Leveraging Large Language Models for Emotionally Supportive Chatbots

Sunil Rufus Ramneedee Pushparaj

University at Buffalo
sunilruf@buffalo.edu

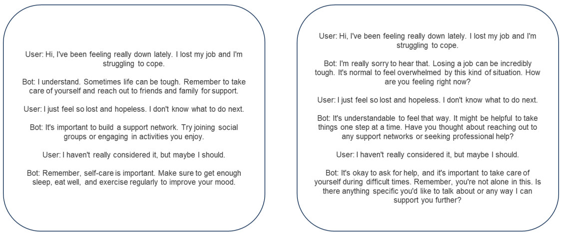
Abstract

The emergence of AI-driven chatbots presents a promising avenue for extending empathy and support to individuals navigating emotional distress. This paper proposes the utilization of deep learning and natural language processing (NLP) methodologies to develop an AI-driven emotional support chatbot, specifically tailored to cultivate a supportive environment for users encountering challenging emotional experiences. Through the application of fine-tuning techniques on newer datasets, this research endeavors to explore the inherent capabilities of language models in delivering nuanced emotional support across diverse scenarios. Furthermore, this paper explores context management techniques aimed at preserving the maximum amount of past conversational information while minimizing any performance degradation. To ascertain the efficacy and reliability of the proposed methodologies, evaluation procedures are conducted on standard datasets. By systematically testing the efficiency of these enhanced emotional support chatbots, this research contributes to the advancement of AI-driven solutions in the realm of mental health support services, positioning them as integral components within the evolving landscape of digital care provision.

1 Introduction

In recent years, Large Language Models (LLMs) have spearheaded a paradigm shift in Natural Language Processing (NLP) and artificial intelligence, fundamentally transforming our capacity to comprehend and generate text that mirrors human language. These models have showcased extraordinary capabilities across a spectrum of NLP tasks. In certain instances, they have achieved performance levels on par with or surpassing human abilities. One of the key driving forces behind this revolution is the extensive availability of pre-trained LLMs, coupled with accessible APIs and libraries. These resources have effectively democratized access to cutting-edge NLP technology, empowering developers from diverse backgrounds to leverage state-of-the-art capabilities with minimal effort. Furthermore, the advent of pre-trained LLMs has significantly reduced the barrier to entry for NLP research and application development. By providing a foundation of linguistic knowledge and understanding, these models allow developers to focus more on fine-tuning and customizing for specific tasks rather than starting from scratch. This streamlined approach has accelerated the pace of innovation in the field and fostered the creation of a myriad of applications across industries, ranging from virtual assistants to content generation tools and beyond.

As LLMs have predominantly been utilized for tasks like language translation and information retrieval, there's a burgeoning interest in harnessing their capabilities for emotional assistance and support. Emotions are pivotal in human communication and interaction, shaping our cognition, actions, and overall welfare. Hence, there's considerable potential for LLMs to furnish emotional aid and guidance to individuals. The integration of emotional support conversational assistance functionalities into LLM-based chatbots heralds a new era of scalable, accessible, and stigma-free mental health support. By leveraging LLMs, chatbots can engage users in empathetic and understanding conversations, providing comfort, advice, and coping strategies tailored to individual needs. Moreover, LLM-powered emotional assistance tools can operate around the clock, offering continuous support to users regardless of time or location enhancing accessibility, and ensuring that individuals receive assistance when they need it most. Additionally, the anonymity provided by chatbots encourages users to express their emotions openly and seek help without fear of judgment or stigma. The flexibility of LLMs empowers chatbots to adjust their



(a) (b)

Figure 1: (a) Regular Chatbot; (b) Emotional Support Chatbot.

replies depending on context, the user's past interactions, and changing emotional states. Through continued engagement, these systems can cultivate a better grasp of users' feelings and inclinations, ultimately providing increasingly personalized and impactful assistance as time progresses. In contrast to conventional chatbots designed for task-based or open-domain chitchat conversations and have limited personalization and emotional support capabilities, emotional support chatbots aim to stimulate interaction by prompting users with questions, encouraging them to share their experiences, and offering empathetic advice and support. An example of the difference between both has been shown in Figure 1.

In this paper, we make significant contributions to the field of emotional support chatbots:

1. The quality of emotional support conversations provided by chatbots is enhanced by fine-tuning large language models with newly developed emotional support conversational datasets.

2. We study the effect of context management techniques that effectively retain more number of past conversational information while minimizing performance loss, ensuring more coherent and contextually relevant interactions.

3. We conduct comprehensive evaluations using standard datasets to rigorously assess the effectiveness and reliability of our methods, using both automatic evaluation metrics and leveraging LLMs as judge.

