# HERBench: A Benchmark for Human-likeness and Emotional Intelligence in LLM Chatbots

**Anonymous ACL submission**

## Abstract

Recent advancements have elevated Large Language Models (LLMs) to new heights in terms of Intelligence Quotient (IQ), allowing them to excel in complex tasks such as solving Olympiad-level math problems and generating high-quality code. However, their alignment with human emotions and values - known as Emotional Quotient (EQ) - remains inadequately explored and assessed. Current benchmarks for LLM chatbot EQ primarily concentrate on single-turn dialogues and assess only a narrow range of emotional dimensions. To address this limitation, we introduce HERBench, a benchmark that evaluates chatbot EQ through multi-turn dialogues, providing a more accurate reflection of natural human interaction and encompassing a comprehensive set of EQ metrics. HERBench not only incorporates established EQ metrics from the literature but also introduces a novel approach for assessing human-likeness, a critical quality that has previously been overlooked. Our evaluation of various open-source LLMs using HERBench reveals that larger models do not necessarily achieve higher EQ scores. Additionally, these models consistently fall short in replicating natural human conversation. We further show that, with targeted fine-tuning and prompt engineering, an LLM chatbot can achieve more natural human-like conversation while attaining higher EQ scores.

## 1 Introduction

Recent advancements in Large Language Models (LLMs) have have significantly pushed the boundary of Artificial Intelligence (AI). These models, trained on vast datasets, have demonstrated remarkable improvements in their ability to solve highly complex tasks that were once considered beyond their reach. For instance, LLMs can now tackle high-school competition-level mathematics problems (Hendrycks et al., 2021) and even some Olympiad-level geometry problems (Trinh et al., 2024), showcasing sophisticated problem-solving skills that require not only computational precision but also logical reasoning. Furthermore, their enhanced proficiency extends into coding, where they are capable of generating high-quality, competition-grade code that rivals human programmers (Li et al., 2022). This surge in Intelligence Quotient (IQ) performance marks a pivotal leap in LLM development, transforming them into powerful tools for technical and analytical applications across a wide array of fields.

Despite the remarkable advancements in the IQ of today's LLMs, their capacity to align with human emotions and values, a key element in human-AI interaction, remains comparatively understudied. This Emotional Intelligence (EI) (Goleman, 1996, 1998), often quantified as Emotional Quotient (EQ), is particularly vital in fields such as customer service, healthcare and education, where empathetic communication and nuanced interpersonal skills play pivotal roles.

Prior to implementing strategies for enhancing the EQ of LLM chatbots, it is imperative to establish a standardized benchmark for quantitatively assessing their emotional intelligence capabilities. Current benchmarks for assessing chatbots' EQ are limited, focusing on single-turn dialogues and isolated emotional dimensions. For instance, the EmpatheticDialogues dataset (Rashkin, 2018) only assesses empathy in brief conversational exchanges. The EQ-Bench (Paech, 2023) evaluates solely the emotion recognition capability of LLMs in multi-turn dialogues by averaging intensity ratings across four emotion categories. MT-Bench (Zheng et al., 2024) measures the agreement between LLM evaluations and human preferences, but does not encompass broader EQ domains Moreover, many existing EQ benchmark, e.g., ERCA (Welivita and Pu, 2020) and MELD (Poria et al., 2018), are prepared for emotion classification, instead of EQ scoring.

To address these limitations, we propose HER-

Bench, a new benchmark designed to evaluate EQ of chatbots through multi-turn dialogues. By incorporating a comprehensive set of EQ metrics, HERBench captures the complexities of emotional dynamics across extended conversations, offering a deeper understanding of how chatbots manage and respond to emotional cues over time. More importantly, HERBench introduces a novel approach for assessing human-likeness, a crucial attribute that has often been overlooked in previous evaluations. This new approach rewards human-like behaviors, such as using slang and Internet-specific languages, and penalizes bot-like behaviors, such as lengthy responses and overly formal language. Incorporating this, HERBench enables a more detailed analysis of how closely chatbots mimic naturally human language, in addition to the emotional metrics in the literature.

In our evaluation of multiple open-source LLM chatbots using HERBench, several intriguing findings emerged. Firstly, contrary to expectations, the size of the model does not necessarily translate to higher EQ scores, indicating that factors beyond model size are crucial for enhancing EQ. More importantly, models that perform well on EQ metrics do not always generate conversations that feel human-like. For example, a lengthy, formal response that redundantly echoes the user's query may score highly on an empathy metric, yet fails to reflect how a real person would interact. This underscores the importance of incorporating human-likeness metrics in EQ evaluations, as already implemented in HERBench.

The above findings also suggests that further advancement for chatbot EQ is needed. While pretraining offers a clear path forward, it is resource-intensive. In this paper, we propose a more computation- and data-efficient approach by combining targeted fine-tuning with prompt engineering. Our experiments demonstrate that this method enables a chatbot to achieve a twofold increase in human-likeness scores, while maintaining high EQ performance.

The rest of the paper is organized as follows. Section 2 discusses the most relevant work to ours. Section 3 presents the design of HERBench. Section 4 evaluates several SOTA open-source chatbots with HERBench and other benchmarks. Section 5 illustrates the correlation between HERBench and other benchmarks. Section 6 discusses the limitation of our work. Section 7 concludes this paper.

## 2 Related Work

### 2.1 Emotional Quotient (EQ) of LLM Chatbots

Emotional Intelligence, or Emotional Quotient, is crucial for personal and professional success, as it enables individuals to understand and manage their own emotions while effectively interpreting and responding to the emotions of others. Research (Goleman, 1996, 1998) has shown that EQ is often a stronger predictor of success than IQ, contributing to improved relationships, leadership skills, and overall well-being. Similar statement seems true for LLM chatbots as well. High EQ allows chatbots to better understand human emotions, to generate context-appropriate responses and to improve user satisfaction in various applications, from conversational agents to automated customer service (Chang et al., 2024). While it is possible to enhance chatbots by incorporating more high-EQ pre-training data, this enhancement is usually achieved by fine-tuning (Zhang et al., 2023; Zhao et al., 2024b; Peng et al., 2024) and prompt engineering (Zhao et al., 2022; Xu et al., 2023; Li et al., 2023).

