Empowering Emotional Support Chatbots with Large Language Models

by

Sunil Rufus Ramneedee Pushparaj

May 9th 2024

A dissertation submitted to the

Faculty of the Graduate School of

the University at Buffalo, The State University of New York

in partial fulfilment of the requirements for the

degree of

Computer Science and Engineering Master of Science in Computer Science and Engineering

Department of Computer Science and Engineering

To my family and friends

Acknowledgments

I want to express my sincere appreciation to all those who have supported and guided me during the course of this research project.

Firstly, I am incredibly thankful to my advisor, Dr. Nalini Ratha, whose invaluable guidance, consistent support, and constructive feedback have been instrumental in shaping and completing this research thesis.

I am also deeply grateful to Dr. Rohini Srihari, a member of my thesis committee, for her insightful suggestions on advancing this research further.

I would like to acknowledge my colleagues in the laboratory for their ongoing assistance and support whenever I needed it.

A heartfelt thank you goes to the Department of Computer Science and Engineering at the University at Buffalo for providing outstanding facilities and resources, which greatly facilitated my research efforts.

Lastly, I wish to convey my heartfelt gratitude to my family and friends for their steadfast support and understanding throughout my academic journey. Their unwavering patience and encouragement have served as my most powerful driving force.

Table of Contents

List of Tables

List of Figures

Abstract

The emergence of AI-driven chatbots presents a promising avenue for extending empathy and support to individuals navigating emotional distress. This study 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, this chatbot aims to leverage Large Language Models with significantly fewer parameters than contemporary state-of-the-art models, such as ChatGPT. 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. Central to its objectives is the refinement of emotional support chatbots by means of fine-tuning existing language models on datasets curated for emotional comprehension. Furthermore, this study undertakes an investigation into knowledge pruning techniques, with the goal of reducing the size and complexity of trainable parameters within these models while ensuring the preservation of their performance metrics. 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.

Chapter 1

Introduction

1.1 The Impact of Large Language Models

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, including but not limited to language translation, text generation, sentiment analysis, and question answering. 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 pretrained LLMs, coupled with accessible APIs (Application Programming Interfaces) 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 pretrained 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 ap-

proach 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.

1.2 Emotional Assistance: A new Frontier for LLMs

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. This innovative approach holds promise in addressing the escalating demand for emotional well-being services. 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. This 24/7 availability enhances accessibility and ensures 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.

Moreover, the flexibility of LLMs empowers chatbots to adjust their 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 chit chat conversa-

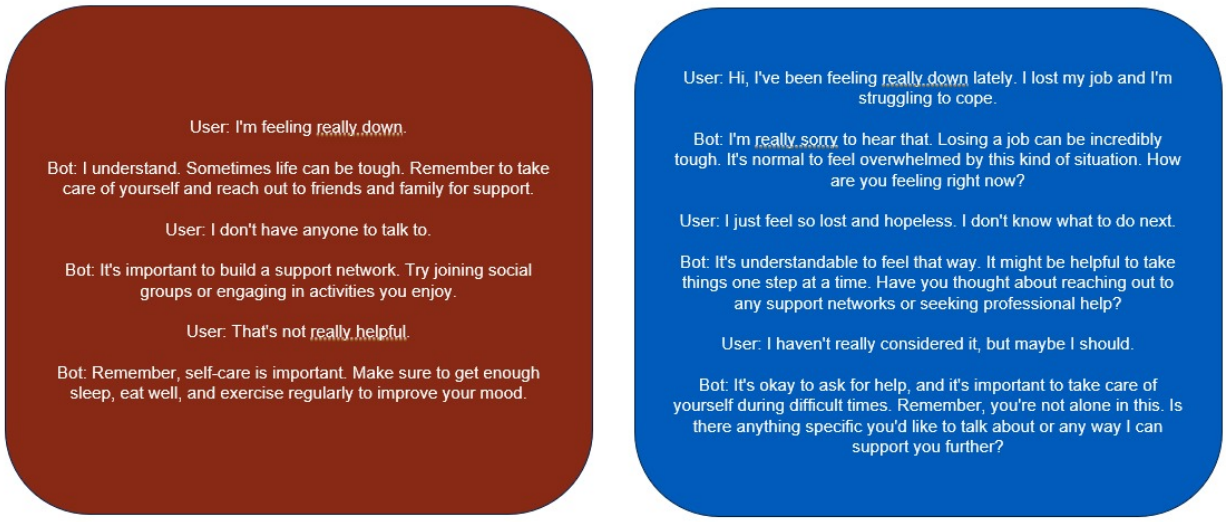


Figure 1.1: Regular chatbot vs Emotional Support Chatbot

tions and have limited personalization and emotion support capabilities, whereas, 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 1.1.

1.3 Leveraging Smaller LLMs for Efficient Inference

In the realm of NLP and AI, Large Language Models (LLMs) such as GPT have garnered significant attention for their impressive capabilities. However, deploying these models for real-time inference poses challenges due to their computational demands, latency concerns, and associated costs.

To address these challenges, there's a growing interest in leveraging smaller LLMs and fine-tuning them effectively to achieve efficient inference. This involves a careful balance between performance and resource constraints. By employing advanced techniques in model architecture design, data preprocessing, and optimization strategies, we can harness the power of these models to provide inferences that are both effective and resource-efficient.