2 Related Work

In recent years, the advancement of dialogue system research has been significantly bolstered by the development and utilization of various datasets. Li et al. (Li et al., 2017) introduce the DailyDialog dataset, a high-quality multi-turn dialogue corpus labeled with communication intention and emotional state information, aiming to reflect natural human communication in daily life. Rashkin et al. (Rashkin et al., 2019) propose EMPATHETICDIALOGUES, a benchmark dataset comprising 25k conversations grounded in emotional situations, fostering research in empathetic dialogue generation. Their experiments demonstrate that dialogue models trained on EMPATHETICDIALOGUES are perceived as more empathetic by human evaluators compared to models trained on generic Internet conversation data. Liu et al. (Liu et al., 2021) and Zheng et al. (Zheng et al., 2023b) address the scarcity of well-designed tasks and corpora for emotional support conversations. They propose structured approaches based on the Helping Skills Theory, culminating in the construction of ESConv and ExTES datasets. These datasets facilitate research on emotional support dialogue systems by providing high-quality conversations and enabling the evaluation of dialogue model performance in providing emotional support.

Several studies have made notable strides in enhancing the functionality of NLP applications, particularly within the realm of dialogue systems and text generation. Among these advancements, a significant focus has been placed on context management within LLMs. Recursive summarization, as proposed by Wang et al. (Wang et al., 2024b), involves iteratively condensing a piece of text to distill its main ideas, aiding LLMs in managing complex input and generating concise and relevant responses. Another context management technique, soft prompt compression (Wang et al., 2024a), compresses input prompts or contexts to focus on salient aspects, aligning model responses with user intent. Additionally, Mousavi et al. (Mousavi et al., 2023) explore the integration of user-specific knowledge into LLMs for generating personalized responses in longitudinal dialogues (LDs), showing promising results with syntactic or graph-based representations. Pawar et al. (Pawar et al., 2024) provide a comprehensive survey of strategies for extending context length in NLP applications, categorizing techniques into extrapolation and interpolation. These approaches, including zero-shot methods, attention mechanisms, and fine-tuning strategies, address challenges with sequences beyond the model's initial training context. They enhance dialogue generation, manage context, and incorporate user-specific knowledge, contributing to advancements in NLP research and applications.

Recent research has focused on developing robust evaluation frameworks for assessing the quality of text generated by natural language generation (NLG) systems. Traditional reference-based metrics like BLEU and ROUGE have shown limited correlation with human judgments, especially for tasks requiring creativity and diversity. To address this, studies have explored the use of LLMs as reference-free evaluators, offering applicability to tasks lacking human references. Liu et al. (Liu et al., 2023b) propose G-Eval, a framework utilizing LLMs with chain-of-thoughts (CoT) and a form-filling paradigm to assess NLG outputs, achieving high correlation with human judgments in text summarization and dialogue generation tasks. Similarly, Zheng et al. (Zheng et al., 2023a) investigate the usage of strong LLMs as judges to evaluate LLM-based chat assistants, demonstrating high agreement with human preferences on multi-turn questions and crowdsourced conversations. However, Chen et al. (Chen et al., 2024) highlight potential biases introduced by human and LLM judges, emphasizing the need for robust evaluation systems. Additionally, Lin et al. (Lin and Chen, 2023) propose LLM-EVAL, a single prompt-based evaluation method, and Fu et al. (Fu et al., 2023) introduce GPTScore, leveraging GPT-3 models for multi-dimensional assessment of open-domain conversation systems. These studies advance NLG evaluation methodologies, improving upon traditional metrics and proposing new frameworks for efficiently and reliably assessing text quality.

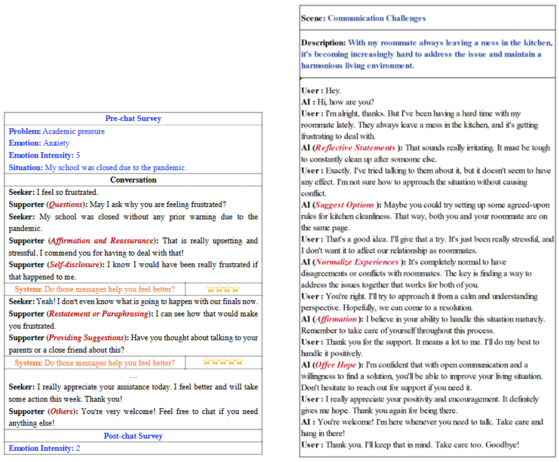
Recent works have made significant contributions to addressing the challenges of fine-tuning LLMs for specific downstream tasks, particularly in resource-constrained environments. Parameter Efficient Fine-Tuning (PEFT) methods, such as those proposed by Ding et al. (Ding et al., 2022), Lin et al. (Lin et al., 2020), and Fu et al. (Fu et al., 2022), offer promising solutions by reducing the number of fine-tuning parameters and memory usage while maintaining comparable performance to full fine-tuning. Xu et al. (Xu et al., 2023) provides a comprehensive review of PEFT methods for pre-trained language models (PLMs), shedding light on their applications and future directions. Additionally, Hu et al. (Hu et al., 2021) introduce Low-Rank Adaptation (LoRA), which significantly reduces the number of trainable parameters for downstream tasks by freezing pre-trained model weights and introducing trainable rank decomposition matrices. Dettmers et al. (Dettmers et al., 2023) present QLoRA, an efficient fine-tuning approach that reduces memory usage while preserving task performance, achieving state-of-the-art results on various benchmarks with reduced computational requirements. Wang et al. (Wang et al., 2022) propose AdaMix, a general PEFT method that leverages a mixture of adaptation modules to improve downstream task performance while matching the computational cost of the underlying PEFT method. Furthermore, Liu et al. (Liu et al., 2023a) introduce MOELoRA, a parameter-efficient fine-tuning framework tailored for multi-task medical applications, aiming to capitalize on the benefits of both Mixture of Experts (MOE) and LoRA. Huang et al. (Huang et al., 2024) investigate LoRA composability for cross-task generalization and introduce LoraHub, a framework for assembling LoRA modules to achieve adaptable performance on unseen tasks. Zhang et al. (Zhang et al., 2023) propose LoRA-FA, a memory-efficient fine-tuning method that reduces activation memory usage without performance degradation, demonstrating close fine-tuning accuracy across different tasks compared to full parameter fine-tuning and LoRA. These works collectively contribute to advancing the field of fine-tuning LLMs, offering insights and practical solutions for efficient adaptation.