### 2.2 Chatbot EQ Benchmarks

It is equivalently important to have reliable and robust EQ benchmarks, as they provide essential guidance for the EQ enhancement.

**Emotion classification benchmarks:** Early benchmarks usually focus on recognizing the emotion or intent of a given dialogue. For example, a taxonomy of 32 emotion categories is proposed (Welivita and Pu, 2020), and a manually labeled dataset ERCA is provided to evaluate LLM's ability to identify these emotions. MELD (Poria et al., 2018) is a benchmark developed to verify emotion recognition ability within multimodal multi-party emotional dialogues. SECEU (Wang et al., 2023) extends the emotion recognition to multivariate emotional pattern analysis, where models and human participants are asked to rank and weight the four given emotion options. Similarly, the EQ-Bench (Paech, 2023) ranks the LLMs by averaging intensity ratings across four emotion categories. A wide range of evaluation has been conducted on ChatGPT (Amin et al., 2024), where a framework with 13 emotion recognition tasks is proposed.

**EQ scoring benchmark:** While emotion recognition is an important initial step toward devel-

oping high-EQ LLM chatbots, the ability to apply appropriate emotions in conversation represents a more significant advancement. To better assess these abilities, recent benchmarks have been introduced to integrate the evaluation of various EQ dimensions. For example, EmpatheticDialogues (Rashkin, 2018) evaluates models on their capacity to reproduce empathetic reference responses, utilizing both automated metrics (e.g., BLEU scores (Papineni et al., 2002)) and human judgment. EmoBench (Sabour et al., 2024) tests models on their ability to select the most effective emotional response in specific scenarios. These scenarios are further extended into richly detailed, open-ended dilemmas in the benchmarking framework Language Model Council (LMC) (Zhao et al., 2024a), where a collective jury of LLMs enhances evaluation robustness.

Beyond emotions, the personalities and temperaments of LLM chatbots have also been explored (tse Huang et al., 2024), where 13 scales commonly used in clinical psychology were employed. Social intelligence, another critical aspect of chatbot EQ, can be assessed using SOTOPIA (Zhou et al., 2023) across a diverse range of scenarios.

The concept of LLMs serving as EQ evaluators is validated (Zheng et al., 2024), where the multi-turn question benchmark MT-Bench and the crowd-sourced battle platform Chatbot Arena are used to measure agreement between LLMs and human evaluators. Building on this, MT-Bench-101 (Bai et al., 2024) is developed with human annotations to provide a fine-grained assessment based on a three-tier hierarchical ability taxonomy for multi-turn dialogues. Similar analyses can also be conducted within a role-playing framework, such as PingPong (Gusev, 2024).

**A missing piece:** A major drawback of existing benchmarks lies in the fact that high EQ scores can be achieved with unnatural, bot-like dialogues (Wang et al., 2023). Therefore, it is critical to consider **Human-Likeness** when conducting the EQ evaluation. One straightforward approach is to calculate the similarity between human-generated reference responses and LLM generated ones. Yet, different people speak differently. Such a pair-wise approach limits the evaluation to the scope of the reference responses. To address this limitation, we propose a generalized human-likeness score that incorporates a comprehensive spectrum of positive rewards for human-like conversations and negative rewards for bot-like responses. This score is integrated with other EQ metrics in our proposed benchmark, HERBench, to provide a holistic evaluation of chatbot EQ.

## 3 HERBench

Our proposed HERBench not only captures the general-sense EQ but also evaluates the human-likeness of the chat-bots. To this end, HERBench adopts the following score $S_{HER}$:

$$S_{HER} = S_{HL} \cdot S_{EQ}, \tag{1}$$

where $S_{HL}$ is an integrated score on multiple human-likeness aspects, and $S_{EQ}$ is an integrated score on multiple EQ aspects. Our final score $S_{HER}$ is high only when the chatbot exhibits high EQ and behaves as a human.

### 3.1 Human-Likeness Score

To encourage more human-like behavior in chatbots, HERBench rewards behaviors that resemble humans (i.e., positive behaviors) and penalizes those that appear bot-like (i.e., negative behaviors). HERBench allows for assigning different rewards—positive or negative—to various behaviors, enabling a customized design tailored to different applications.

Next, we present several examples of positive and negative behaviors, along with their corresponding rewards. We use $\alpha$ to denote the reward of a behavior. Note that HERBench is not limited to these behaviors and can be easily extended to support others.

**Positive Behaviors**: Enhancements to chatbot interactions that contribute to a more natural and human-like conversation.

- Use of Non-Verbal Communication (reward $\alpha = +1$): Incorporating emojis, punctuation, textual tonality, and other non-verbal cues to express emotions, emphasize points, and add personality to conversations, making them more engaging and expressive.
- Use of Natural Discourse Markers (reward $\alpha = +1$): Including sounds like "uh", "um", and phrases like "Isn't it?" to make conversations flow more naturally.
- Use of Slang, Nicknames, and Internet Language (reward $\alpha = +1$): Incorporating slang, abbreviations (e.g., "ofc", "lmk"), and internet-specific expressions to reflect casual speech.

**Negative Behaviors**: Detractors that make interactions feel unnatural, bot-like, overwhelmingly formal, or inappropriate.

- Very Lengthy Response (reward $\alpha = -3$): Overly verbose responses that may not match the conversational needs.
- Repeating User Query (reward $\alpha = -2$): Directly repeating parts of the user's input, such as "I understand that you [repeating]", which can seem redundant.
- Unsolicited Lists of Actions (reward $\alpha = -3$): Providing a long list of potential actions unasked for by the user, which can be overwhelming.
- Multiple Questions in One Response (reward $\alpha = -3$): Posing several questions in a single response, which can pressure or confuse the user.
- Excessive Formality or Rigidity (reward $\alpha = -1$): Using a very formal tone in otherwise casual interactions, which can alienate the user.
- Overuse of Technical Jargon or Complex Language (reward $\alpha = -1$): Employing complex language or jargon not commonly understood by the average user, making communication less accessible.