One key aspect of this approach is the careful selection and adaptation of model archi-

tectures. Smaller LLMs, while lacking the vast parameter space of their larger counterparts, can still be fine-tuned to excel in specific tasks. By carefully curating and preprocessing the training data, we can ensure that the model learns efficiently from the available information, leading to faster convergence and improved performance.

Furthermore, optimization strategies such as quantization, pruning, and efficient hardware utilization play a crucial role in enhancing inference efficiency. By quantizing model parameters to lower precision, pruning unnecessary connections, and leveraging specialized hardware accelerators, we can significantly reduce inference time and computational requirements while maintaining acceptable levels of accuracy.

1.4 Enhancing Emotional Support with a Humanoid robot

We have deployed this application through two distinct channels: a web-based chat interface and integration with a humanoid robot. The web-based chat interface offers accessibility and convenience, allowing users to engage with the emotional support bot from any device with internet access. Through this platform, individuals can seek support, share their feelings, and receive empathetic responses in real-time, fostering a sense of connection and reassurance.

Simultaneously, by integrating the emotional support bot with a humanoid robot, we aim to enhance the user experience through physical presence and interaction. This approach adds a layer of embodiment to the support provided, enabling the robot to engage with users in a more immersive and impactful manner. Whether in healthcare settings, educational environments, or personal spaces, the humanoid robot serves as a tangible companion, offering comfort, encouragement, and companionship to those in need. By deploying the application across these two channels, we strive to maximize accessibility and effectiveness, ensuring that individuals can receive the support they need, wherever they are, and however they choose to engage with the technology.

Chapter 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. [1] introduce the DailyDialog dataset, a high-quality multi-turn dialogue corpus labeled with communication intention and emotion information, aiming to reflect natural human communication in daily life. Rashkin et al. [2] 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. [3] and Zheng et al. [4] 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. [5], 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 [6], compresses input prompts or contexts to focus on salient aspects, aligning model responses with user intent. Additionally, Mousavi et al. [7] 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. [8] provide a comprehensive survey of strategies for extending context length in NLP applications, categorizing techniques into extrapolation and interpolation. These approaches, encompassing zero-shot methods, attention mechanisms, and fine-tuning strategies, address challenges associated with handling sequences beyond the model's initial training context. Collectively, these works offer insights into enhancing dialogue generation, managing context, and incorporating user-specific knowledge, contributing to the broader landscape of 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. [9] 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. [10] 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. [11] highlight potential biases introduced by human and LLM judges, emphasizing the need for robust evaluation systems. Additionally, Lin et al. [12] propose LLM-EVAL, a single prompt-based evaluation method, and Fu et al. [13] introduce GPTScore, leveraging GPT-3 models for multi-dimensional assessment of open-domain conversation systems. These studies collec-

tively contribute to advancing NLG evaluation methodologies, addressing the limitations of traditional metrics, and proposing novel frameworks for assessing text quality efficiently and reliably.

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. [14], Lin et al. [15], and Fu et al. [16], 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. [17] 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. [18] introduce Low-Rank Adaptation (LoRA), which significantly reduces the number of trainable parameters for downstream tasks by freezing pretrained model weights and introducing trainable rank decomposition matrices. Dettmers et al. [19] 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. [20] 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. [21] 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. [22] 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. [23] 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.

Chapter 3

Datasets

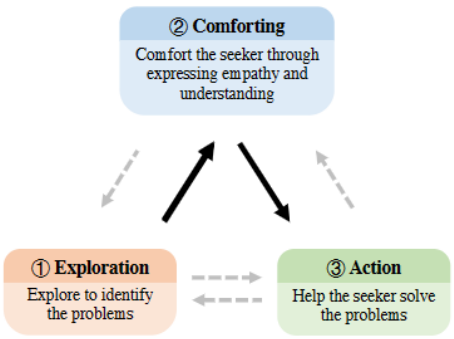
3.1 Emotional Support Conversation Dataset (ESConv)

3.1.1 Framework

The authors define the task of Emotional Support Conversation (ESC) , with the goal of reducing users' emotional distress and helping them navigate challenges through empathetic interactions. They propose an ESC Framework 3.1, grounded in the Helping Skills Theory[24], tailored for dialog systems. The framework consists of three stages (Exploration, Comforting, and Action) with various support strategies. To facilitate research in an emotional support conversation, they construct the ESConv dataset, ensuring rich annotation and quality examples through crowdworker interactions. Training tutorials based on the ESC framework are provided to supporters, and multiple mechanisms are employed to ensure the effectiveness of emotional support. Evaluation of state-of-the-art models reveals significant improvement in emotional support when utilizing various strategies, with the Joint model mimicking human supporter behaviors effectively. This work aims to advance data-driven approaches in building dialog systems capable of providing effective emotional support.

