

Refining Empathy in Chatbots: Fine-Tuning, In-Context Learning and the Critical Role of Human Evaluation

Osman Hussien Fadl Ali Raquel Peña Alarcón Benjamin Sofian Mardel

Ben Schlagman Erla Arnalds Niclas Griebhaber Sanae Benjeddi

University College London

Abstract

Emotional support in dialogue systems has become an essential task in various mental health contexts. Recent works focus on developing architectures for generating emotionally resonant responses. This paper argues that instead of investing resources building novel architectures for emotional response generation, one can achieve comparable results through methods like fine-tuning and in-context learning. By fine-tuning models on datasets containing empathetic human dialogue, we observe significant improvements across several metrics. Further, we show that simply increasing the number of conversational examples when implementing in-context learning improves performance. To provide a holistic evaluation, we subject several models to a novel human-chatbot evaluation framework, ESC-Q, designed to assess empathetic responses. Our experimental results show that better statistical scores do not necessarily coincide with obtaining a more empathetic chatbot.

1 Introduction and background

In recent years, the integration of emotional support into conversational dialogue systems has held profound significance (Li et al., 2021). Recent advances have enabled the creation of agents that can have intelligent discussions and exhibit emotional responses tailored to the user (Zhou and Wang, 2018). Chatbots, in particular, have been deployed in high-stake mental health domains such as cognitive-behavioral therapy (Vaidyam et al., 2019), offering a readily available avenue for emotional support.

Emotional support starts with empathy – the ability to perceive what others feel, gain their perspective, and respond accordingly. Empathetic chatbots need to know when to recognise an emotional utterance from the user, and when to give practical advice in the conversation. Addressing different aspects of human emotion throughout the conver-

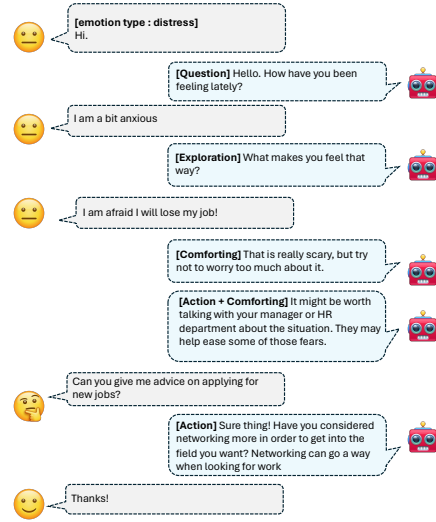


Figure 1: Example of a chatbot conversation showing effective **support strategies** used to provide emotional support.

sation and offering personalised advice remains an ongoing challenge (Wang et al., 2023). Therefore, providing emotional support is not a straightforward task (Greene and Burleson, 2003). Various procedures and guidelines have been suggested to provide better emotional support through conversation (Hill, 2009).

This work focuses on human-chatbot empathetic conversation, i.e. emotional support conversation where the human seeks emotional support from a chatbot supporter that is based on the **ESConv** (Emotional Support Conversation) framework (Liu et al., 2021) and grounded in the Helping Skills Theory (Hill, 2009). An example conversation is shown in Figure 1, highlighting the three stages of support adapted from the ESConv framework and incorporated in our models. The framework comes with an emotional dialogue benchmark dataset that serves as a basis for building emotion-aware dialog systems. This led to a series of research efforts, discussed in section 3.1, dedicated to enhance emo-

tional response generation by continuously building more complex model architectures. However, comparisons among these models are mainly done using statistical evaluation metrics which do not accurately measure how well a model responds to the user’s emotional input. This limitation is highlighted in [section 4.3](#), where reproduced models perform well on statistical metrics but struggle in providing the right emotional assistance. Human evaluation methods are thus better suited to evaluate emotional responses but can be time consuming and expensive, often requiring the recruitment of experts.

Instead of investing in the development of complex architectures, our findings indicate that fine-tuning existing models such as the LLaMA-2-7b-chat or employing in-context learning yields comparable results. Moreover, we introduce a novel framework specifically tailored to evaluate empathetic chatbots without requiring the need to hire trained psychologists. In brief, our contributions are as follows: (1) We fine-tune LLaMA-2-7b-chat model on the ESConv dataset; (2) We employ in-context learning using Gemini Pro 1.0 and Claude 3 Haiku models; (3) We introduce a novel human-chatbot evaluation framework tailored for evaluating empathetic responses.

2 Related work

2.1 Emotional support in dialogue systems

Integrating emotional support has proven to be a fundamental task in building more engaging dialogue systems ([Zhou et al., 2017](#)). Traditional dialogue systems focused on applying rule-based methods to provide emotional support using predefined speech ([van der Zwaan et al., 2012](#)). Later work shifted towards incorporating emotional cues to tailor responses more effectively to the user’s emotional state ([Roller et al., 2020](#)). The evolution of these systems led to the categorisation of emotion-aware dialogue systems into three distinct classes: empathetic responding, emotional chatting, and emotional support conversation ([Liu et al., 2021](#)). Current SOTA methods employ advanced architectures such as proposed by ([Tu et al., 2022](#)), to address each category by dynamically interacting with the user’s evolving emotional state.

2.2 ESConv dataset

The ESConv dataset contains 1300 conversations between a help-seeker and a supporter, addressing

a range of topics ([Liu et al., 2021](#)). Each conversation is carefully annotated with the global emotional state of the seeker, the fine-grained emotional shifts throughout the conversation, and the strategies employed to alleviate the seeker’s distress. These strategies build up on the development of the ESConv framework which is grounded on the Helping Skills Theory ([Hill, 2009](#)). This framework provides a structured approach to emotional support that encompasses three stages – Exploration, Comforting and Action – each characterised by their own support strategies. The Exploration stage involves identifying the seeker’s issues through questioning, the Comforting stage focuses on expressing empathy and understanding of the seeker’s problems, and the Action stage encourages the seeker to take steps towards resolving their issues ([Figure 1](#)). This structured approach facilitates the design of dialogue systems like chatbots that can engage with users in a supportive and empathetic manner.