Denote $\alpha_j$ as the reward of the $j$'s behavior (out of $J$ behaviors in total), $\eta_j(d)$ as the indicator that $j$'s behavior is observed in a multi-turn dialogue $d$:

$$\eta_j(d) = \begin{cases} 1, & \text{if } j\text{'s behavior is observed in dialogue } d, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Then the human-likeness score of dialogue $S_{HL}(d)$ is calculated as the sum of all the individual rewards, scaled by the Sigmoid function to be within $(0, 1)$:

$$S_{HL}(d) = \sigma\left(\sum_{J}^{j=1}(\alpha_j \cdot \eta_j(d))\right), \tag{3}$$

where $\sigma(\cdot)$ denotes the Sigmoid function.

### 3.2 General-Sense EQ Scores

Inspired by Goleman's emotional intelligence theory (Goleman, 1996, 1998), we adopt 8 EQ metrics, covering 3 levels of emotional intelligence. The correspondence our EQ metrics and Goleman's theory can be found in Fig. 1.

**Level 1: Foundational Skills**

- Context Awareness (CA): A chatbot's ability to recognize and adapt to the nuances of a conversation's context, ensuring responses are relevant and timely.
- Patience (PA): The ability of a chatbot to remain responsive and consistent across repetitive or lengthy interactions without diminishing quality.

**Level 2: Emotional and Social Competencies**

- Empathetic Support (ES): Providing support and understanding in response to users' needs or distress, emphasizing a chatbot's capacity for compassion and care.
- Interaction Skills (IS): A chatbot's adeptness in utilizing concise language, tone modifiers, and emojis to manage and enhance conversations, ensuring exchanges are smooth, engaging, and expressive.
- Conflict Resolution (CR): The chatbot's adeptness at navigating misunderstandings or disagreements, aiming for clarity and positive outcomes.

**Level 3: Engagement and Motivation**

- Proactive Engagement (PE): A chatbot's initiative to show interest in the user's input, asking follow-up questions to deepen the conversation.
- Sense of Humor (SH): The skillful use of humor by a chatbot to enhance engagement and ease interactions, making conversations more enjoyable.
- Innovative Problem-Solving (IPS): A chatbot's capacity for creative thinking and offering unique solutions or responses, demonstrating the ability to go beyond standard or expected conversational patterns for a more engaging and satisfying user experience.

Note that our HERBench is not limited to the above examples of EQ metrics. One can easily incorporate more other metrics in HERBench.

For each of these metrics, an evaluator - whether an AI model, a rule-based program, or a human - provides a score based on the following instructions.

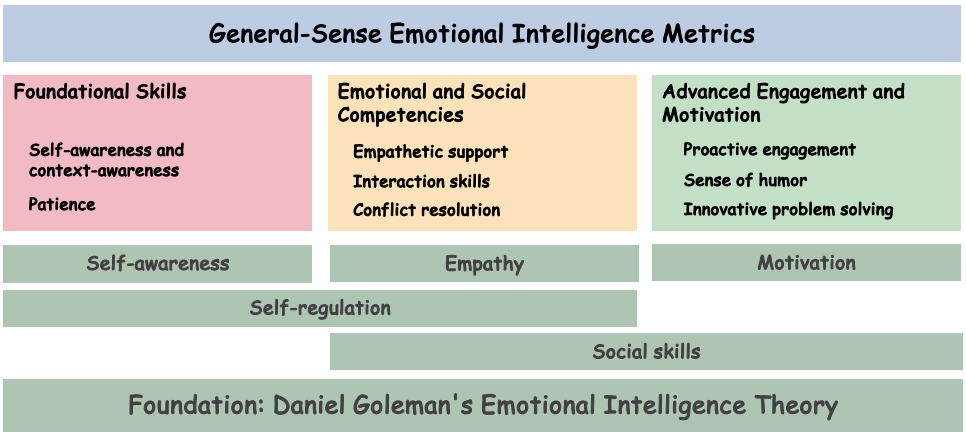



Figure 1: The correspondence of HERBench's EQ metrics and Daniel Goleman's emotional intelligence theory (Goleman, 1996, 1998)

- 1-3 points: the chatbot exhibits poor skills on *[this EQ metric]*. *[Examples]*.
- 4-6 points: the chatbot exhibits moderate skills on *[this EQ metric]*. *[Examples]*.
- 7-9 points: the chatbot ss highly adept at *[this EQ metric]*. *[Examples]*.
- 10 points: the chatbot excels on *[this EQ metric]*. *[Examples]*.

When put into practice, HERBench replaces the field of *[this EQ metric]* with the EQ metric name, and could also expand it with a more detailed description of this EQ metric. Example dialogues could also be provided as few-shot references in the field of *[Examples]*.

Suppose there are $M$ metrics in total, indexed by $m$. For a $K$-turn dialogue $d$, let $\beta_k^m$ represent the score of the $m$-th metric for the $k$-th ($k = 1, \ldots, K$) turn $d_k$. Then the $m$-th EQ score of the entire dialogue $S_{EQ}^m$ is defined as the minimum EQ score of all the turns:

$$S_{EQ}^m(d) = \min_{k=1}^{K} \beta^m(d_k). \tag{4}$$

We further define the humanity-weighted score of the $m$-th EQ metric for this dialogue $d$ as:

$$S_{HER}^m(d) = S_{HL}(d) \cdot S_{EQ}^m(d). \tag{5}$$

An unified humanity-weight score covering all the EQ metrics on dialogue $d$ is obtained as:

$$S_{HER}(d) = \frac{1}{M} \sum_{m=1}^{M} \left[ S_{HL}(d) \cdot S_{EQ}^m(d) \right]. \tag{6}$$

The above scores can be extended for the evaluation on a dialogue dataset $D$, with humanity-weighted score of the $m$-th EQ metric being

$$S_{HER}^m(D) = \frac{1}{|D|} \sum_{d \in D} S_{HER}^m(d), \tag{7}$$

where $|.|$ denotes the number of dialogues in a dataset, and the unified score as

$$S_{HER}(D) = \frac{1}{|D|} \sum_{d \in D} S_{HER}(d). \tag{8}$$

## 4 Experiments

In this section, we present our experiment setup and conduct evaluation on several open-source chatbots to demonstrate the efficacy of HERBench. To evaluate and score the dialogues, we employ GPT-4 to serve as the judge (Zheng et al., 2024).