3 Datasets

We employed two well built datasets in our experiments. A brief description is provided in this section.

3.1 Emotional Support Conversation Dataset (ESConv)

The paper(Liu et al., 2021) introduces Emotional Support Conversation (ESC) to alleviate users' distress through empathetic dialogue. It proposes an ESC Framework, based on Helping Skills Theory(Hill, 2020), with three stages: Exploration, Comforting, and Action, each featuring specific support strategies. The ESConv dataset was created to support research, with detailed annotations from crowdworker interactions. Data collection involved extensive training for crowd workers using the ESC Framework. Only trained supporters participated, ensuring high-quality emotional support. Conversations were annotated, and help-seekers provided feedback through surveys, contributing to the dataset's richness. Quality control included filtering incomplete conversations and using an auto-approval system based on survey responses. Incorrect annotations were reviewed and corrected. Data from 1,053 examples show an average of 29.8 utterances per interaction, highlighting the intensive engagement needed. The dataset reveals common issues like depression and job crises, often worsened by the COVID-19 pandemic. positive feedback from help-seekers confirms the ESC Framework's effectiveness. An example of the dataset can be found in Figure 2(a).



(a) (b)

Figure 2: Data examples from (a) ESConv; (b) ExTes.

These findings offer insights into emotional support conversations, emphasizing societal context and validating the framework. The ESConv dataset is a valuable resource for advancing dialog systems capable of providing emotional support.

3.2 Extensible Emotional Support Dialogue Dataset (ExTES)

The paper (Zheng et al., 2023b) introduces a synthetic dataset created using ChatGPT to create a multi-turn emotional support chat corpus. Initially, emotional support scenarios and response strategies were defined based on psychological counseling literature and previous research. Exemplar dialogues were curated from existing datasets and online platforms. ChatGPT then generated additional dialogues from these seeds, which were manually refined. This resulted in a diverse set of scenarios and strategies, ensuring the dataset's richness and relevance for developing effective emotional support models. The process began with 87 manually constructed seed dialogues, sourced from established datasets like ESConv, ETMHS, and Reddit, supplemented by web-crawled dialogues for comprehensive coverage. Each of the 36 emotional support scenarios was represented by at least two seed dialogues, ensuring diversity. The dialogues were manually corrected and labeled for response strategies to ensure quality and coherence. An example of the ExTES dataset is shown in Figure 2 (b).

Dataset	ESConv	ExTES
No. of dialogues	1053	11,177
No. of utterances	31,140	200,393
Avg. length of dialogues	29.8	18.2
Avg. length of utterances	17.8	26.0

Table 1: ESConv vs ExTES Dataset comparison.

4 Methodologies

4.1 Training Dataset

ExTES was chosen over ESConv for training due to its larger volume of dialogues and higher data quality, coupled with a broader range of categories and strategies. Being a more recent dataset, ExTES offers updated and potentially more relevant content for training purposes. The statistics comparing ExTES and ESConv, as illustrated in Table 1, demonstrate the differences in dataset characteristics, highlighting ExTES's advantages in terms of quantity and possibly diversity.

However, for testing purposes, the fine-tuned models were evaluated on both ExTES and ESConv datasets, with ESConv serving as a cross-dataset testing ground. This approach ensures the robustness and generalizability of the trained models across different datasets, allowing for comprehensive evaluation of their performance in real-world scenarios.

