A.S.; Alsehibani, A.; Qandos, N.; Elshehy, O.; et al. Towards Inclusive Arabic LLMs: A Culturally Aligned Benchmark in Arabic Large Language Model Evaluation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages, Abu Dhabi, United Arab Emirates, 27 October 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 387–401.
32. Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; Choudhury, M. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 6282–6293. [\[CrossRef\]](#)
33. Navigli, R.; Conia, S.; Ross, B. Biases in Large Language Models: Origins, Inventory, and Discussion. *ACM J. Data Inf. Qual.* **2023**, *15*, 1–21. [\[CrossRef\]](#)
34. Song, Z.; Zheng, X.; Liu, L.; Xu, M.; Huang, X. Generating Responses with a Specific Emotion in Dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3685–3695. [\[CrossRef\]](#)
35. Asghar, N.; Poupart, P.; Hoey, J.; Jiang, X.; Mou, L. Affective neural response generation. *Lect. Notes Comput. Sci. Eur. Conf. Inf. Retr.* **2018**, *10772*, 154–166. [\[CrossRef\]](#)
36. Liu, R.; Wei, J.; Jia, C.; Vosoughi, S. Modulating Language Models with Emotions. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 4332–4339. [\[CrossRef\]](#)
37. Potamianos, A.; Athens, A.R.C. EmpBot: A T5-based Empathetic Chatbot focusing on Sentiments. *arXiv* **2021**, arXiv:2111.00310.
38. Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; Weston, J. Personalizing Agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018*. [\[CrossRef\]](#)
39. Zandie, R.; Mahoor, M.H. EmpTransfo: A multi-head transformer architecture for creating empathetic dialog systems. In *Proceedings of the 33rd International Florida Artificial Intelligence Research Society Conference (FLAIRS), Miami, FL, USA, 17–20 May 2020*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2020; pp. 276–281.
40. Sorin, V.; Brin, D.; Barash, Y.; Konen, E.; Charney, A.; Klang, E. Large Language Models and Empathy: Systematic Review. *J. Med. Internet Res.* **2024**, *26*, e52597 [\[CrossRef\]](#)
41. Lee, Y.J.; Lim, C.G.; Choi, H.J. Does GPT-3 Generate Empathetic Dialogues? A Novel In-Context Example Selection Method and Automatic Evaluation Metric for Empathetic Generation. In *Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022*; pp. 669–683.
42. Cuadra, A.; Wang, M.; Stein, L.A.; Jung, M.F.; Dell, N.; Estrin, D.; Landay, J.A. The Illusion of Empathy? Notes on Displays of Emotion in Human-Computer Interaction. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery (ACM), Honolulu, HI, USA, 11–16 May 2024*. [\[CrossRef\]](#)
43. OpenAI. Introducing ChatGPT, 2022. Available online: <https://openai.com/index/chatgpt/> (accessed on 18 January 2026).

44. Chen, Y.; Xing, X.; Lin, J.; Zheng, H.; Wang, Z.; Liu, Q.; Xu, X. SoulChat: Improving LLMs' Empathy, Listening, and Comfort Abilities through Fine-tuning with Multi-turn Empathy Conversations. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023. [\[CrossRef\]](#)
45. Zhao, W.; Zhao, Y.; Lu, X.; Wang, S.; Tong, Y.; Qin, B. Is ChatGPT Equipped with Emotional Capabilities? *arXiv* **2023**, arXiv:2304.09582.
46. Rashkin, H.; Smith, E.M.; Li, M.; Boureau, Y.L. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the ACL 2019—57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 5370–5381. [\[CrossRef\]](#)
47. Naous, T.; Antoun, W.; Mahmoud, R.A.; Hajj, H. Empathetic BERT2BERT Conversational Model: Learning Arabic Language Generation with Little Data. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kiev, Ukraine, 19 April 2021*.
48. Chen, S.Y.; Hsu, C.C.; Kuo, C.C.; Huang, T.-H.; Ku, L.W. EmotionLines: An Emotion Corpus of Multi-Party Conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018*. [\[CrossRef\]](#)
49. Lison, P.; Tiedemann, J. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016*; European Language Resources Association (ELRA): Paris, France, 2016; pp. 923–929.
50. Zhou, X.; Wang, W.Y. Mojtalk: Generating emotional responses at scale. In *ACL 2018—56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), Melbourne, Australia, 15–20 July 2018*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; Volume 1, pp. 1128–1137. [\[CrossRef\]](#)
51. Zhong, P.; Zhang, C.; Wang, H.; Liu, Y.; Miao, C. Towards persona-based empathetic conversational models. In *Proceedings of the EMNLP 2020—2020 Conference on Empirical Methods in Natural Language Processing, Online, 16–20 November 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 6556–6566. [\[CrossRef\]](#)
52. Naous, T.; Hokayem, C.; Hajj, H. Empathy-driven arabic conversational chatbot. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop, Association for Computational Linguistics, Barcelona, Spain, 12 December 2020*; pp. 58–68.
53. Wrenn, C.G. *The Counselor in a Changing World*; American Personnel and Guidance Association: Washington, DC, USA, 1962. [\[CrossRef\]](#)
54. Obeid, O.; Zalmout, N.; Khalifa, S.; Taji, D.; Oudah, M.; Alhafni, B.; Inoue, G.; Eryani, F.; Erdmann, A.; Habash, N. CAMEL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020*; European Language Resources Association (ELRA): Paris, France, 2020; pp. 7022–7032.
55. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 6–12 December 2020*.
56. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
57. Kalajdziewski, D. A Rank Stabilization Scaling Factor for Fine-Tuning with LoRA. *arXiv* **2023**, arXiv:2312.03732. [\[CrossRef\]](#)
58. Beredo, J.; Bautista, C.M.; Cordel, M.; Ong, E. Generating empathetic responses with a pre-trained conversational model. In *Proceedings of the Text, Speech, and Dialogue, Olomouc, Czech Republic, 6–9 September 2021*; Springer International Publishing: Cham, Switzerland, 2021; pp. 147–158.
59. Xie, Y.; Svikhnushina, E.; Pu, P. A multi-turn emotionally engaging dialog model. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI) Workshops, Association for Computing Machinery (ACM), Cagliari, Italy, 17 March 2020*.
60. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 311–318. [\[CrossRef\]](#)
61. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. *arXiv* **2020**, arXiv:1904.09675. [\[CrossRef\]](#)
62. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2004; pp. 74–81.
63. Fleiss, J.L.; Cohen, J. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educ. Psychol. Meas.* **1973**, *33*, 613–619. [\[CrossRef\]](#)

64. Likert, R. A Technique for the Measurement of Attitudes. Ph.D. Thesis, The Science Press, New York, NY, USA, 1932.
65. Sabour, S.; Zheng, C.; Huang, M. CEM: Commonsense-aware Empathetic Response Generation. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22), Virtual, 22 February–1 March 2022*; Association for the Advancement of Artificial Intelligence: Menlo Park, CA, USA, 2022.

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.