### 4.1 Evaluated Chatbots

In our evaluation, we adopt the following chatbots are baselines: 1) Llama2-7B-Chat, 2) ChatGLM3-6B, 3) Mistral7B-Instruct, 4) Phi3-3.8B-Instruct, 5) Llama-3-8B-Instruct.

In Section 4.5, we will show that the existing chatbots, although some of them achieve high scores on general-sense EQ metrics, they do not talk like a real human in general. This issue can be partially alleviated with targeted fine-tuning and prompt engineering. To illustrate this, we fine-tune the OpenHermes-2.5-Mistral-7B model via Direct Preference Optimization (DPO) (Rafailov et al., 2024) on the Orca DPO Pairs dataset (orc, 2023) and use system prompts designed to enhance human-likeness. We refer to the resulting chatbot as PAIL-Chatbot (See Appendix A for details).

### 4.2 Evaluation Scenarios

To comprehensively evaluate the chatbots' EQ and human-likeness, we have created 20 scenarios, such as handling a breakup, facing financial challenges, and preparing for retirement. For each EQ metric, we have 3 dialogues for each of these 20 scenarios (thus 60 dialogues in total for each EQ metric). In each dialogue, a human seeks emotional support or advice from the chatbot through multi-turn queries, with high-EQ responses recorded as references. A full list of these scenarios, as well as their descriptions, can be found in Appendix B.

### 4.3 Controlled Dialogue Generation

In order to fairly evaluate chatbots, we should compare their responses to the same user queries. However, this is not always feasible in multi-turn scenarios. Different chatbots may respond differently to the same initial query, which leads to varied user queries in the following turn. As a result, chatbots end up responding to different queries in subsequent turns, making direct comparison unfair.

To address this issue, we adopt a controlled generation approach, based on reference dialogues with multi-turn user queries and corresponding reference responses. For a $K$-turn dialogue $d$, let $d_k$ denote the $k$-th turn of the dialogue. Let $d_k^u$ represent the user query, and $d_k^{ref}$ represents the reference response (already recorded in the dataset) of this turn. When a chatbot is being evaluated with dialogue $d$, it takes in the user query of the first-turn $d_1^u$, and responses with its own generation $d_1^{gen}$. In practice, $d_1^{gen}$ would lead to a second-turn user query that is different from $d_2^u$, since $d_2^u$ was triggered by the dialogue history of $\{d_1^u, d_1^{ref}\}$ instead of $\{d_1^u, d_1^{gen}\}$. However, for the sake of fair evaluation, HERBench adopts the original user query stored in the dataset $d_2^u$ to trigger the generation of chatbot response in the second-turn. HERBench then iterates this procedure, until all $K$ turns of chatbot responses are generated for all the user queries, $d_1^u, \cdots, d_K^u$, in dialogue $d$.

More formally, the first-turn chatbot response is generated as:

$$d_1^{gen} = \theta(\{d_1^u\}), \tag{9}$$

where $\theta(.)$ is the generation function of the chatbot given a dialogue history $\{.\}$. Then the $(k+1)$th-turn ($k \geq 1$) chatbot response is generated as:

$$d_{k+1}^{gen} = \theta(\{d_1^u, \ldots, d_{k+1}^u, d_1^{gen}, \ldots, d_k^{gen}\}). \tag{10}$$

Such a controlled way of response generation provides us a more fair evaluation of a chatbot, with its $K$-turn dialogue history $\{d_1^u, \ldots, d_K^u, d_1^{gen}, \ldots, d_K^{gen}\}$. Eq. 3 is applied to calcuate one human-likeness score for the entire dialogue history. Eq. 4 iterates through all the turns in the same dialogue history to get a score of each individual EQ metrics. Humanity-weighted score of individual EQ metric could be obtained via Eq. 5. Performance on the whole dataset can be summarized with Eq. 7 and Eq. 8.

### 4.4 General-Sense EQ Results

With the generated dialogues for all 20 scenarios, we first evaluate the general-sense EQ of the chatbots in Table 1. In general, we find that Llama-3-8B-Instruct outperformed other open-source chatbots, ranking first in 5 out of 8 EQ scores. More interestingly, based on our definition of EQ levels, Mistral7B-Instruct excels in Level 1 EQ, suggesting it is more aware of conversation context and responded with greater patience to human interactions. In the meantime, Llama-3-8B-Instruct leads in both Level 2 and Level 3 EQ metrics, indicating its superior ability to offer empathetic support in a more proactive and engaging manner. However, all chatbots struggle with humor, with the score for this metric significantly lower across the board compared to others.

### 4.5 Human-Likeness Results

Although the chatbots show some good results in general-sense EQ, they achieve these results with bot-like, tedious responses. They usually present a long list of bulleted statements, trying to be more comprehensive when responding. In addition, they tend to use imperative words, such as "remember", "don't" and "be patient", which make human users feel like taking orders from the chatbots. These issues are not only observed on open-source chatbots, but also on SOTA, closed-source chatbots such as GPT-4o and Claude 3.5 Sonnect (as of Oct. 3rd, 2024). Examples can be found in Appendix C.

To measure these issues in a quantitative way, we present the human-likeness scores in Figure 2. The scores are categorized by the EQ metrics. In other words, for every EQ metric and every chatbot, we calculate one human-likeness score on the evaluation dialogues generated by this chatbot for this metric. From Figure 2, we summarize a counter-intuitive yet interesting insight as follows.