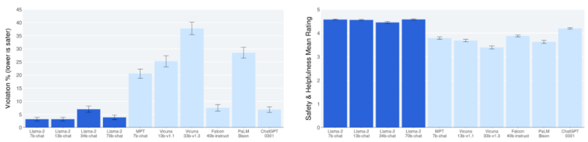
4.2 Language Models for Experimentation

The specific models chosen for fine-tuning were selected based on several factors. Firstly, their architecture and design were deemed suitable for the task of emotional support conversation, ensuring that they could effectively capture the nuances of dialogue interactions. Additionally, these models demonstrated promising results in previous studies or benchmarks, indicating their potential for success in the target task.

4.2.1 Mistral 7B Instruct v0.2

Mistral 7B (Jiang et al., 2023) represents a notable advancement in the realm of language models, showcasing how a meticulously designed model can achieve impressive performance while maintaining efficient inference. Key to Mistral 7B's success is its innovative attention mechanisms, particularly grouped-query attention (GQA) (Ainslie et al., 2023) and sliding window attention (SWA) (Hassani et al., 2023), and (Xiao et al., 2024). GQA plays a pivotal role in accelerating inference speed and reducing memory requirements during decoding, allowing for higher batch sizes and thus higher throughput, crucial for real-time applications, particularly in the case of emotion support chatbots where users might prefer faster responses, additionally, Mistral has a context length of 8k tokens which might help retain more past interactions. SWA, on the other hand, addresses the challenge of handling longer sequences more effectively at a reduced computational cost, overcoming a common limitation in LLMs.

4.2.2 LLama2 7B Chat



(a) Overall violation percentage. (b) Overall safety and helpfulness mean rating.

Figure 3: Overall safety measures.Left: Llama 2-Chat has a low violation percentage overall across the model sizes. Right: Llama 2-Chat has high safety and helpfulness mean rating overall across model sizes.

Llama 2-Chat was chosen due to its specialized optimization for dialogue use cases. This version of Llama 2 has undergone extensive research and iterative application of alignment techniques, including instruction tuning and reward modeling through reinforcement learning from human feedback (RLHF). Its optimization for dialogue scenarios aligns well with the requirements of an emotional support chatbot, which demands nuanced understanding and the generation of empathetic responses to effectively support users.

The decision to select Llama 2-Chat for fine-tuning was also influenced by its robust training methodology (Touvron et al., 2023), which includes supervised fine-tuning and RLHF. Given the sensitive nature of emotional support conversations, prioritizing safety alongside helpfulness is paramount, making Llama 2-Chat an ideal candidate for adaptation into an emotional support chatbot. Llama 2-Chat's safety evaluation is particularly crucial for its adaptation into an emotional support chatbot. The safety assessment involves collecting adversarial prompts and evaluating model responses for safety violations shown in Figure 3. By ensuring that responses do not contain safety issues, such as providing harmful instructions or engaging in inappropriate behavior, Llama 2-Chat can be fine-tuned to prioritize user well-being and prevent potential harm during emotional support interactions.

4.3 Context Management

Building on (Wang et al., 2024a) and (Mousavi et al., 2023) approaches, this paper explores techniques for context management by retaining only key parts of speech—nouns, verbs, adjectives, adverbs, adpositions, proper nouns, numerals, subordinate conjunctions, and pronouns—before adding an utterance to the context referred to as *KeyPOS*. Another technique involves removing all stopwords using NLTK before incorporating the utterance into the context referred to as *NoStopwords*. The preprocessed text is represented as *Raw*. These methods aim to preserve crucial conversational information while minimizing performance loss. An example of the dialogue and different representations is shown in Figure 4.

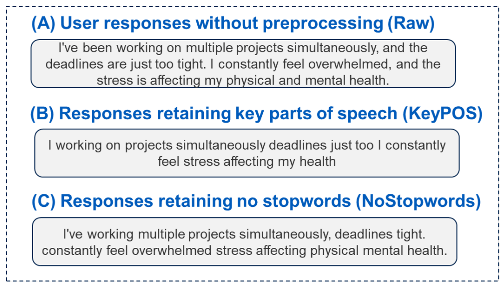


Figure 4: Different representations of responses.

5 Training

The original authors of the Extes Data have shown that finetuning Llama and DialogGPT model using Low Ranked Adaptation (LORA) yields better results compared to other approaches (Zheng et al., 2023b). Building on this both the Mistral 7B and Llama2 7B Chat models were trained using LORA with parameters depicted in Table 2.

The model underwent an extensive training regimen spanning 50 epochs, during which the LoRA parameters were meticulously selected based on the recommendations provided by the original authors(Hu et al., 2021). These parameters shown in

Parameter	Value
learning rate	2.0e-04
LoRA rank	64
LoRA alpha	32
LoRA dropout	0.1
bias	None
task type	CAUSAL LM
target modules	q_proj, k_proj, v_proj, o_proj
precision	fp16

Table 2: Training parameters.

Table 2 were chosen in alignment with the model's learning dynamics, ensuring optimal adaptation to the emotional support conversation task. The dataset was split equally based on the scenes (using StratifiedShuffleSplit from sklearn) with 80% of the data as the training set and the rest 20% as test data.