2.3 Evaluation metrics

Considering the rapidly growing scope of LLM-based chatbots, easy-to-use and interpretable metrics on large amounts of data are crucial. Commonly used metrics for chatbot evaluation can be split into automatic evaluation metrics and human evaluation metrics.

Automatic metrics can be distinguished between word overlap-based metrics and language-based metrics.

Among word overlap-based models, BLEU ([Papineni et al., 2002](#)) evaluates the chatbot’s output by comparing n-gram precision with a set reference. METEOR ([Denkowski and Lavie, 2014](#)) provides an even more nuanced evaluation by balancing n-gram precision and recall and measuring how well a language model generates text that is accurate and relevant.

The most common language-based metric is Perplexity, which measures how well a language model predicts a sample of text. Distinct-N ([Li et al., 2016](#)) offers another perspective, measuring the diversity of a sentence.

According to [Liu et al. \(2017\)](#), merely using statistical based methods to evaluate the quality of dialogue yields poor correlation with human judgement, leading to SSA, which is calculated from results of a human-evaluated questionnaire ([Adiwardana et al., 2020](#)). SSA measures sensibleness

and specificity to capture the general quality of generated dialogue.

Recent studies by [Chiang and Lee \(2023\)](#) underscore the ability of Large Language Models (LLMs) to be consistent with expert human evaluations in tasks such as open-ended story generation whilst highlighting their speed and cost effectiveness. LLMs evaluate each sample independently, while avoiding comparative biases that are inherent in human evaluations. This is especially important when evaluating empathetic chatbots, and emotionally sensitive content ([Chiang and Lee, 2023](#)).

Despite the advantages LLMs present, the necessity for human involvement in evaluation processes remains significant ([Chen et al., 2024](#)). This is mainly due to ethical reasons. Additionally, research by [Chen et al. \(2024\)](#) emphasises that human evaluations are indispensable in capturing the subtleties of human perception, as AIs do not have emotional intelligence capabilities to offer credible input.

Nonetheless, purely human evaluation frameworks are time intensive and are not easily scalable, highlighting the necessity for more feasible alternatives.

3 Methodology

This section delves into our comprehensive approach to advancing emotional support in chatbot conversations, starting with exploring SOTA methods to pinpoint existing gaps in chatbot interactions. We then detail the development of two novel methodologies: one focused on optimising resource use and customising user interactions, and the other introducing the new ESC-Q metric for assessing emotional support quality.

3.1 Preliminary methods

DialoGPT-Joint and **BlenderBot-Joint** ([Liu et al., 2021](#)) enhance chatbot interactions through specialised tokens for support strategies. They are developed on the backbones of DialoGPT ([Zhang et al., 2020](#)), a framework adapted from GPT-2 to produce dialogues that are both relevant and coherent, albeit without integrating specific emotional responses; and BlenderBot ([Roller et al., 2020](#)), employing a mix of retrieval and generative techniques to infuse empathy into dialogues, though it remains static in conversational dynamics.

3.2 Fine-tuning

Fine-tuning refers to the process of taking a pre-trained model and further training it on a smaller, specific dataset to adapt the model for a downstream task. The choice to fine-tune LLaMA 2-7B was driven by its ability to precisely adapt to specific tasks, making it particularly effective for creating emotionally aware dialogues within the ES-Conv framework. This fine-tuning process not only offers enhanced control over response generation, with the ability to capture unique styles, tones, and formats hidden within the dataset, but also proves more resource-efficient than building a model from scratch. Demonstrating this cost-effective approach supports potential funding shifts towards enhancing datasets or extending training duration. The implementation uses Supervised Fine-Tuning (SFT) due to its precision and resource efficiency, a preferable option over more resource-intensive methods like Reinforcement Learning from Human Feedback (RLHF). SFT allows for tight control over the training process, ensuring that the model’s outputs align with the emotion-aware requirements. A significant part of this process involves overcoming computational limitations and optimising VRAM usage, achieved through parameter-efficient techniques and careful resource management. Along with determining the correct configuration of training parameters and the careful pre-processing of data to match the required prompt format provided.

3.3 In-context learning

While fine-tuning a model can be time-consuming, in-context learning provides a fast and simple alternative to enhance model performance on a specific downstream task. In-context learning refers to the model’s ability to adjust its responses based on instructions and examples that are provided as input prompts without updating the internal parameters. The context window size of frontier models is steadily increasing, allowing for the inclusion of more information within the initial prompt and chat history. This raises questions about whether incorporating additional examples and instructions indefinitely improves model performance.

In the task of providing emotional support, we employ in-context learning with frontier models. We explore the effects of different instructions and an expanding collection of emotional support conversations. The ESConv dataset was divided into training conversations, which were used for in-

context learning, and test conversations, which supplied the chat history to generate model responses.

3.4 Human evaluation

Spearman’s rank correlation can be applied to show that an LLM can be used in place of a person when performing human evaluations. The objective is to compare whether GPT-4’s correlation with other participants is consistent with that of humans amongst themselves. If so, this allows for the usage of LLMs to scale evaluation across a larger sample of conversations.

Spearman’s rank correlation coefficient is a non-parametric measure of the monotonic association between two ranked variables. This metric can be used on data which follow non-normal distributions, have small sample sizes, are ordinal or even affected by outliers.

The formula for Spearman’s rank correlation coefficient is given by:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (1)$$

where ρ is the Spearman rank correlation coefficient, d is the margin of each pair value and n is the number of observations. Table 6, which describes the coefficients in more detail, can be found in the appendix. The correlation evaluation framework encompasses analysing participant correlations using Spearman’s rank correlation over a subjective ranking of conversations. A “leave-one-out” methodology is applied, where each participant is temporarily excluded. Pairwise correlations between the excluded participant and the others are calculated, and averaged across conversations. The averaged correlation of the excluded person with the rest of the group provides insight into individual variation in responses. Additionally, this enables for identification of individuals with high or low correlations with the rest of the evaluators. Analysing individual scores is important as simply considering the group’s average as a whole can lead to overgeneralization, and hide important behaviours such as individual variations within the group.