3.1.2 Data Collection

The data collection process for the ESConv dataset involved meticulous planning and implementation to ensure the acquisition of high-quality conversation examples. At the outset, a comprehensive tutorial based on the ESC Framework was developed to train crowdworkers



Strategies	Stages			Examples	Lexical Features
Question				*Can you talk more about your feelings at that time?*	do you (15.0), are you (13.8), how (13.7), what (12.3), do (11.5)
Restatement or Paraphrasing				*It sounds that you feel like everyone is ignoring you. Is it correct?*	is that (8.2), so you (8.2), it sounds (7.1), correct (7.1), so (6.6)
Reflection of Feelings				*I understand how anxious you are.*	can tell (7.4), understand how (5.8), are feeling (5.1), tell (5.1), understand (4.9)
Self-disclosure				*I feel the same way! I also don't know what to say to strangers.*	my (15.3), was (10.5), me (10.2), had (9.7), myself (7.8)
Affirmation and Reassurance				*You've done your best and I believe you will get it!*	its (5.7), thats (5.6), will (5.4), through this (5.1), you will (4.7)
Providing Suggestions				*Deep breaths can help people calm down. Could you try to take a few deep breaths?*	maybe (7.3), if (6.5), have you (6.4), talk to (5.8), suggest (5.8)
Information				*Apparently, lots of research has found that getting enough sleep before an exam can help students perform better.*	there are (4.4), will (3.8), available (3.7), seen (3.3), possible (3.3)
Others				*I am glad to help you!*	welcome (9.6), hope (9.6), glad (7.3), thank (7.0), hope you (6.9)

Figure 3.1: An Overview of ESC Framework

in providing effective emotional support. This framework served as a guiding principle for supporters, outlining various strategies and stages of emotional support.

Only crowdworkers who successfully passed an examination based on the tutorial were admitted to the task, ensuring that supporters were adequately trained and equipped with the necessary skills. These trained supporters then engaged in conversations with help-seekers, adhering to the ESC Framework. Throughout these interactions, support strategies were annotated to structure the resulting dataset effectively, providing valuable insights into the dynamics of emotional support conversations.

After each conversation, supporters rated the level of detail provided by the seeker about their problems, contributing to rich annotations and enhancing the dataset's quality. Meanwhile, help-seekers completed pre-chat surveys, offering valuable information on their problems, emotions, and the origin of their situation.

Feedback from help-seekers was collected during conversations, enabling the assessment of the helpfulness of supporter messages on a 5-star scale. Additionally, post-chat surveys were conducted to evaluate the help-seeker's emotion intensity after the conversation and the performance of the supporter, providing further insights into the effectiveness of the support provided.

To maintain data quality, preliminary filtering mechanisms were employed to remove in-

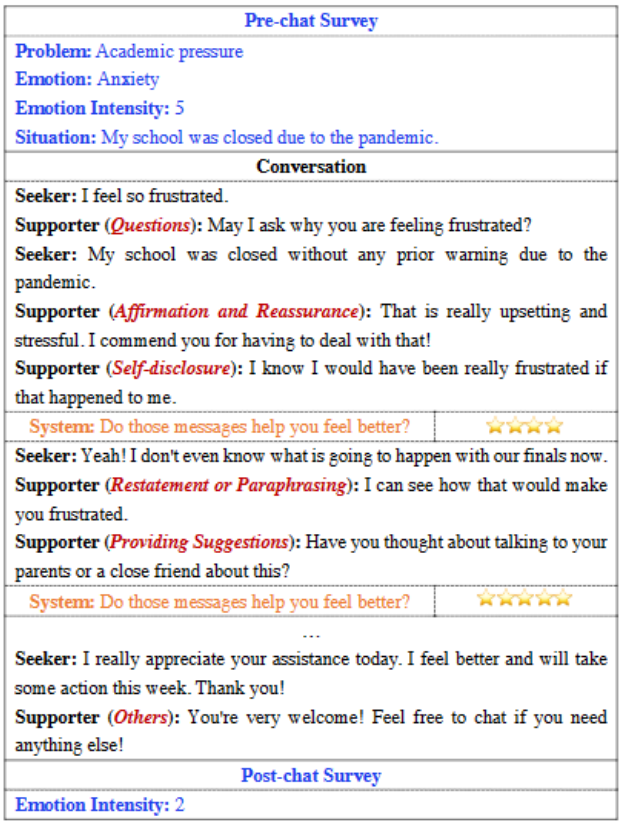


Figure 3.2: Data example from ESConv

complete or short conversations. An auto-approval program was also meticulously designed, utilizing criteria based on post-chat survey responses and utterance length to ensure only high-quality conversations were included in the dataset. An example of the dataset can be found in figure 3.2.

Furthermore, annotation correction processes were implemented to review and revise incorrect annotations of support strategy and seeker's emotion intensity, further enhancing data accuracy and reliability. Overall, the data collection process for the ESConv dataset was a comprehensive and rigorous undertaking aimed at providing researchers with valuable insights into emotional support skills in dialog systems.

Category	Total	Supporter	Seeker
# dialogues	1,053	-	-
Avg. Minutes per Chat	22.6	-	-
# Workers	854	425	532
# Utterances	31,410	14,855	16,555
Avg. length of dialogues	29.8	14.1	15.7
Avg. length of utterances	17.8	20.2	15.7

Figure 3.3: Statistics of ESConv

3.1.3 Data Statistics

The statistics for the 1,053 ESConv examples 3.3 reveal the substantial engagement required in effective emotional support conversations, with an average of 29.8 utterances per interaction, surpassing previous datasets' typical lengths. Analysis of figure 3.4 further highlights prevalent societal challenges, with ongoing depression and job crises being the most common problems expressed by help-seekers, accompanied by emotions such as depression and anxiety, likely exacerbated by factors like the COVID-19 outbreak. The positive feedback received from help-seekers underscores the efficacy of the ESC Framework-based training, indicating supporters' success in providing satisfactory emotional support.