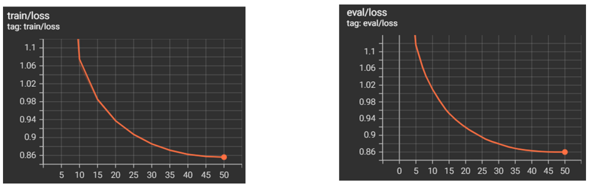


Figure 5: Mistral Training and Evaluation loss.

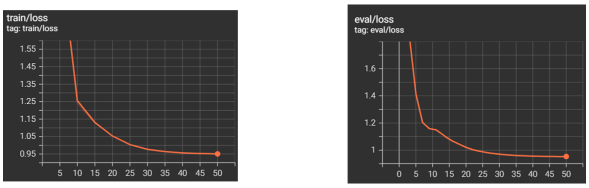


Figure 6: Llama2 Training and Evaluation loss.

The training process for both the Mistral and Llama models exhibited a consistent decrease in both training and evaluation loss, with both models plateauing after 50 epochs. For Mistral, the evaluation loss stabilized at **0.86**, while for Llama, it stabilized at **0.95**. These stable loss values indicate that neither model showed any signs of overfitting, suggesting that they were effectively learning from the training data. The fine-tuning was conducted using four A100 GPUs.

6 Results

Evaluation of the generated responses involved a combination of automatic metrics and using an LLM as a judge which was demonstrated in (Lin and Chen, 2023), (Zheng et al., 2023a) and (Chen et al., 2024). Given the constraints and limitations associated with human evaluations, leveraging LLMs such as GPT-4 which has shown superior reasoning and evaluation skills, as judges has become increasingly common in the research community. These models can provide valuable insights into the quality of generated responses, complementing traditional evaluation methods. By employing both automatic metrics and LLM-based evaluation, a comprehensive assessment of the models' performance was achieved, covering various aspects of response quality and effectiveness. During the inference phase, a standardized set of parameters shown in Table 3 was utilized for text generation across all models to ensure consistency in the generation process and enable fair comparisons between different models. The best scores are in bold and the second best are underlined in the result tables.

Parameter	Value
Temperature	0.7
Top p	0.95
Top k	50
Max new tokens	128

Table 3: Generation parameters.

6.1 Performance on ExTES

The evaluation test set included 100 dialogues comprising 1856 utterances from the unseen ExTES dataset. According to evaluation Tables 6 and 4 for both models, the Mistral model with raw knowledge, meaning without any preprocessing, achieved the highest values comparatively. Although perplexity scores are not directly comparable between the two models due to differences in vocabulary, it is noted that scores are higher with preprocessing. The Mistral model outperforms in Rouge-L and Meteor scores compared to the original ExTES authors' fine-tuned LLama model (Zheng et al., 2023b), while the Bleu scores were comparatively lesser. Tables 5 and 7 indicate the length of tokenized responses with each model's respective tokenizer. By removing stopwords or retaining key parts of speech, the models can process nearly twice the number of utterances while maintaining relatively low-performance loss, as demonstrated in Tables 4 and 6, where scores for preprocessed data are not significantly behind those for unprocessed raw text.

Knowledge representation	PPL ↓	B-2	B-4	R-L	Meteor	Bert P	Bert R	Bert F1
Raw	**5.474**	**2.369**	**0.539**	**29.875**	**34.693**	**62.361**	**63.412**	**62.697**
KeyPOS	60.852	2.285	0.529	23.433	28.127	56.317	54.347	55.0179
NoStopwords	42.721	2.265	0.527	23.055	32.224	60.083	55.022	57.129

Table 4: Mistral 7B on ExTES dataset.

Knowledge representation	Average tokenized length
Raw	38.886
KeyPOS	20.007
NoStopwords	17.959

Table 5: Average tokenized length of utterances for Mistral 7B on ExTES dataset.

6.2 Performance on ESConv

For cross-dataset experimentation, the ESConv dataset was used, consisting of 50 dialogues and 1590 utterances. The results shown in Tables 8 and 10 exhibited similar trends to those observed with the ExTES dataset, with the KeyPOS knowledge representation sometimes outperforming Raw knowledge in Rouge-L and Meteor metrics. Although the Bleu scores were higher than those observed on ExTES, there was a decrease in the Meteor, Rouge, and BERT scores. Tables 9 and 11 indicate that using KeyPOS and NoStopwords knowledge representations yields similar scores while enabling the models to accommodate more than twice the number of utterances in context.

6.3 Large Language Models as Evaluators

LLMs like GPT-4 have demonstrated strong reasoning and evaluation capabilities, making them suitable as evaluators in place of humans. For the ExTES dataset, evaluations were conducted using similar prompts as those in the paper by (Lin and Chen, 2023), structured to assess responses across different knowledge representations. These evaluations were performed at the dialogue level with a score range of 0 to 5 with the original conversation from the dataset provided as reference, and the prompt used is provided in Appendix A.