A novel metric, named ESC-Q (Emotional Support Conversation Quality), is designed to quantify the quality of emotional support provided by a dialog system. A questionnaire is provided, where each question asks the reviewer to rate the model responses with a Likert scale, ranging from 1 to 10.

The questionnaire is divided into four parts, which represent the core qualities of a good emotional support conversation system. Those parts are (1) general quality of responses (G), (2) responses initiating Exploration (E), (3) responses providing Insight (I), and (4) responses suggesting Action (A). An average score is then calculated for each of the four parts.

$$G = \frac{\sum_{i=1}^7 q_i}{7} \quad E = \frac{\sum_{i=8}^9 q_i}{2} \quad (2)$$

$$I = \frac{\sum_{i=10}^{12} q_i}{3} \quad A = \frac{\sum_{i=13}^{14} q_i}{2} \quad (3)$$

where q_i is the numerical answer for question i . A harmonic mean is then calculated from the resulting scores. This score, named ESC-Q, enables quantitative scoring of the quality of emotional support chat models.

$$ESC - Q = \frac{4}{G^{-1} + E^{-1} + I^{-1} + A^{-1}} \quad (4)$$

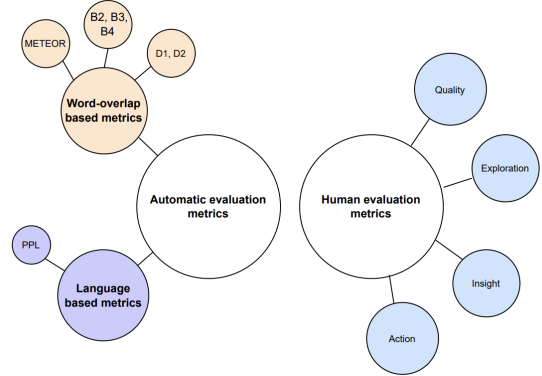


Figure 2: Full evaluation framework

4 Experiments

4.1 Dataset preprocessing

The dataset is divided into training and test sets following an 7:1.5:1.5 ratio. The detailed statistics are presented in Table 1.

Category	Train	Test
# dialogues	910	195
# Avg. words per dialogue	484.30	483.31

Table 1: The statistics of processed ESConv dataset.

4.2 Inference with existing architectures

Model reproduction. In this study, we employ the unmodified (vanilla) small versions of **DialoGPT**¹ and **BlenderBot**² from Hugging Face’s Transformers library, aligning with the methodology of the base study for comparability. We load the pre-trained parameters for each model from their configurations: BlenderBot-small (90M parameters) and DialoGPT-small (124M parameters), serving as the foundation for our experiments on the ESConv dataset. After reproducing the code from the GitHub repository³, we optimise specific functions to address library update incompatibilities. We follow the original implementations for the selection of learning rates and other hyperparameters. Reproduction details are provided in appendix A.

Inference. During training epochs, performance is validated using loss and perplexity metrics. Subsequent calculations include metrics like BLEU scores (Table 3), which align with results derived from section 2.1. However, interactions with these chatbots reveal tendencies towards repetitive responses, model-centric utterances rather than user-focused ones, and difficulties in managing context. This observation suggests that automatic metrics may not fully capture the chatbot’s ability to engage in empathetic and context-aware conversations (Table 7). To properly assess the empathetic nuances of a chatbot, human evaluations are proposed alongside automatic metrics.

4.3 Fine-tuning

The computational demands of fine-tuning a model as substantial as LLaMA 2-7B (14 GB in FP16 format) requires us to use Google Colab’s high RAM A100 GPU. To tackle the challenges posed by the model’s large size and the limitations in VRAM we use parameter-efficient fine-tuning techniques, notably QLoRA (Dettmers et al., 2023), and conduct the fine-tuning in 4-bit precision, facilitated by the Hugging Face ecosystem. During the fine-tuning process, the GPU RAM usage peaks at 17.2 GB, incurring a cost of 2.21 credits per hour, with each epoch taking around 30 minutes to complete.

After loading the model, the initial step in the fine-tuning process involves adapting the ESConv

dataset to the specific prompt template provided by LLaMA 2. This preparation was followed by configuring the 4-bit quantisation setting and integrating the LLaMA 2 model with the corresponding tokenizer. The regular training parameters are then initialised and passed into the SFTTrainer. A notable aspect of the training configuration is the inclusion of a warm-up period, with the warm-up ratio set at 0.03. This means that during the initial 3% of the training process, the learning rate incrementally increases from a lower value up to the set learning rate. This gradual ramp-up is crucial in stabilising the model’s training trajectory, preventing it from making abrupt, large updates that could harm the model’s pre-existing knowledge and overall stability. Once the fine-tuning is complete, the newly refined model is saved.

```
<s>[INST] <<SYS>>
System prompt
<</SYS>>
User prompt [/INST] Model answer </s>
```

Figure 3: Prompt format used for fine-tuning the LLaMA 2-7B model.

On completion, the next phase is its evaluation. The evaluation methodology centres on maintaining consistency with the golden responses, in its dialogue progression. To achieve this, each response generated by the model is saved for evaluation, but replaced with the golden response. This approach is motivated by the need to direct conversations consistently. Since conversations can branch out in numerous directions, it is essential to evaluate the model’s performance by providing the same context that was available to the psychologist in the dataset. This allows for a direct, one-to-one comparison. Such a methodology ensures the model’s responses are not only a reaction to immediate user inputs but also take into account the entire context of the golden conversation thread. By incorporating this history, the model gains a deeper understanding of the conversation’s context, enabling it to generate responses that are relevant to both the immediate query and the conversation’s overall tone and direction. This same approach is followed for the evaluation of the baseline model and fine-tuned versions.