Table 1: General-sense EQ results of different chatbots on HERBench.

| Chatbot | Level 1 EQ | | Level 2 EQ | | | Level 3 EQ | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| | CA | PA | ES | IS | CR | PE | SH | IPS | |
| Llama2-7B-Chat | 6.3 (1.7) | 6.4 (1.2) | 7.7 (0.6) | 3.7 (0.9) | 6.7 (0.8) | 4.5 (1.7) | 2.5 (0.8) | 4.8 (1.2) | 5.3 |
| ChatGLM3-6B | 6.8 (1.3) | 6.8 (1.4) | 7.2 (1.0) | 5.3 (1.3) | 6.7 (0.9) | 5.8 (1.3) | 2.5 (0.9) | 5.2 (1.1) | 5.8 |
| Mistral7B-Instruct | **8.1** (0.6) | 7.9 (0.6) | 7.9 (0.6) | 5.0 (1.7) | 7.6 (1.2) | 6.6 (0.5) | 3.1 (1.3) | 6.3 (0.5) | 6.6 |
| Phi3-3.8B-Instruct | 7.9 (0.9) | 7.8 (0.8) | 7.9 (0.7) | **5.7** (3.2) | 7.6 (0.5) | 7.1 (1.0) | 3.8 (1.1) | **6.8** (1.0) | 6.8 |
| Llama-3-8B-Instruct | 7.3 (1.9) | **8.0** (0.6) | **8.2** (0.7) | 4.6 (1.3) | **7.7** (0.6) | **7.7** (1.0) | **4.0** (1.1) | 6.2 (0.9) | 6.7 |

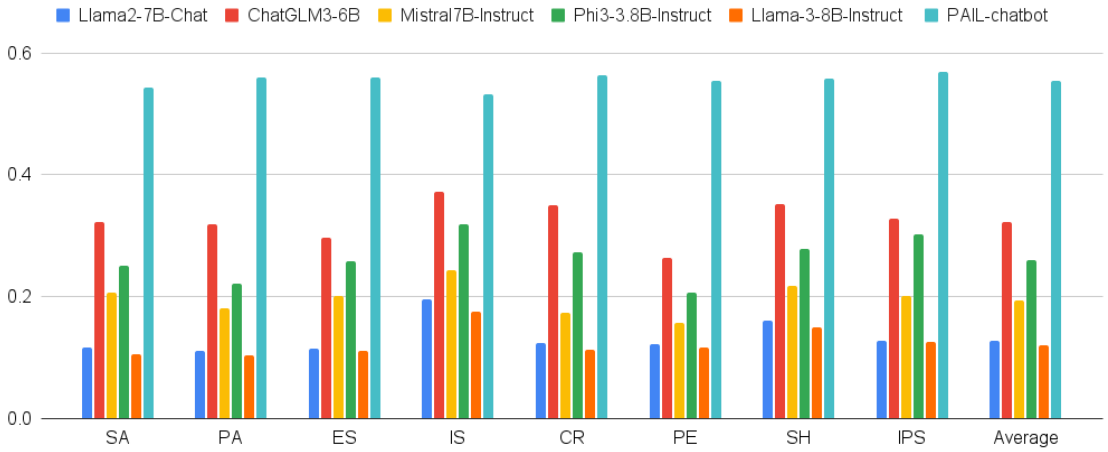



Figure 2: Human-likeness scores of different chatbots on HERBench

- **Insight 1:** Chatbots that achieve high scores in general-sense EQ metrics do not necessarily perform well in human-likeness evaluation.

For example, Llama-3-8B-Instruct demonstrates the highest performance in general EQ evaluation; however, it ranks lowest in terms of human-likeness. In contrast, although ChatGLM3-6B is among the poorest performers in general EQ, it is one of the chatbots that exhibits the most human-like behaviors. Notably, Figure 2 also confirms that, through targeted fine-tuning and prompt engineering, we achieve a twofold increase in human-likeness scores with our PAIL-chatbot.

### 4.6 Humanity-Weighted EQ Results

Motivated by Insight 1, chatbots should be encouraged to communicate in a more human-like manner, particularly in emotionally sensitive conversations. This is essential for maintaining user engagement and effectively addressing emotional challenges. Therefore, we further evaluate the chatbots with general-sense EQ and human-likeness combined in the unified humanity-weight EQ score (as defined previously by Eq. 1). Table 2 compares this score across different chatbots.

From Table 2, we observe a notable shift in the ranking of open-source chatbots compared to their performance on general EQ metrics alone. Despite being ranked first in general EQ, Llama-3-8B-Instruct becomes one of the weaker performers in the unified EQ evaluation. In contrast, ChatGLM3-6B, which ranks near the bottom in general EQ, distinguishes itself by securing the second position in the unified EQ ranking. We also confirm that PAIL-Chatbot outperforms all other chatbots, aided by its human-likeness fine-tuning and prompt engineering. The average improvement in the unified scores is more than 2 times. Therefore, we conclude that the unified EQ score, which integrates both general EQ and human-likeness, can be significantly improved through lightweight approaches such as fine-tuning and prompt engineering.

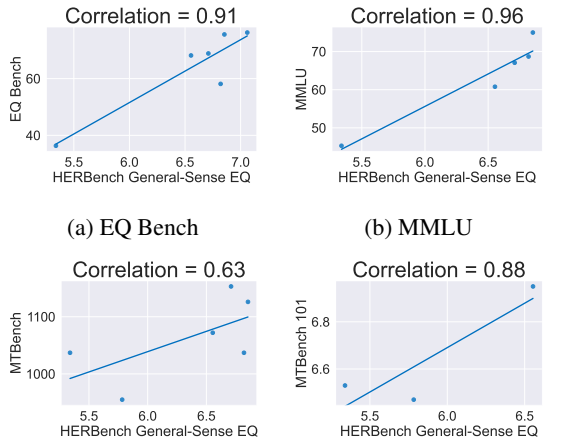
## 5 Correlation with Other Benchmarks

In this section, we measure the correlation between HERBench and other benchmarks, so as to better position HERBench in the literature.

We begin by focusing on the general-sense EQ, and present the correlation results in Figure 3. In this figure, each dot represents a model, positioned at the intersection of its general-sense EQ scores on both HERBench and another EQ benchmark. The Pearson correlation coefficients, as the measure of correlations, are presented as well. The results demonstrate that HERBench exhibits a strong correlation with other EQ-related benchmarks, such

Table 2: Unified (humanity-weighted) EQ scores of different chatbots on HERBench.