These insights offer researchers valuable groundwork for understanding the dynamics of emotional support conversations within the ESConv dataset. The prevalence of specific problems and emotions sheds light on the societal context influencing help-seekers, while the positive feedback validates the effectiveness of the ESC Framework in guiding supportive interactions. Overall, these statistics provide a foundation for further exploration and research in the domain of emotional support in dialog systems.

	Categories	Num	Proportion
Seeker's Problem	Ongoing Depression	306	29.1%
	Job Crisis	233	22.1%
	Breakup with Partner	216	20.5%
	Problems with Friends	159	15.1%
	Academic Pressure	139	13.2%
	Overall	1,053	100.0%
Seeker's Emotion	Anxiety	281	26.7%
	Depression	276	26.2%
	Sadness	250	23.7%
	Anger	96	9.1%
	Fear	88	8.4%
	Disgust	32	3.0%
	Shame	30	2.8%
	Overall	1,053	100.0%
Seeker's Feedback	1 (Very Bad)	71	1.1%
	2 (Bad)	183	2.9%
	3 (Average)	960	15.5%
	4 (Good)	1,855	29.9%
	5 (Excellent)	3,144	50.6%
	Overall	6,213	100.0%
Support Strategy	Question	3,109	20.9%
	Restatement or Paraphrasing	883	5.9%
	Reflection of Feelings	1,156	7.8%
	Self-disclosure	1,396	9.4%
	Affirmation and Reassurance	2,388	16.1%
	Providing Suggestions	2,323	15.6%
	Information	904	6.1%
	Others	2,696	18.1%
	Overall	14,855	100.0%

Figure 3.4: Statistics of all the annotations, including the help-seekers' problems, emotions, feedback, and the support strategies

3.2 Extensible Emotional Support Dialogue Dataset (ExTES)

3.2.1 Dataset Collection

The dataset collection process heavily relied on constructing a sophisticated multi-turn emotional support chat corpus utilizing ChatGPT (gpt-3.5-turbo)3.7. Initially, comprehensive

Category	Dialogues	Proportion	Category	Dialogues	Proportion
Breakups or Divorce	710	6.3%	Navigating Gender Identity and Transitioning	202	1.8%
Conflicts or Communication Problems	1,109	9.9%	Moving to a New City or Country	202	1.8%
Communication Challenges	1,008	9.0%	Career Transitions	202	1.8%
Coping with the Death of a Loved One	593	5.3%	Parenthood and Parenting Challenges	202	1.8%
Dealing with the Loss of a Pet	601	5.4%	Low Self-Esteem or Lack of Confidence	302	2.7%
Work-related Stress and Burnout	403	3.6%	Body Image Concerns and Eating Disorders	101	0.9%
Financial Worries and Uncertainty	403	3.6%	LGBTQ+ Identity	101	0.9%
Unemployment-related Stress	403	3.6%	Cultural Identity and Belonging	101	0.9%
Academic Stress	403	3.6%	Academic Stress or Pressure	202	1.8%
Spirituality and Faith	202	1.8%	Job Loss or Career Setbacks	202	1.8%
Managing Bipolar Disorder	202	1.8%	Parenting Challenges and Parental Guilt	202	1.8%
Anxiety and Panic	202	1.8%	Sibling Rivalry or Family Conflict	403	3.6%
Depression and Low Mood	403	3.6%	Surviving and Recovering from Physical or Emotional Abuse	101	0.9%
Adjusting to a New Job or Role	302	2.7%	Healing from Sexual Assault or Domestic Violence	101	0.9%
Chronic Illness or Pain Management	302	2.7%	Post-Traumatic Stress Disorder (PTSD)	101	0.9%
Coping with a Diagnosis or Medical Treatment	202	1.8%	Healing from Abuse	202	1.8%
Caregiver Support	202	1.8%	Addiction and Recovery	202	1.8%
Finding Meaning and Purpose in Life	202	1.8%	Support for Loved Ones or Friends	202	1.8%

Figure 3.5: Statistics of all 36 emotional support scenarios covered in ExTES dataset

emotional support scenarios and response strategies were defined, drawing from literature on psychological counseling and previous emotional support research. Exemplar dialogues were then meticulously curated from existing datasets and online platforms. Subsequently, ChatGPT was employed to generate additional dialogues based on these seed exemplars, with manual correction applied wherever necessary to refine the generated dialogues. The resulting dataset encompassed a diverse set of emotionally varied scenarios and response strategies, as depicted in figures 3.5 and 3.6. This approach ensured the dataset's richness and relevance, facilitating the development of effective models for emotional support chat.