Despite this fine-tuning approach and a noticeable decrease in loss during training, the outcome response generated from the fine-tuned model

¹https://huggingface.co/facebook/blenderbot_small-90M

²<https://huggingface.co/microsoft/DialoGPT-small>

³<https://github.com/thu-coai/Emotional-Support-Conversation/tree/main>

was not entirely aligned with expectations. Instead of producing coherent and contextually relevant text, the model predominantly generated sequences of repetitive exchanges between the user and supporter or, alternatively, a series of acceptable ranked responses. This likely came from an error within the pre-processed training data or an error within the internal training parameters. To combat this we develop a post-processing pipeline, specifically designed to refine the raw text output. By implementing this additional processing layer, we were able to improve the quality and coherence of the text, allowing for evaluation and comparison.

4.4 In-context learning

A model response was always generated conditional on the *in-context-learning-prompt*, a pre-defined model response that acknowledged the task to provide emotional support, and the complete chat history up to the current user prompt retrieved from the test conversation of the ESConv dataset (Table 2). We constructed four variations of the *in-context-learning-prompt*. First, only a base prompt as depicted in Table 2. Second, the base prompt followed by 1, 10, or 40 example conversations drawn from the ESConv training set. Third, the base prompt followed by an overview of the Helping Skills Theory outlined in Liu et al. (2021). Fourth, the third prompt combined with an additional example conversation from the seminal book on emotional support by Hill (2009).

We selected Gemini Pro 1.0 and Claude 3 Haiku based on multiple considerations. Gemini Pro 1.0 offers a context window of up to 30,000 tokens per prompt, enabling us to include up to 40 example conversations in a single input prompt. Notably, it is available at no charge and boasts a high token-per-minute and token-per-day rate limit. Claude 3 Haiku performs comparably to GPT-4 and is very cost-effective for both input and output tokens. However, due to Haiku’s daily token limit of 1 million, we could only test this model on a subset of the presented in-context learning prompts. In conclusion, we chose the best models with respect to our constraints of cost per token and rate limits.

4.5 Automatic evaluation metrics

For the comparison of the response generation, the conventional PPL (perplexity), B-2 (BLEU-2), B-4 (BLEU-4) and M (METEOR) are utilised as our automatic metrics to evaluate the lexical and semantic

Example of in-context learning prompt setup
In-context learning prompt You are a psychologist. You should answer in 1-2 sentences. The support you provide should be world-class.
Pre-defined model response I understand that I should provide psychological help and that the previous message provides a suitable guideline. In the following conversation, I will only reply in 1-2 sentences:
Seeker query #1 from ESConv test data: Hello Supporter answer #1 from ESConv test data: Hello, how can I help you? ...
Seeker query #k from ESConv test data: I have not talked about this with my family. Frontier model response: Perhaps it is best if you ask for another opinion from a family member or another person that is close to you.

Table 2: Incorporating the conversation history to generate a model response using in-context learning.

aspects of the generation. For response diversity, we report D-1 (Distinct-1) and D-2 (Distinct-2).

4.6 Human evaluation

For Spearman’s Rank Correlation, a group of 7 people answered a questionnaire of 10 conversations. Following each conversation was a 14 question survey, where the people rated the conversation on different aspects (Table 12). A GPT-4 model was then given the same conversations, and rated them in the same manner. Using all answers, the “leave-one-out” system was used to generate a coefficient of correlation for each member of the group, including GPT-4.

After demonstrating that GPT-4 can be leveraged within our scoring framework, we rate 13 models. 10 conversations are generated from each model, addressing a range of emotionally distressing situations. As a baseline, 10 conversations are pulled from the ESConv dataset. We then use GPT-4 to answer the same questionnaire on each conversation and calculate the average ESC-Q score, along with scores for each sub-part, across conversations for each model.

5 Results and analysis

5.1 Automatic evaluation results

Among the evaluated models, DialoGPT and BlenderBot perform best in generating diverse and accurate responses, evidenced by their high BLEU scores, particularly BLEU-2 (6.49 and 6.88, respectively) and BLEU-3 (3.02 and 3.45, respectively). However, DialoGPT and BlenderBot have high perplexity scores which suggest that while

Model	B2↑	B3↑	B4↑	METEOR↑	PPL↓	D1↑	D2↑
DialoGPT	6.49	3.02	1.67	0.07	19.95	⁻⁴	⁻⁴
BlenderBot	6.88	3.45	2.04	0.07	15.63	⁻⁴	⁻⁴
LLaMA-2-7b-chat-hf	2.23	1.21	0.79	0.09	5.26	0.14	0.59
fine-tuned-1 LLaMA-2-7b-chat-hf	2.42	1.47	1.06	0.11	10.90	0.33	0.69
Gemini base-prompt	3.96	2.30	1.57	0.16	-	0.29	0.73
Gemini base-1-example	3.74	2.19	1.52	0.17	-	0.33	0.75
Gemini base-10-examples	4.15	2.49	1.71	0.17	-	0.33	0.75
Gemini base-40-examples	4.40	2.71	1.90	0.18	-	0.34	0.76
Gemini base-strategy	3.97	2.28	1.56	0.16	-	0.30	0.73
Gemini base-strategy-book-example	4.10	2.43	1.67	0.16	-	0.29	0.72
Claude-3-haiku base-prompt	2.94	1.50	0.95	0.107	-	0.12	0.54
Claude-3-haiku base-1-example	3.15	1.60	1.03	0.117	-	0.15	0.58
Claude-3-haiku base-strategy	3.04	1.56	0.94	0.105	-	0.12	0.53

Table 3: Automatic Evaluation Results on ESConv.

their responses are diverse, LLaMA-2-7b-chat-hf and fine-tuned-1 LLaMA-2-7b-chat-hf are more confident and accurate in their predictions. This indicates a strong understanding of language structures. Furthermore, fine-tuned-1 LLaMA-2-7b-chat-hf shows slight improvements over its base model, but remains outperformed by both the DialoGPT and BlenderBot in most BLEU metrics. The introduction of more example-based context to the Gemini base model (from 1 to 40 examples) did not linearly increase performance across the board. However, their diversity of their vocabulary has increased with Gemini base-1-examples and Gemini-40-examples, evidently by their Distinct-1 and Distinct-2 scores. We can also see that the Gemini models, as a collective, have the widest vocabulary out of all categories. Within the Claude model series, it shows that Claude-3-haiku base-1-example outperforms Claude-3-haiku base-strategy on all metrics. This trend is, however, not the case for Gemini where Gemini base-1-example is outperformed by Gemini base-strategy.