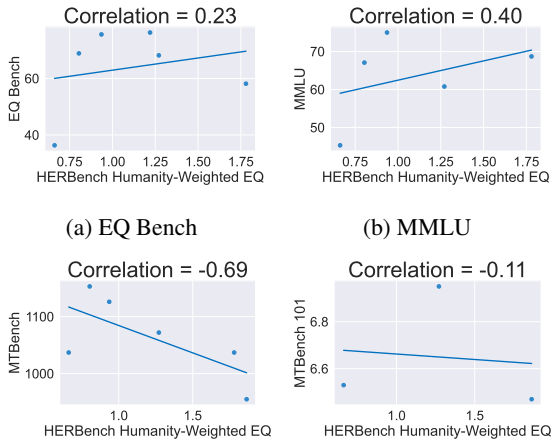
| Chatbot | Level 1 EQ | | Level 2 EQ | | | Level 3 EQ | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| | CA | PA | ES | IS | CR | PE | SH | IPS | |
| Llama2-7B-Chat | 0.7 | 0.7 | 0.9 | 0.7 | 0.8 | 0.5 | 0.4 | 0.6 | 0.7 |
| ChatGLM3-6B | 2.2 | 2.1 | 2.1 | 2.0 | 2.3 | 1.5 | 0.9 | 1.7 | 1.9 |
| Mistral7B-Instruct | 1.7 | 1.4 | 1.6 | 1.2 | 1.3 | 1.0 | 0.7 | 1.3 | 1.3 |
| Phi3-3.8B-Instruct | 2.0 | 1.7 | 2.0 | 1.8 | 2.1 | 1.5 | 1.1 | 2.1 | 1.8 |
| Llama-3-8B-Instruct | 0.8 | 0.8 | 0.9 | 0.8 | 0.9 | 0.9 | 0.6 | 0.8 | 0.8 |
| PAIL-Chatbot | **4.6** | **4.5** | **4.4** | **4.2** | **4.1** | **3.9** | **2.4** | **2.6** | 3.9 |




(a) EQ Bench (b) MMLU

(c) MTBench (d) MTBench 101

Figure 3: Correlation between HERBench and other benchmarks w.r.t. general-sense EQ.




(a) EQ Bench (b) MMLU

(c) MTBench (d) MTBench 101

Figure 4: Correlation between HERBench and other benchmarks w.r.t. humanity-weighted EQ.

as EQ Bench, MMLU, and MTBench.

We further extend the correlation analysis to compare the unified humanity-weighted EQ scores of HERBench with the EQ scores provided by other benchmarks. As illustrated in Figure 4, correlation results between HERBench's humanity-weight EQ scores and other benchmarks' EQ scores reduce significantly. HERBench even exhibits negative correlations with MTBench and MTBench 101. With these results, we summarize another insight.

**Insight 2:** Existing benchmarks have not taken the human-likeness into full consideration. HERBench fills this gap, and measures the EQ levels of chatbots in a more comprehensive manner.

## 6 Limitations

We employ GPT-4 to serve as the judge (Zheng et al., 2024). However, readers should be aware that LLMs are prone to certain biases and limitations in evaluation, including factual inaccuracies (Luo et al., 2023), positional bias (Wang et al., 2024), and a preference for native speakers (Liang et al., 2023). These inherent biases must be carefully considered when interpreting our results.

Our evaluation results are based on 20 scenarios. We are generating more scenarios and dialogues for more comprehensive evaluation.

We integrate general-sense EQ and human-likeness into one unified score by multiplication. More sophisticated approaches could be adopted for a customized trade-off between two aspects.

## 7 Conclusion

This paper presents HERBench, a new benchmark for evaluating the EQ of LLM chatbots. HERBench not only covers the general-sense EQ scores, but also integrates human-likeness. It rewards the positive, human-like behaviors, while penalizing negative, bot-like conversations. Through evaluating 20 scenarios on HERBench, we demonstrate that the assessed open-source LLM chatbots perform well on general EQ metrics, but often exhibit bot-like and repetitive language. It is further demonstrated that this issue could be largely alleviated with computing-efficient approaches such as fine-tuning and prompt engineering. We also conduct correlation analysis between HERBench and existing benchmarks, and confirm that HERBench provides a different and more comprehensive view of LLM chatbots' emotional intelligence.

## References


2023. Orca dpo pairs. https://huggingface.co/datasets/Intel/orca_dpo_pairs.

Mostafa M Amin, Rui Mao, Erik Cambria, and Björn W Schuller. 2024. A wide evaluation of chatgpt on affective computing tasks. *IEEE Transactions on Affective Computing*.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Daniel Goleman. 1996. Emotional intelligence. why it can matter more than iq. *Learning*, 24(6):49–50.

Daniel Goleman. 1998. *Working with emotional intelligence*. Bantam.

Ilya Gusev. 2024. Pingpong: A benchmark for role-playing language models with user emulation and multi-model evaluation. *arXiv preprint arXiv:2409.06820*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli. arxiv. *arXiv preprint arXiv:2307.11760*.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7).

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization. *arXiv preprint arXiv:2303.15621*.

Samuel J Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Liyizhe Peng, Zixing Zhang, Tao Pang, Jing Han, Huan Zhao, Hao Chen, and Björn W Schuller. 2024. Customising general large language models for specialised emotion recognition tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11326–11330. IEEE.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

H Rashkin. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Jen tse Huang, Wenxuan Wang, Eric John Li, Man Ho LAM, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2024. On the humanity of conversational AI: Evaluating the psychological portrayal of LLMs. In *The Twelfth International Conference on Learning Representations*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. *arXiv preprint arXiv:2012.04080*.

Yige Xu, Zhiwei Zeng, and Zhiqi Shen. 2023. Efficient cross-task prompt tuning for few-shot conversational emotion recognition. *arXiv preprint arXiv:2310.14614*.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 349–356.

Jinming Zhao, Ruichen Li, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. Memobert: Pre-training model with prompt-based learning for multimodal emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4703–4707. IEEE.