The data collection process commenced with the manual construction of 87 seed dialogues, each meticulously crafted to reflect authentic emotional support interactions. These seed dialogues were sourced from established emotion support datasets such as ESConv, ETMHS, and Reddit, providing a foundation of real-world scenarios. To ensure comprehensive coverage, web crawling techniques were employed to supplement the collection process, capturing additional dialogues that might not have been present in existing datasets. By combining dialogues from real emotion support datasets with those obtained through web crawling, the dataset achieved a balanced representation of authentic scenarios, ensuring di-

Category	Dialogues	Proportion
Reflective Statements (RS)	14,560	14.8%
Clarification (Cla)	2,898	2.9%
Emotional Validation (EV)	19,367	19.8%
Empathetic Statements (ES)	8,482	8.7%
Affirmation (Aff)	16,539	16.9%
Offer Hope (OH)	4,665	4.8%
Avoid Judgment And Criticism (AJC)	1,767	1.8%
Suggest Options (SO)	6,079	6.2%
Collaborative Planning (CP)	3,534	3.6%
Provide Different Perspectives (PDP)	3,322	3.4%
Reframe Negative Thoughts (RNT)	2,050	2.1%
Share Information (SI)	3,181	3.3%
Normalize Experiences (NE)	2,403	2.6%
Promote Self-Care Practices (PSP)	2,686	2.7%
Stress Management (SM)	2,474	2.5%
Others (Oth)	3887	3.9%
Overall	97,893	100%

Figure 3.6: Statistics of response strategies used in ExTES

versity and richness. Each of the 36 identified emotional support scenarios was represented by at least two seed dialogues, guaranteeing a wide spectrum of emotional situations and response strategies. Finally, to uphold the quality and relevance of the dataset, collected dialogues underwent rigorous manual correction and labeling of response strategies, ensuring accuracy and coherence for model training and analysis. An example of the data has been shown in figure 3.9.

3.2.2 Dialogue Quality Evaluation

The evaluation of dialogue quality in the ExTES dataset involved comprehensive human assessment, comparing the augmented dialogues with seed dialogues through crowd-sourcing. Following established metrics from previous studies, including informativeness, understanding, helpfulness, consistency, and coherence, each dialogue was rated on a four-level Likert scale by recruited college students. The results, as presented in figure 3.8, indicates that the method employed to generate dialogues, leveraging ChatGPT, yielded high-quality emotional

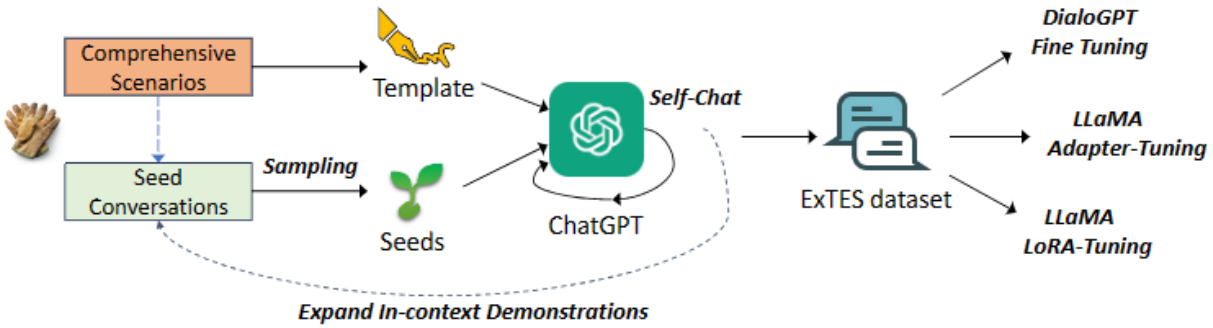


Figure 3.7: The pipeline for collecting the ExTES conversation dataset via our proposed extendable looping scheme. Based on the dataset, we benchmark and analyze the effect of fine-tuning the chat model with various techniques.

	Crowd-sourced	ExTES	κ
Informativeness	2.39	**2.53**	0.51
Understanding	**2.64**	2.52	0.46
Helpfulness	2.48	**2.61**	0.44
Consistency	**2.75**	2.67	0.39
Coherence	2.38	**2.45**	0.52

Figure 3.8: Human evaluation of ExTES quality. The scores (from 0 to 3) are averaged over all the samples rated by three annotators. κ denotes Fleiss' Kappa, indicating fair to moderate inter-annotator agreement ($0.2 < \kappa < 0.6$)

support interactions akin to those obtained through crowd-sourcing. Notably, the dialogues generated by ChatGPT exhibited comparable or even superior scores in informativeness and helpfulness when contrasted with the crowd-sourced seed dialogues. This suggests that ChatGPT's responses tended to offer more substantial and comprehensive content, contributing to the overall quality of the generated dialogues. These findings underscore the effectiveness of utilizing advanced language models like ChatGPT in constructing emotionally supportive conversational datasets. The comparison between ESConv data has been shown in figure 4.1

Scene: Communication Challenges

Description: With my roommate always leaving a mess in the kitchen, it's becoming increasingly hard to address the issue and maintain a harmonious living environment.