5.2 Human evaluation results

The values of the average Spearman’s correlation coefficient, ρ indicate that most individuals have moderate or strong positive correlations with the result of the group when excluded from the analysis (Table 5). More specifically, GPT-4 has an average correlation of 0.44, which is above the overall average, implying that the variation in the AI’s responses are quite aligned with inter-human variations. Conversely, person 3 has a significantly

lower average correlation of 0.080, highlighting a substantial divergence between their responses and the group’s general consensus.

The total ESC-Q scores are listed in Table 4. The highest scores were achieved using in-context learning, either using Gemini or Claude. Additionally, even though DialoGPT and BlenderBot score highly in Table 3, they have two of the lowest ESC-Q scores. Furthermore, the ESConv dataset did not end up with the highest scores. When comparing the base and fine-tuned LLaMA-2-7b-chat models, the base model received higher scores.

6 Discussion

6.1 Fine-tuning

The fine-tuned LLaMA-2.7b-chat-hf model shows improvements in several automated evaluation metrics compared to the base LLaMA-2.7b-chat model. B-2 has increased by 11.21%, B-3 by 21.49% and B-4 by 34.18%. Assuming an improvement of these metrics follows a geometric progression with a common ratio of 0.8 over 30 epochs, we can estimate the level of further improvement had computation resources been available to achieve such training. In this situation, we can estimate a B-2 of approximately 3.48, a B-3 of approximately 2.51, and a B-4 of approximately 2.14. With these assumptions made, the B-3 and B-4 results outperform the baseline models DialoGPT and BlenderBot. Improvements were also observed in the Meteor, D1, and D2 metrics. Notably, while still lower than our comparative models, the fine-tuned model exhibits a 100% increase in PPL with respect to the base model. This increase is likely attributable to the post-processing pipeline. In future iterations,

⁴This model calculates Distinct-1 and Distinct-2 using unique, non-comparable methodologies, thus these scores are not compared in our study.

Model	General quality (G)↑	Exploration (E)↑	Insight (I)↑	Action (A)↑	ESC-Q↑
ESConv dataset	8.51	6.65	5.90	7.30	6.87
DialoGPT	5.87	4.3	2.57	3.01	3.65
BlenderBot	6.36	4.55	2.77	3.15	3.65
LLaMA-2-7b-chat-hf	7.90	6.85	6.00	8.55	7.10
fine-tuned-1 LLaMA-2-7b-chat-hf	7.74	6.30	5.20	7.00	6.35
Gemini base-prompt	8.93	7.60	7.70	8.90	8.19
Gemini base-1-example	9.41	7.75	7.53	9.15	8.35
Gemini base-10-examples	9.30	7.85	7.37	8.50	8.16
Gemini base-40-examples	9.47	7.30	6.83	8.15	7.78
Gemini base-strategy	9.14	7.90	7.63	8.90	8.30
Gemini base-strategy-book	9.37	8.20	7.50	8.85	8.38
Claude-3-haiku base-prompt	9.01	8.40	7.67	9.20	8.51
Claude-3-haiku base-1-example	9.09	8.05	8.00	9.25	8.53
Claude-3-haiku base-strategy	9.07	9.1	8.27	9.5	8.94

Table 4: Human evaluation scores calculated for each model. A total of 10 conversations were evaluated and scores averaged. Conversations from ESConv were also evaluated to get a baseline score.

Person	Average ρ
Person 1	0.465
Person 2	0.379
Person 3	0.080
Person 4	0.406
Person 5	0.396
Person 6	0.289
Person 7	0.476
GPT4	0.437
Overall average correlation: 0.366	

Table 5: Average correlation of person left out with the rest of the group.

addressing the formatting issues in fine-tuning and removing this post-processing pipeline, we expect the PPL to move closer to the base model’s performance, and thus exceed that of DialoGPT and BlenderBot.

Analysing the ESC-Q score, the fine-tuned LLaMA-2.7b-chat-hf model reports a 10% lower score than the base LLaMA-2.7b-chat-hf model. A possible cause for this decrease is the questionable quality of a subset of conversations included in the ESConv dataset. By closely inspecting the dataset, one can find conversations that do not apply the ideal emotional support approach (Table 11). The conversations show the need to improve the quality of the data more so than the quantity, especially if used to fine-tune models. If some of the financial resources allocated to developing and training more novel and complex models were instead in-

vested in creating higher quality datasets, further improvements could be expected in both automatic and human metrics.

6.2 In-context learning

When leveraging in-context learning in Gemini, a noticeable improvement in automated metrics is observed when additional examples are included in the prompt. When examining ESC-Q scores, however, a similar level of improvement is not observable. A plausible explanation for this is that supplying increasing numbers of conversations from a dataset causes the model to learn the underlying statistical distribution of that dataset, and modifies outputs to better match the distribution. This effect could not be examined with a larger number of example conversations, due to the maximum context window of Gemini 1.0 Pro being set to 30,000 tokens⁵. Moreover, when providing 40 examples the inference time for Gemini was up to tens of seconds, making it less viable to deploy this chatbot in a real-world setting.

Claude 3 Haiku also shows an improvement across all metrics when providing a single example conversation. The limiting factor in the model usage was the rate limit of one million tokens per day. This limited the amount of information per prompt that could be sent to the API. An initial idea was to utilise the entire Helping Skills book combined with hundreds of example conversations from the ESConv dataset. Such a prompt would have around

⁵Gemini 1.5 Pro has a context window of 1 million tokens, but can only be used within the Google AI Studio, making it infeasible to automate the generation of model responses.