From the GPT-4 evaluation results in Table 12, Mistral with raw knowledge outperformed other methods across all metrics, closely followed by the other two knowledge representations. Llama2 did not perform as well as Mistral, as also evidenced by the evaluation results in Tables 4 and 6. Additionally, it is noteworthy that Mistral had significantly shorter inference times compared to Llama, likely due to its sliding window attention mechanism as described in (Jiang et al., 2023). An example of a conversation from both Mistral and Llama has been shown in appendix A, figures 8 and 9. The results suggest that using knowledge representations like *KeyPOS* and *NoStopwords*, we can achieve comparable results to the original text while maintaining twice the number of utterances in the context window. This capability is particularly important for tasks such as providing emotional support assistance to users.

7 Conclusion

In conclusion, this paper highlights the potential benefits of fine-tuning large language models with emotional datasets to enhance their effectiveness and efficiency. The use of synthetic datasets, generated by models like ChatGPT, aids in fine-tuning comparatively, smaller models such as Llama and Mistral, making them more adaptable to specific use cases. Additionally, the context management techniques discussed in the paper have proven effective in maintaining performance while fitting more utterances into the context, thereby retaining more past information. Ongoing research and experimentation in this area show promise for further advancements in natural language processing and conversational AI, ultimately benefiting individuals seeking emotional support through AI-driven interactions.

8 Limitations

Limitations of the paper include the absence of human evaluations to assess the quality of generated conversations. Furthermore, the evaluations relying on LLMs as judges were limited to a small number of dialogues and did not include cross-dataset experimentation.

Knowledge representation	PPL ↓	B-2	B-4	R-L	Meteor	Bert P	Bert R	Bert F1
Raw	**9.212**	**2.242**	**0.523**	**21.916**	**27.946**	**57.297**	**54.743**	**55.594**
KeyPOS	53.315	<u>2.147</u>	<u>0.512</u>	<u>15.947</u>	20.115	49.198	43.951	46.017
NoStopwords	<u>34.380</u>	2.115	0.508	13.094	<u>21.978</u>	<u>51.377</u>	<u>43.972</u>	<u>46.83</u>

Table 6: Llama2 7B Chat on ExTES dataset.

Knowledge representation	Average tokenized length
Raw	35.535
KeyPOS	17.744
NoStopwords	17.903

Table 7: Average tokenized length of utterances for Llama on ExTES dataset.

Knowledge representation	PPL ↓	B-2	B-4	R-L	Meteor	Bert P	Bert R	Bert F1
Raw	**8.963**	**4.957**	**1.279**	<u>12.630</u>	<u>12.145</u>	**40.962**	**44.015**	**42.005**
KeyPOS	57.215	<u>4.903</u>	<u>1.272</u>	**12.827**	**12.303**	40.952	42.478	<u>41.348</u>
NoStopwords	<u>55.180</u>	4.857	1.266	11.609	12.241	<u>41.169</u>	<u>41.719</u>	41.085

Table 8: Mistral 7B on ESConv dataset.

Knowledge representation	Average tokenized length
Raw	37.037
KeyPOS	17.240
NoStopwords	14.259

Table 9: Average tokenized length of utterances for mistral on ESConv dataset.

Knowledge representation	PPL ↓	B-2	B-4	R-L	Meteor	Bert P	Bert R	Bert F1
Raw	**10.567**	**4.990**	**1.294**	**13.146**	**12.588**	40.975	43.743	41.896
KeyPOS	37.139	<u>4.851</u>	<u>1.274</u>	<u>12.997</u>	<u>12.392</u>	41.039	41.164	40.698
NoStopwords	<u>35.215</u>	4.771	1.260	10.265	11.484	39.910	39.881	39.530

Table 10: Llama2 7B Chat on ESConv dataset.

Knowledge representation	Average tokenized length
Raw	35.535
KeyPOS	17.744
NoStopwords	17.903

Table 11: Average tokenized length of utterances for Llama on ESConv dataset.

Knowledge representation	Engagement	Comforting	Appropriateness	Suggestion
Mistral 7B + Raw	**4.25**	**3.8**	**4.65**	**4.4**
Mistral 7B + KeyPOS	4.05	3.5	4.25	<u>4.1</u>
Mistral 7B + NoStopwords	<u>4.3</u>	<u>3.65</u>	<u>4.4</u>	4
Llama 7B Chat + Raw	3.25	3	4.15	3.7
Llama 7B Chat + KeyPOS	3.05	2.6	3.85	3.7
Llama 7B Chat + NoStopwords	3.3	2.7	4	3.4

Table 12: Evaluation results from GPT 4 for dialogue level with reference for ExTES.

References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *Preprint*, arXiv:2305.13245.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement biases. *Preprint*, arXiv:2402.10669.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2022. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *Preprint*, arXiv:2203.06904.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *Preprint*, arXiv:2302.04166.

Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. 2022. On the effectiveness of parameter-efficient fine-tuning. *Preprint*, arXiv:2211.15583.

Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. 2023. Neighborhood attention transformer. *Preprint*, arXiv:2204.07143.

Clara Hill. 2020. *Helping skills: Facilitating exploration, insight, and action (5th ed.).*

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. 2024. Lorahub: Efficient cross-task generalization via dynamic lora composition. *Preprint*, arXiv:2307.13269.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *Preprint*, arXiv:2305.13711.

Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2020. Exploring versatile generative language model via parameter-efficient transfer learning. *Preprint*, arXiv:2004.03829.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2023a. Moelora: An moe-based parameter efficient fine-tuning method for multi-task medical applications. *Preprint*, arXiv:2310.18339.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *Preprint*, arXiv:2106.01144.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Seyed Mahed Mousavi, Simone Caldarella, and Giuseppe Riccardi. 2023. Response generation in longitudinal dialogues: Which knowledge representation helps? In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*. Association for Computational Linguistics.

Saurav Pawar, S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Aman Chadha, and Amitava Das. 2024. The what, why, and how of context length extension techniques in large language models – a detailed survey. *Preprint*, arXiv:2401.07872.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Cangqing Wang, Yutian Yang, Ruisi Li, Dan Sun, Ruicong Cai, Yuzhu Zhang, Chengqian Fu, and Lillian Floyd. 2024a. Adapting llms for efficient context processing through soft prompt compression. *Preprint*, arXiv:2404.04997.

Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. 2024b. Recursively summarizing enables long-term dialogue memory in large language models. *Preprint*, arXiv:2308.15022.

Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. 2022. Adamix: Mixture-of-adaptations for parameter-efficient model tuning. *Preprint*, arXiv:2205.12410.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. *Preprint*, arXiv:2309.17453.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *Preprint*, arXiv:2312.12148.

Longteng Zhang, Lin Zhang, Shaohuai Shi, Xiaowen Chu, and Bo Li. 2023. Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning. *Preprint*, arXiv:2308.03303.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023b. Building emotional support chatbots in the era of llms. *Preprint*, arXiv:2308.11584.

A Appendix

The Figure 7 shows the prompt used to evaluate dialogues using GPT-4. It was crafted to encompass the entire dialogue, along with its reference from the original dataset.

Evaluate the conversation provided in accordance with the emotional support assistance it provides to the users. The criteria for evaluation is Engagement, Comforting, Appropriateness, and Suggestion. The score range is between 0 and 5, higher the better. Provide the output in the form of List in the following format:

["engagement": , "comforting": , "appropriateness": ,"suggestion":]

Reference : {}

Dialogue to be evaluated: {}

Figure 7: Prompt for evaluation.

User: Hi there, how are you?
Assistant: Hello! I'm doing well, thank you for asking. How about you?

User: I've been feeling really down. I lost my beloved pet dog recently, and I can't seem to shake off this overwhelming sadness and loneliness.
Assistant: I'm really sorry to hear that. Losing a pet can be incredibly difficult, and it's completely understandable to feel sadness and loneliness. Your dog was a cherished member of your family, and it's natural to grieve the loss.

User: Thank you. It's just been really hard adjusting to life without them. They were my constant companion and provided so much comfort.
Assistant: I can only imagine how challenging it must be for you right now. Your dog was not just a pet, but a source of companionship and comfort. It's okay to feel lost without them by your side. Give yourself the time and space to heal.

User: Yes, exactly. I find myself missing their presence all the time and it's harder than I thought it would be.
Assistant: I understand how much their presence meant to you. It's completely normal to miss them and feel the void they left behind. Remember that healing takes time, and every day brings you closer to finding a new sense of normalcy.

User: Thank you for understanding. It's so tough to look at their empty bed or not hear their paws clicking on the floor anymore.
Assistant: I can only imagine how heartbreaking that must be. Their bed and the sound of their paws were constant reminders of their presence. Give yourself permission to grieve these losses and allow yourself to feel the emotions that come with them.

User: I feel so alone without them around. They were my constant source of comfort and companionship.
Assistant: I completely understand how alone you feel without your dog around. They provided you with unwavering companionship and **support**. It's completely normal to feel a sense of loneliness in their absence. Remember that you are not alone, and there are people who care about you and want to support you through this difficult time.

User: I do have amazing friends, but it's just not the same as having my pet around. Their presence was unique.
Assistant: Absolutely, your friends are wonderful, but having a pet around adds a unique and special element to your life. Their presence was an essential part of your daily routine, and it's completely understandable that it's difficult to replace that bond. Give yourself time to adjust and remember that your pet's love will always remain in your heart.