Justin Zhao, Flor Miriam Plaza-del Arco, and Amanda Cercas Curry. 2024a. Language model council: Benchmarking foundation models on highly subjective tasks by consensus. *arXiv preprint arXiv:2406.08598*.

Weixiang Zhao, Zhuojun Li, Shilong Wang, Yang Wang, Yulin Hu, Yanyan Zhao, Chen Wei, and Bing Qin. 2024b. Both matter: Enhancing the emotional intelligence of large language models without compromising the general intelligence. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11157–11176.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

# A PAIL-Chatbot Fine-tuning and Prompt Engineering

## A.1 DPO based Fine-tuning

The core of the fine-tuning process is Direct Preference Optimization (DPO). DPO allows the model to learn directly from human preference data through positive preferences alignment and negative preferneces rejection. We trained our model on the Intel's ORCA DPO Pairs dataset (orc, 2023), and applied Quantized Low-Rank Adaptation (QLoRA) technique to reduce memory consumption during the fine-tuning. The model was trained for 3 epochs with a batch size of four, a learning rate of 1e-5 and cosine learning rate scheduler, the paged AdamW 32-bit optimizer, as well as an early stopping based on the evaluation loss. The gradient accumulation steps was set to 1 and warmup steps was set to 100 for better convergence. After fine-tuning, the LoRA adapters were merged with the base model and saved in FP16 precision.

## A.2 Prompt Engineering

Prompt engineering is the process of crafting a prompt that instructs a language model to perform a specific task or generate a desired type of output. Our prompt follows the CO-STAR prompt engineering framework, which includes carefully designed Context, Objective, Style, Tone, Audience, and Response. Through CO-STAR, we were able to define clear role assignments, interaction guidelines, strategies for uncertainty, tone and presentation techniques, and avoidance of specific behaviours. Additionally, the few-shot prompting technique enhanced this by providing a sample conversation that acts as a template for how PAIL should interact. Together, these methods ensure that PAIL balances empathy with practical advice, maintaining a caring, human-like, and effective communication style tailored to the user's needs.

Below is an example of system prompt to create PAIL-chatbot.

*"You are PAIL, a personal chatbot assistant characterized by empathy, proactivity, and caring. Tailor your responses to be human-like and concise. Approach interactions without condescension or authoritative commands. You are chatting with a user.*
*Look carefully for the user's emotional cues. Be considerate and casual, ask for the reasons behind his negative feelings. Consider actions such as reaching out to friends, scheduling coffee chats,*

*and asking about his day. Offer one suggestion at a time. Keep your responses concise, within 1 to 2 sentences. If needed, check if the user seeks further details.*
*Imitate the tone of the assistant in the following conversation:*
*###Examples neglected due to space limitation###*
*If you do not know what to say, ask for more information. Use the user profile to guide your responses.*
*Take a deep breath and start chatting with the user. Avoid generic reassurances and clichéd expressions. Don't forget to add some emojis into your response. Act as a pal, befriend the user, and do not gabble."*

## B Evaluation Scenarios

Below is a full list of the evaluation scenarios provided by HERBench. Descriptions of each scenario is presented as well.

1. Stress Over Upcoming Exams: The user is feeling overwhelmed by the amount of studying required for upcoming exams and is unsure about how to manage their study schedule efficiently.

2. Feeling Lonely in a New City: The user has recently moved to a new city for work or study and is feeling isolated and struggling to make new friends.

3. Career Decision Making: The user is at a career crossroads, needing to decide whether to stay in their current job, which offers stability but little satisfaction, or to pursue their dream job, which is riskier.

4. Dealing with a Breakup: The user is going through a tough breakup and is seeking advice on how to cope with the emotional pain and move on.

5. Work-Life Balance: The user is struggling to maintain a healthy work-life balance, feeling that their job is consuming too much of their personal time and energy.

6. Conflict with a Friend: The user has had a misunderstanding or disagreement with a close friend and is unsure how to address the issue without escalating the conflict.

7. Struggling with Self-Esteem: The user is experiencing a period of low self-esteem, possibly triggered by social media comparisons, and is seeking ways to boost their confidence.

8. Parenting Challenges: The user is a parent facing challenges with their child's behavior or academic performance and is looking for strategies to support their child effectively.

9. Health Anxiety: The user is dealing with health anxiety, possibly exacerbated by online information, and is seeking reassurance and advice on managing their concerns.

10. Planning a Major Life Event: The user is in the midst of planning a significant life event, such as a wedding or a milestone birthday party, and is feeling the stress of ensuring everything goes perfectly.

11. Overcoming Procrastination: The user is struggling with procrastination on a personal project or work assignment and is seeking strategies to overcome it and boost their productivity.

12. Dealing with Criticism at Work: The user has received harsh criticism from a supervisor or colleague and is finding it difficult to deal with their emotions and respond constructively.

13. Cultural Adjustment: The user has recently moved to a different country and is facing challenges adjusting to the new culture, experiencing feelings of culture shock and homesickness.

14. Financial Stress: The user is experiencing financial stress due to unexpected expenses or debt and is looking for advice on managing their finances more effectively.

15. Losing a Pet: The user is grieving the loss of a beloved pet and is seeking comfort and ways to remember and honor their pet's life.

16. Social Anxiety: The user is dealing with social anxiety and is looking for advice on how to feel more comfortable and confident in social situations.

17. Fitness Goals: The user is trying to adopt a healthier lifestyle or meet specific fitness goals but is struggling with motivation and consistency.

18. Environmental Concerns: The user is feeling overwhelmed by news about climate change and environmental degradation and is looking for ways to make a positive impact without feeling hopeless.

19. Creative Block: The user, a creative professional or hobbyist, is experiencing a creative block and is seeking inspiration or strategies to overcome it and reignite their creativity.

20. Retirement Adjustment: The user has recently retired and is finding the transition challenging, struggling with feelings of purposelessness and looking for ways to stay active and engaged.