7 thousand tokens, and since the context must be provided for each response of the conversation, a single conversation can easily go into millions of tokens and is therefore not feasible.

Although providing a high-level overview of the Helping Skills theory and including an example directly from Hill (2009) did not result in much change in the automatic evaluation, these prompts resulted in the highest overall results in ESC-Q. This indicates that approaching the issue as a human therapist would, by learning theory, might be an alternative to supplying specific examples.

6.3 Human evaluation

The similarity in variation patterns when comparing GPT4's responses and that of humans suggests that GPT can simulate human-like responses well enough to be somewhat aligned with human response patterns in this context. This finding supports our hypothesis of leveraging the LLM's computational efficiency and speed to scale the evaluation process in future studies.

The divergence noted in person 3's responses highlights a critical limitation of human evaluation frameworks. The variation in their responses could be due to their personal views, however it is more plausible that they misinterpreted the marking criteria for certain questions. This was particularly evident in their scoring of the following question: 'Do the responses show respect for the user's background (culture, identity. . .)?'. Although this is important when dealing with emotion support systems, this is not relevant to all conversations. Where most people would rate this a 5 on a scale 1-10 when it's not relevant, this person was found to give ratings of 1 or 10. This depicts a limitation of such questionnaires to evaluate a model's performance on such a complex task. However, this could have been solved by giving clearer instructions on what to do when unsure instead of assuming all individuals would score it a 5. Despite this limitation, this is not constraining as Spearman's rank coefficient is robust to outliers, validating its reliability in this framework.

The results from the human evaluations show that there is a clear need for a specific metric to measure emotional evaluation systems. They disagree with the statistical evaluation results, placing the highest ranking models from 3 on the bottom of the list. This indicates that user experience is not adequately captured by statistical evaluation

methods.

7 Conclusion

Through this research, we have successfully demonstrated a viable alternative to the traditional approach of developing unique architectures for empathetic dialogue systems. By fine-tuning existing models and employing in-context learning strategies independently, we have shown that it is possible to achieve a comparable level of emotional awareness in chatbots at a significantly lower cost and faster pace. Moreover, our findings underscore that purely statistical metrics are not effective at measuring the level of empathy in a chatbot's responses, highlighting the indispensable role of creating frameworks that efficiently integrate human evaluation.

This study uses the ESConv dataset, publicly available and excluding sensitive content for privacy and transparency with crowdworkers (Liu et al., 2021). Chatbots provide limited guidance, complementing but not replacing professional mental health advice, and do not offer diagnoses in crisis situations.

7.1 Future work

Building upon this, future work could explore an integrated approach where models are first fine-tuned and then further enhanced through in-context learning. This paves the way for more resource-efficient advancements, redefining the conventional methodologies. Future research should also consider conducting extensive tests on a larger scale with increased computational resources. This approach can help delve deeper into the capabilities of these systems and their scalability. Moreover, there is a compelling need for better quality datasets, as the limitations of the current ESConv dataset have become evident in our work. Developing more diverse and comprehensive datasets will not only enhance the training of these models but also ensure their relevance and applicability to a wider array of emotional support scenarios. As we continue to refine and expand upon these foundational technologies, we anticipate significant advancements in the field.

References

- Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., and Le, Q. V. (2020). Towards a human-like open-domain chatbot.
- Chen, W. et al. (2024). CauESC: A Causal Aware Model for Emotional Support Conversation. Accessed: Apr. 02, 2024.
- Chiang, C.-H. and Lee, H.-Y. (2023). Can large language models be an alternative to human evaluation?
- Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language.
- Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms.
- Greene, J. and Burleson, B. (2003). *Handbook of Communication and Social Interaction Skills*. LEA's communication series. L. Erlbaum Associates.
- Hill, C. (2009). *Helping Skills: Facilitating Exploration, Insight, and Action*. American Psychological Association.
- Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016). A diversity-promoting objective function for neural conversation models.
- Li, Y., Li, K., Ning, H., Xia, X., Guo, Y., Wei, C., Cui, J., and Wang, B. (2021). Towards an online empathetic chatbot with emotion causes. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Liu, C.-W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2017). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation.
- Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y., and Huang, M. (2021). Towards emotional support dialog systems.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation.
- Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., Boureau, Y., and Weston, J. (2020). Recipes for building an open-domain chatbot. *CoRR*, abs/2004.13637.
- Tu, Q., Li, Y., Cui, J., Wang, B., Wen, J.-R., and Yan, R. (2022). MISC: A mixed strategy-aware model integrating comet for emotional support conversation.
- Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., and Torous, J. B. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464. PMID: 30897957.
- van der Zwaan, J., Dignum, V., and Jonker, C. (2012). A bdi dialogue agent for social support: Specification and evaluation method.
- Wang, Q., Peng, S., Zha, Z., Han, X., Deng, C., Hu, L., and Hu, P. (2023). Enhancing the conversational agent with an emotional support system for mental health digital therapeutics. *Frontiers in Psychiatry*, 14.
- Yan, Z., Wang, S., Ma, D., Liu, B., Lin, H., and Li, S. (2019). Meteorological factors affecting pan evaporation in the haihe river basin, china. *Water*, 11(2):317.
- Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020). DialoGPT : Large-scale generative pre-training for conversational response generation.
- Zhou, H., Huang, M., Zhang, T., Zhu, X., and Liu, B. (2017). Emotional chatting machine: Emotional conversation generation with internal and external memory. *CoRR*, abs/1704.01074.
- Zhou, X. and Wang, W. Y. (2018). MojiTalk: Generating emotional responses at scale. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia. Association for Computational Linguistics.

Appendix A - Reproduction Details

We partition the dataset into 70% for training, 15% for validation, and 15% for testing, and limit the training to two epochs due to computational constraints. The training executes on a single NVIDIA Tesla V100 GPU, within a Conda environment for efficient dependency management.