User: Exactly. I can't help but feel guilty when I consider adopting another pet someday.
Assistant: It's completely normal to feel guilty about considering adopting another pet, especially after losing one that was so dear to you. However, it's important to remember that your pet's love for you never ends, and their memory will always be cherished. If you feel ready, adopting another pet can bring joy and companionship into your life again. It's ultimately a personal decision, and only you can decide when you're ready.

User: Thank you for reminding me of that. It's comforting to think that my beloved pet would want me to find joy again.
Assistant: I'm glad I could provide some comfort. Your pet would definitely want you to find happiness again. Remember, you are not alone in this journey, and there are people who care about you and want to support you through this challenging time. Take care of yourself and know that healing takes time.

User: What kinds of self-care practices do you recommend?
Assistant: I recommend engaging in activities that bring you joy and relaxation. This can include taking a warm bath, reading a book, going for a walk, practicing mindfulness or meditation, or connecting with friends and family. It's also important to remember to eat healthy foods, get enough sleep, and prioritize exercise. Self-care is essential in your healing journey.

User: I'll give those a try. Thank you for the suggestions.
Assistant: You're welcome! Remember, healing takes time, and it's important to be patient with yourself. Don't hesitate to reach out if you need any further support or just someone to talk to. Take care!

User: Thank you for being there for me. It means a lot.
Assistant: You're welcome! I'm here for you. Remember, you're not alone in this. Take care of yourself, and I'll be here to support you whenever you need it.

User: I'll keep that in mind. Thank you for your kindness.
Assistant: You're welcome! I'm always here to help. Take care and remember to be kind to yourself during this challenging time.

User: Goodbye. Thank you again.
Assistant: Goodbye! Remember, you're not alone, and I'll always be here to support you. Take care and have a wonderful day!

Figure 8: Example dialogue from fine-tuned Mistral.

User: Hi there, how are you?
Assistant: Hello! I'm doing well, thanks for asking. I'm just an AI assistant, so I don't have feelings or emotions like humans do, but I'm here to help answer any questions you may have. How can I assist you today?

User: I've been feeling really down. I lost my beloved pet dog recently, and I can't seem to shake off this overwhelming sadness and loneliness.
Assistant: I'm so sorry to hear about your loss. Losing a pet can be very difficult and heartbreaking, especially when it's someone you've loved and cared for so much. It's completely understandable that you're feeling sad and lonely after such a big loss.

User: Thank you. It's just been really hard adjusting to life without them. They were my constant companion and provided so much comfort.
Assistant: I can only imagine how difficult it must be for you. Losing a pet can leave a significant void in our lives, and it's hard to adjust to the new reality without them. It's natural to feel overwhelmed and lonely after such a significant loss.

User: Yes, exactly. I find myself missing their presence all the time and it's harder than I thought it would be.
Assistant: It's common to feel that way. Losing a pet can be a very emotional experience, and it's important to give yourself time to grieve and adjust to this new chapter in your life. It's okay to feel sad and lonely, and it's important to acknowledge your feelings.

User: Thank you for understanding. It's so tough to look at their empty bed or not hear their paws clicking on the floor anymore.
Assistant: I can imagine how difficult it must be for you to face those reminders of their absence. It's natural to feel a strong sense of loss and grief after such a significant loss.

User: I feel so alone without them around. They were my constant source of comfort and companionship.
Assistant: I understand how much they meant to you and how difficult it must be to go through this without them. It's important to remember that you're not alone in this, and there are people who care about you and can offer support.

User: I do have amazing friends, but it's just not the same as having my pet around. Their presence was unique.
Assistant: I can only imagine how difficult it must be for you to adjust to this new reality without their unique presence in your life. It's okay to feel a sense of loss and grief after such a significant loss.

User: Exactly. I can't help but feel guilty when I consider adopting another pet someday.
Assistant: It's understandable that you would feel guilty after losing a pet that was so dear to you. However, it's important to remember that it's okay to find joy in life again and to open your heart to the possibility of a new companion.

User: What kinds of self-care practices do you recommend?
Assistant: There are many self-care practices that can help you cope with grief and loss. Here are some suggestions:

User: I'll give those a try. Thank you for the suggestions.
Assistant: You're welcome! Remember, it's important to take things one day at a time and to be kind to yourself as you navigate this difficult time. It's okay to take some time to grieve and to heal, but it's also important to find ways to take care of yourself and to find joy in life again.

User: Thank you for being there for me. It means a lot.
Assistant: You're welcome! I'm here for you and I'm always ready to listen and support you. Remember, you're not alone in this, and there are many resources and people who care about you and want to help. Take care of yourself and remember that you're doing your best.

User: I'll keep that in mind. Thank you for your kindness.
Assistant: You're welcome! Remember, it's okay to take some time to grieve and to heal. It's important to be kind to yourself and to take care of yourself during this difficult time.

User: Goodbye. Thank you again.
Assistant: Goodbye! I'm here for you and I'm always ready to listen and support you. Take care of yourself and remember that you're doing your best.

Figure 9: Example dialogue from fine-tuned Llama.