User : Hey.
AI : Hi, how are you?
User : I'm alright, thanks. But I've been having a hard time with my roommate lately. They always leave a mess in the kitchen, and it's getting frustrating to deal with.
AI (*Reflective Statements*): That sounds really irritating. It must be tough to constantly clean up after someone else.
User : Exactly. I've tried talking to them about it, but it doesn't seem to have any effect. I'm not sure how to approach the situation without causing conflict.
AI (*Suggest Options*): Maybe you could try setting up some agreed-upon rules for kitchen cleanliness. That way, both you and your roommate are on the same page.
User : That's a good idea. I'll give that a try. It's just been really stressful, and I don't want it to affect our relationship as roommates.
AI (*Normalize Experiences*): It's completely normal to have disagreements or conflicts with roommates. The key is finding a way to address the issues together that works for both of you.
User : You're right. I'll try to approach it from a calm and understanding perspective. Hopefully, we can come to a resolution.
AI (*Affirmation*): I believe in your ability to handle this situation maturely. Remember to take care of yourself throughout this process.
User : Thank you for the support. It means a lot to me. I'll do my best to handle it positively.
AI (*Offer Hope*): I'm confident that with open communication and a willingness to find a solution, you'll be able to improve your living situation. Don't hesitate to reach out for support if you need it.
User : I really appreciate your positivity and encouragement. It definitely gives me hope. Thank you again for being there.
AI : You're welcome! I'm here whenever you need to talk. Take care and hang in there!
User : Thank you. I'll keep that in mind. Take care too. Goodbye!

Figure 3.9: An example of ExTES data

Chapter 4

Methodologies

4.1 Training Dataset

Category	ExTES	ESConv
Dialogues	11,177	1,053
Utterances	200,393	31,410
Avg. length of dialogues	18.2	29.8
Avg. length of utterances	26.0	17.8
Num. of support strategise	16	8
Num. of scenarios	36	5

Figure 4.1: The statistics of ExTES vs ESConv

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 Figure 4.1, 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. By testing on both datasets, researchers can assess how well the models generalize to unseen data and whether they can effectively adapt to varying conversation contexts and characteristics.

4.2 Language Models for Experimentation

In the process of fine-tuning models for the task at hand, recent models with significantly fewer parameters and lower inference time were prioritized. The decision to opt for such models was driven by the need for efficiency without compromising performance. These models offer the advantage of faster inference times, making them more practical for real-world applications where speed is crucial[25].

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. Furthermore, their reduced parameter count made them more feasible for fine-tuning with limited computational resources while still achieving competitive performance.

By selecting models with lower parameter counts and faster inference times, we aimed to strike a balance between computational efficiency and model effectiveness. This approach allows for efficient training and deployment of models for emotional support conversation tasks, ultimately enhancing the scalability and accessibility of such systems.

4.2.1 Mistral 7B Instruct v0.2

Mistral 7B [26] 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) [27] and sliding window attention (SWA) [28], and [29]. 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

Parameter	Value
dim	4096
n layers	32
head dim	128
hidden dim	14336
n heads	32
n kv heads	8
window size	4096
context len	8192
vocab size	32000

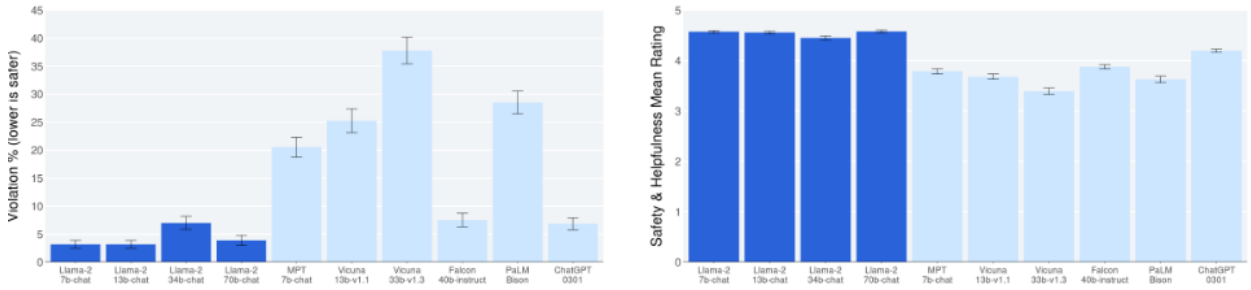
Table 4.1: Mistral model architecture

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.

SWA operates by leveraging the stacked layers of a transformer to attend to information beyond a specified window size, effectively extending the model's attention span. By recursively accessing tokens from the input layer within a certain range, SWA enables Mistral 7B to maintain a theoretical attention span of approximately 131K tokens, significantly enhancing its ability to process longer sequences. Additionally, the implementation of a rolling buffer cache further optimizes memory usage by limiting the cache size based on the window size, resulting in an 8x reduction in cache memory usage without compromising model quality.

Moreover, Mistral 7B employs strategies such as pre-fill and chunking to enhance sequence generation efficiency. By pre-filling the cache with the known prompt or breaking down large prompts into smaller chunks, Mistral 7B streamlines the token prediction process, enabling smoother generation of sequences. This combination of innovative attention mechanisms, efficient memory management, and sequence generation strategies collectively contributes to the enhanced performance and efficiency of Mistral 7B, making it a significant milestone in the development of language models.

4.2.2 LLama2 7B Chat



(a) Overall violation percentage.

(b) Overall safety and helpfulness mean rating.

Figure 4.2: Overall safety measures.Left: Llama 2-Chat has a low violation percentage overall across model sizes. Right: Llama 2-Chat has a high safety and helpfulness mean rating overall across model sizes

Llama 2-Chat was chosen as the base model for fine-tuning into an emotional support chatbot 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 [25], which includes supervised fine-tuning and RLHF. The incorporation of human preference data for reward modeling ensures that the model is trained to generate responses that are not only helpful but also safe. 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. The iterative nature of collecting preference data allows for continuous improvement of the reward model, ensuring that the fine-tuned emotional support chatbot maintains an accurate understanding of user preferences and safety considerations over time.