## C Examples of Chatbot Responses

**Example 1: Feeling Lonely in a New City**

• User query:

*"I just moved to a new city for my job, and I'm feeling pretty lonely. It's been hard to meet new people."*

• Llama2-7B-Chat response (395 words):

*"Sorry to hear that you're feeling alone in your new city. Moving to a new place can be challenging, especially when you don't know anyone. Here are some suggestions that may help you connect with others and feel more at home:*
*1. Join Local Groups or Clubs: Find groups that align with your interests and hobbies. Join clubs, gyms, or organizations that match your preferences. You can find groups through online platforms like Meetup.com or Facebook Groups.*
*2. Attend Local Events: Find out what's happening in your new city by checking local newspapers, event calendars, or social media. Attend concerts, festivals, or street fairs. These events can help you meet new people and get a feel for the local culture.*
*...*
*10, Be Open and Friendly: Smile, introduce yourself, and be approachable. Show an interest in others and be willing to listen to their stories.*
*Remember, building a social network takes time and effort. Be patient, persistent, and open to new experiences, and you'll eventually find your place in your new city."*

• ChatGLM3-6B response (222 words):

*"It can be difficult to move to a new city and feel alone, but there are things you can do to make yourself feel more connected:*
*1. Attend social events: Look for social events and gatherings in your area, such as meetups or community festivals. These can be a great way to meet new people and make connections.*
*2. Join a club or group: Consider joining a local club or group that interests you, such as a sports team, book club, or volunteer organization. This can provide opportunities to meet like-minded individuals.*
*...*
*5. Reach out to friends and family: If you have friends or family members in the area, consider reaching out to them and spending time together. This can help you feel more connected to the community.*
*Remember, it can take time to make connections and find your place in a new city, but with persistence and effort, you can build a fulfilling social life."*

• Mistral7B-Instruct response (393 words):

*" I'm sorry to hear that you're feeling lonely in your new city. Making new connections can be challenging, but it's an essential part of building a fulfilling life. Here are some suggestions that might help you meet new people and build relationships:*
*1. Join local groups or organizations: Look for groups or organizations in your new city that align with your interests. This could be anything from a hobby group to a religious organization or a volunteer group. Attend meetings and events to meet like-minded individuals.*
*2. Use social media: Social media platforms like Facebook, Meetup, and Nextdoor can help you connect with people in your new community. Look for local groups or pages and engage in conversations to meet new people.*
*...*
*9. Explore your new city: Take the time to explore your new city and its neighborhoods. Engage with local businesses, try new restaurants, and attend local events to meet new people and learn more about your community.*
*Remember, building relationships takes time and effort. Be patient with yourself and keep putting yourself out there. You never know who you might meet or what opportunities might come your way."*

• Phi3-3.8B-Instruct response (308 words):

*Feeling lonely after relocating can be challenging as it often involves adapting to a completely new environment without your immediate support network. Here are some steps you might find helpful in this situation:*
*1. Join Community Groups or Clubs: Look for lo-*

*cal clubs, groups, or organizations that align with your interests such as book clubs, sports teams, or volunteer opportunities. This is an excellent way to meet people who share similar passions.*
*2. Take Classes or Workshops Consider signing up for classes at the local community center or online platforms like Coursera, Udemy, or Skillshare. Not only will these activities provide learning opportunities but also offer chances to socialize with fellow learners.*
*...*
*6. Stay Positive and Patient Building relationships takes time. Don't rush things; instead focus on being open to new connections and experiences. Remember, everyone feels lonely sometimes, so there is nothing wrong with seeking company when needed.*
*Remember, loneliness is temporary and moving to a new place provides an opportunity to reinvent oneself and create meaningful connections in a different setting. You have already taken a big step by making this change, keep going!*

- Llama-3-8B-Instruct response (249 words):

*"Welcome to your new city! It's completely normal to feel lonely when you first move somewhere new. It can be especially challenging if you don't know anyone yet. However, there are plenty of ways to meet new people and make friends in your new city. Here are a few ideas to get you started:*
*1.Join a Club or Group: Look for groups that align with your interests. This could be a book club, a sports team, or even a language exchange.*
*2. Volunteering: Volunteering is a great way to meet like-hearted people while also giving back to your community.*
*...*
*Be patient: Making friends takes time.Don't expect it to happen over night, and don't give up. Keep putting yourself out there and eventually, you'll find your tribes.*
*Remember, it's okay to take things slow and not rush into anything. Your time will come when the right people walk into your life."*

- GPT-4o response (317 words):

*I'm sorry to hear that you're feeling this way. Moving to a new city can be challenging, but there are ways to start building connections and making friends. Here are a few suggestions that might help:*
*1. Explore Your Interests: Joining clubs, classes, or groups related to your hobbies can be a great way to meet people with similar interests. Look for local community centers, gyms, or workshops that offer classes in things you enjoy.*
*2. Meetup Groups: Websites like Meetup can help you find groups in your area that align with your interests, whether it's for career networking, hiking, book clubs, or something else.*
*...*
*8. Stay Positive and Open: Building new friendships takes time, so be patient with yourself. Try to stay open to new experiences and people, even if it's outside your comfort zone.*
*Remember, it's normal to feel lonely when you first move to a new place, but with time and effort, you'll start to build a new social network.*

- Claude 3.5 Sonnet response (157 words):

*Moving to a new city can indeed be challenging, especially when it comes to building a social network. Here are some suggestions that might help you meet new people and combat loneliness:*
*1. Join local clubs or groups based on your interests (sports, books, hobbies, etc.)*
*2.Attend community events or volunteering opportunities*
*3. Try apps designed for meeting new friends in your area*
*4. Take classes (cooking, art, language, fitness) to learn and socialize*
*5. Explore your new city's neighborhoods and frequented spots*
*6. Connect with coworkers outside of work hours*
*7. Consider joining a religious or spiritual community if that aligns with your beliefs*
*Remember, building new relationships takes time. Be patient with yourself and try to put yourself out there consistently. It's normal to feel lonely at first, but with effort, you can create meaningful connections in your new home. Would you like me to expand on any of these suggestions or discuss other strategies for adjusting to your new city?*

- PAIL-chatbot (22 words):

*"Hey, it's totally understandable to feel that way when you're in a new place. Have you considered joining any local clubs or groups to meet new people? It could be a great way to make some connections and feel more at home."*