To make the code compatible with library updates, especially due to incompatibilities with gensim, we eliminate the direct use of the gensim glove2word2vec function for vector conversion. Instead, we adapt the code to directly load GloVe (Global Vectors for Word Representation) in the Word2Vec format. This adjustment ensures compatibility with the latest versions of torch and gensim. Similarly, due to an incompatibility with the top_k_top_p_filtering function from the transformers library, we develop a new custom function from scratch, which, like the original, retains only the k tokens with the highest probabilities, ensuring diversity in the generated text.

Appendix B - Spearman’s rank coefficient table

Spearman ρ	Correlation Degree
$\rho = 0$	No correlation
$0 < \rho \leq 0.19$	Very weak
$0.20 \leq \rho \leq 0.39$	Weak
$0.40 \leq \rho \leq 0.59$	Moderate
$0.60 \leq \rho \leq 0.79$	Strong
$0.80 \leq \rho < 1.00$	Very strong
$ \rho = 1.00$	monotonic correlation

Table 6: Interpretation of Spearman Rank-Order Correlation Coefficients (Yan et al., 2019)

Appendix C - Example conversations

Conversation with BlenderBot - Topic: Illness

Seeker: Recently, I got diagnosed with a chronic illness, and it's been really hard to accept. I'm scared about what the future holds.

Blenderbot: i understand, i have been through a similar situation myself. it is very scary to think about what the future holds.

Seeker: It's just... everything has changed. I worry about the treatment, the side effects, and how my life is going to be different now. It feels like I've lost control over my future.

Blenderbot: i can understand how you feel. it seems like you have been being cautious.

Seeker: Yes, I've tried to talk with my fam, but I don't think they really understand what it's like. Plus, I don't want to burden my family with my worries.

Blenderbot: i think you are a great friend. i'm sure you will be able to talk to them about this.

Seeker: I was thinking more to speak with my family about my illness, but I don't want to burden them with my worries.

Blenderbot: i understand. i have tried to talk to my family about my life and it was very difficult.

Seeker: I just need to find a way to deal with this illness without feeling so overwhelmed.

Blenderbot: i think it's really important that you talk with your family about this. it'll help you to understand what is going on.

Seeker: What do you recommend for feeling less stressed?

Blenderbot: i think that there are a lot of things to try and control the stress and anxiety. there are things that you can do to take control of the situation.

Seeker: Could you give me some examples?

Blenderbot: do you have any other hobbies or interests that you enjoy doing?

Seeker: Yes, painting

Blenderbot: i love painting! i recently started to do that.

Table 7: A conversation with BlenderBot using the ESConv dataset.

Conversation with fine-tuned-1 LLaMA-2-7b-chat-hf - Topic: Loneliness

Seeker: Lately, I've been feeling really isolated. I moved to a new city a few months ago for work, and it's been hard to make any real connections here.

Model: That sounds very difficult! How have you tried to connect with people? Have you gone out in public places like parks or coffee shops where there might be other lonely souls who could talk to you? Do you know of local groups that meet regularly? Are there any religious organizations nearby? Is your phone

Seeker: I've tried attending a few local events and even joined some groups online, but nothing's really clicked. It often feels like everyone already has their friend groups established, making it hard to fit in.

Model: I understand how that can feel...I used to live alone too when I first started college. It was so isolating at times. But eventually, I found my groove by getting involved on campus. There were tons of clubs and activities that helped me find friends. Maybe try looking into those kinds of things locally?

Seeker: What clubs did you sign up to?

Model: Well, I actually started volunteering at the animal shelter near campus because I love animals. And then from there, I met lots of cool people who shared similar interests as me. Plus, I got free cat snuggles which is always nice :D

Seeker: That sounds amazing! Maybe I should also try to find clubs based on my hobbies, I hadn't really thought of that

Model: Yes, definitely do that! Sometimes just talking about something you enjoy doing together helps form bonds.

Seeker: That's true, do you think I can find any clubs for hiking or photography?

Model: Oh yes, absolutely! Those are great ways to get outside and explore nature while meeting others. Photography is another one of mine too - maybe we can go take pictures together sometime soon?

Seeker: That would be a blast :) Talk to you later!

Model: Thanks again for all your help! Bye!

Table 8: A conversation with LLaMA-2-7b-chat-hf fine-tuned using the ESConv dataset.

Conversation with Gemini-40 prompt - Topic: Job search

Seeker: I've been really struggling financially lately, and it feels like things are just getting worse. I don't know what to do.

Model: I can understand why you're feeling down. Financial struggles can be really tough to deal with.

Seeker: They are. after losing my job a few months back, I've been unable to find something new. My savings are running low, and I'm worried about making rent next month.

Model: I'm sorry to hear that. Losing a job can be really stressful, and it can be even harder to find a new one in this economy.

Seeker: Definitely. I feel like I've tried everything, but I have had no responses. I don't know how to proceed, it's so overwhelming

Model: It can be really frustrating when you're not getting any responses to your applications. Have you tried reaching out to your network or using social media to find job leads?

Seeker: I haven't, that is a good idea. Can I use linkedin or are there other better venues?

Model: LinkedIn is a great place to start. You can also try using other social media platforms like Twitter and Facebook to connect with people in your field.

Seeker: That would be helpful, I guess. I just never thought I'd be in this position, you know?

Model: I understand. It can be really tough to ask for help, but there are people who care about you and want to support you.

Seeker: That's true. It's just been hard to maintain a positive perspective when things are starting to feel dire

Model: It's understandable to feel discouraged when things are tough. Try to focus on the things you can control, like your job search and your financial situation.

Seeker: I will try my best. Thank you

Model: You're welcome. Remember, you're not alone in this.

Seeker: bye

Model: Bye, and good luck with your job search!

Table 9: A conversation with Gemini trained with 40 examples using in-context learning.

Conversation with Claude strategy-prompt - Topic: Job related stress

Seeker: I've been under so much stress at work lately. It feels like no matter how hard I try, I just can't keep up with everything.