Llama 2-Chat's safety evaluation is particularly crucial for its adaptation into an emo-

tional support chatbot. The safety assessment involves collecting adversarial prompts and evaluating model responses for safety violations shown in 4.2. 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. This emphasis on safety aligns with the ethical considerations inherent in providing emotional support and underscores the importance of selecting a model that not only generates helpful responses but also upholds user safety and well-being as paramount concerns. Therefore, the comprehensive safety evaluation process of Llama 2-Chat makes it a suitable candidate for the development of an emotionally supportive chatbot that prioritizes both effectiveness and user safety.

4.2.3 Phi-3-Mini-4k-Instruct

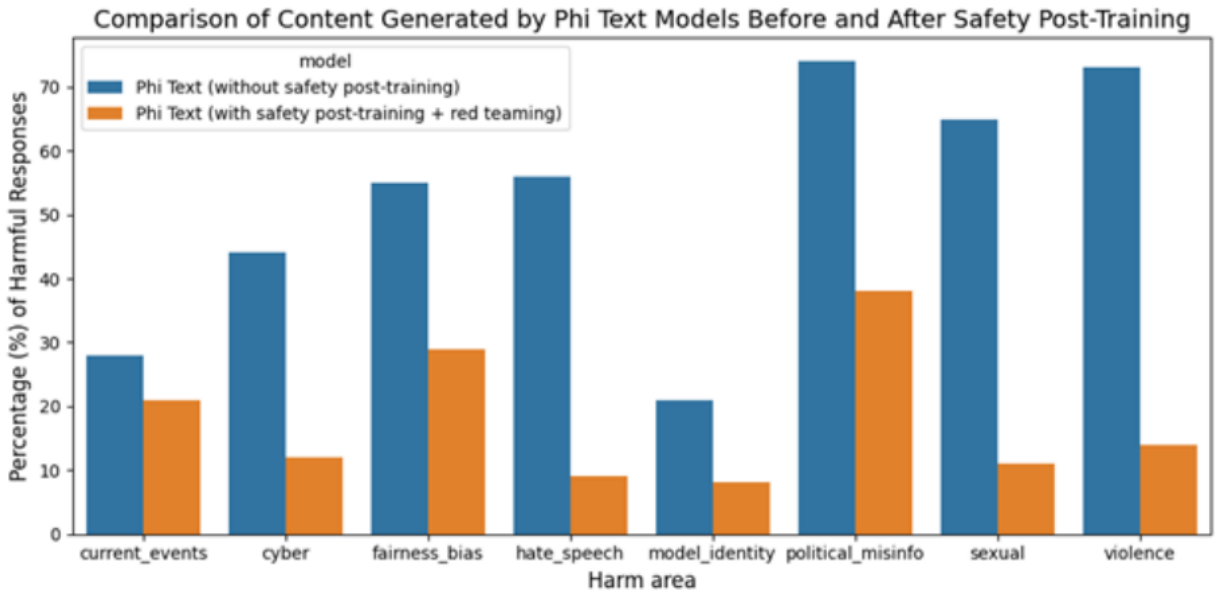


Figure 4.3: Comparison of harmful response percentages by Microsoft AI Red Team between phi-3-mini before and after the safety alignment.

Phi-3-mini [30] emerges as a compelling choice for fine-tuning an emotional support chatbot due to several key factors, particularly its compact size and mobile-friendly deployment capabilities. With a transformer decoder architecture and a default context length of 4K,

phi-3-mini offers a suitable framework for generating responses in dialogue settings. Moreover, the introduction of a long context version via LongRope extends the context length to 128K, enhancing its capacity to handle longer conversational contexts, a critical aspect in emotional support interactions where the conversation may delve into deeper and more complex topics.

What makes phi-3-mini particularly appealing is its optimization for deployment on mobile devices. With a size that allows for quantization to 4-bits, phi-3-mini occupies only approximately 1.8GB of memory, making it highly suitable for running on smartphones. This capability enables users to access emotional support resources conveniently and efficiently, directly from their mobile devices, without the need for extensive computational resources or constant internet connectivity.

Furthermore, phi-3-mini's training methodology underscores its suitability for emotional support applications. By leveraging high-quality training data and focusing on data optimization for small-scale models, phi-3-mini achieves a level of performance comparable to larger models like GPT-3.5 or Mistral, despite its significantly reduced parameter count. The emphasis on supervised fine-tuning and direct preference optimization during post-training ensures that the model's responses are not only accurate but also safe and helpful, aligning with the ethical considerations inherent in providing emotional support.

Phi-3-mini, chosen for fine-tuning an emotional support chatbot, prioritizes safety through rigorous post-training alignment, red-teaming exercises, and evaluations across responsible AI harm categories, adhering to Microsoft's principles. Leveraging preference datasets focused on helpfulness and harmlessness, along with in-house generated datasets, the model filters potentially harmful responses 4.3, ensuring positive interactions. Ongoing optimization efforts aim to maintain safety and reliability, making phi-3-mini well-suited for sensitive domains like emotional support, where user well-being is paramount. Additionally, its smaller size and deployability on mobile devices make it accessible even with limited computational resources.