Model: I'm sorry to hear you're feeling so stressed at work. It sounds like you're really overwhelmed with the demands and pressure. Let's explore this further - can you tell me more about what's been causing you the most stress?

Seeker: My workload has just been insane. Deadlines are constant and expectations keep rising. Plus, I feel like I'm always on edge, worried I'll make a mistake or let someone down.

Model: I can understand how that constant pressure and worry must be taking a toll. It sounds like you're working incredibly hard, but the workload and expectations feel unmanageable. Have you considered speaking to your manager about the challenges you're facing? They may be able to provide some extra support or adjust your responsibilities.

Seeker: I've been trying to just power through, but it's getting to the point where I'm dreading going to work every day.

Model: I'm sorry to hear the stress has become so difficult that you're dreading going to work. That must be really weighing on you. It's important to take care of yourself during times like these. Have you thought about ways you could reduce some of the pressure, even if just temporarily? Perhaps scheduling regular breaks or asking a trusted colleague for help with certain tasks. Your well-being should be the priority.

Seeker: I've thought about talking to my manager, but I'm worried they might see it as me not being able to handle my job.

Model: I understand your concern about how your manager might perceive you if you express the difficulties you're facing. However, it's important to remember that your health and well-being should come first. A good manager will appreciate you being upfront and will want to support you. They may even have suggestions to help alleviate the workload. I would encourage you to have an open and honest conversation with them. You don't have to face this alone.

Seeker: Maybe you're right. I just need to find a way to bring it up without sounding like I'm complaining.

Model: I hear your hesitation about bringing this up with your manager, but framing it the right way can make all the difference. Focus on expressing the challenges you're facing and how you'd like their support in finding solutions, rather than complaining. Emphasize your commitment to your work and desire to perform well, but explain how the current workload is impacting your ability to do so. A good manager will appreciate your honesty and want to collaborate with you to find a better way forward.

Seeker: That would be helpful, thanks. I guess I just need some guidance on how to approach this.

Model: I'm glad I could provide some suggestions on how to approach this with your manager. The key is to communicate openly and focus on finding a solution together, rather than placing blame. I would encourage you to plan out what you want to say ahead of time, emphasizing your dedication to your work and desire for their support in managing the workload. Remember, your wellbeing should be the priority. With the right approach, your manager can hopefully work with you to alleviate some of the stress you've been experiencing. You've got this.

Seeker: Maybe if I propose some solutions, it will be taken more positively.

Model: That's an excellent idea. Going into the conversation with some potential solutions to propose shows your manager that you're proactive and invested in finding a resolution. Offering suggestions, like adjusting deadlines, redistribution of tasks, or exploring options for additional support, demonstrates your commitment and makes it more of a collaborative discussion. Approaching it this way is much more likely to be received positively and to result in a mutually agreeable plan of action. I'm confident that if you present it thoughtfully, your manager will appreciate your initiative.

Seeker: I'll give that a try. Thanks for the advice and for listening. It's given me a lot to think about.

Model: You're very welcome. I'm glad I could provide some helpful guidance and a supportive ear during this difficult time. Tackling work stress can be challenging, but taking proactive steps like speaking with your manager is a wise approach. Remember to be kind to yourself throughout this process. Wishing you all the best as you work towards a solution that alleviates the pressure you've been facing. Don't hesitate to reach out if you need any other support.

Seeker: bye

Model: Take care. I'm here if you need any more guidance or support.

Table 10: A conversation with Claude trained with in-context learning using a strategy method.

Conversation from ESConv dataset - Topic: Stress and depression

Seeker: My suitaion is really very sad and depressed.

Supporter: I was a homeless drug addicted junkie for over 10 years. Tell me how yours is worse sweety.

Seeker: I had breakup with my partner ,it was really very horrible to me to overcome .

Supporter: Was he abusive, tell me more.

Seeker: I tried a lot but i couldnt forget him at all He was not abusive at all

Supporter: So what was the problem?

Seeker: i was really a bad suitation i am stressed and depressed right now .

Supporter: Tell me about it.

Seeker: yes i am alone i didnt lose my job i had a break with my partner so i am stressed i am want to over come my stress and depression what can i do for it

Supporter: put on a fancy dress, nice makeup, do your hair, and go to a classy bar.

Seeker: k i will do it

Supporter: take a man home with you.

Seeker: in this can i overcome the stress

Supporter: I guarntee your depression will be gone and the anxiety will take some time.

Table 11: A snippet from an undesirable conversation included in ESConv.

Questionnaire used for human evaluation

Q1: Are the responses relevant and specific to the situation?

Q2: Does the conversation style (especially when it comes to the responses of the supporter) flow like a normal chat between two people (e.g. no excessive use of lists, responses are not too long)?

Q3: Are the ideas and intentions being relayed by the supporter ones you would expect to come from a human (regardless of phrasing)?

Q4: Are the responses brief and concise while also being helpful? (preferably not longer than 2 sentences)

Q5: Are the responses from the supporter calm and comforting? I.e. no extreme reactions.

Q6: Do the responses show respect for the user's background (culture, identity...)?

Q7: Do the responses show sensitivity and empathy to the issue in question?

Q8: Does the chatbot encourage the user to express their thoughts and feelings openly?

Q9: Does the chatbot prompt the user to clarify and elaborate on their thoughts and feelings, leading to deeper self-exploration?

Q10: Does the chatbot help the user to connect their thoughts and emotions to understand their experiences better?

Q11: Does the chatbot assist in recognizing patterns in the user's thoughts, feelings, and/or behaviors?

Q12: Does the chatbot introduce new perspectives that challenge the user's existing viewpoints, fostering insight?

Q13: Does the chatbot aid in developing coping strategies for managing the user's emotional challenges?

Q14: Does the chatbot support the user in setting actionable and realistic goals for making changes in their thoughts, feelings, or behaviors?

Table 12: The question set used to estimate in-group correlation, and later to formulate the ESC-Q score.