4.3 Low Ranked Adaptation (LORA)

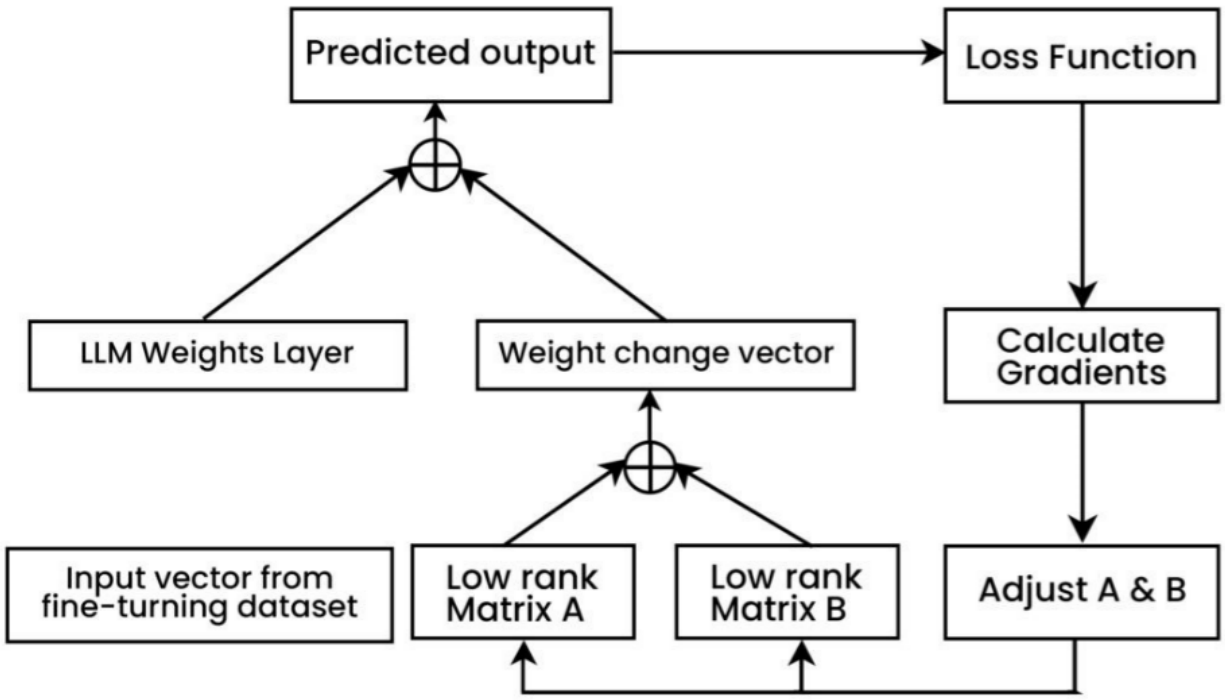


Figure 4.4: LoRA training loop

The Low-Rank Adaptation (LoRA) [18] technique offers a promising approach for fine-tuning language models in the domain of emotional support while addressing key challenges associated with traditional fine-tuning methods. By freezing the pretrained model weights and introducing trainable rank decomposition matrices into each layer of the Transformer architecture, LoRA significantly reduces the number of trainable parameters for downstream tasks. For instance, compared to fine-tuning GPT-3 175B with Adam, LoRA can reduce the number of trainable parameters by 10,000 times and the GPU memory requirement by 3 times, making it more efficient for deployment in resource-constrained environments.

In the context of emotional support, LoRA enables the adaptation of large-scale pre-trained language models to effectively address users' emotional needs while minimizing computational resources and maintaining model quality. The figure 4.4 shows LoRA training loop in which instead of gradients being applied directly to the hidden weight matrix, they

are applied to lower rank matrices A and B. Once A and B matrices are trained, the change matrix constructed from A and B is simply added to the weight matrix at end of training and this process saves billions of matrix multiplication operations for LLM. By constraining the updates to the weights with low-rank decompositions, LoRA facilitates efficient training and lowers the hardware barrier to entry, making it accessible for deployment on a variety of platforms, including mobile devices. Furthermore, the linear design of LoRA allows for seamless integration with existing methods, such as prefix-tuning, without introducing additional inference latency.

Moreover, LoRA facilitates efficient task switching during deployment by allowing only the swapping of LoRA weights rather than all parameters. This capability enables the creation of customized models tailored to specific tasks, which can be dynamically swapped in and out as needed, even on devices with limited VRAM. Additionally, LoRA offers a speedup during training, with observed improvements of up to 25% compared to full fine-tuning on models like GPT-3 175B, as it eliminates the need to calculate gradients for the majority of parameters.

Applying LoRA to Mistral, Llama, and Phi models offers significant benefits for emotional support chatbot fine-tuning. LoRA reduces the number of trainable parameters while maintaining or improving model quality, enhancing efficiency and resource utilization. Mistral, prioritizing efficiency and speed, could leverage LoRA to further optimize its inference speed and memory requirements, enhancing its suitability for real-time emotional support interactions. Similarly, Llama's focus on dialogue applications could benefit from LoRA by streamlining adaptation processes and enabling the model to handle a wider range of emotional support scenarios. For Phi models, designed for deployment on mobile devices, integrating LoRA could optimize efficiency and resource usage, making emotional support chatbots more accessible on mobile platforms while maintaining high performance. Overall, LoRA enhances the efficiency and effectiveness of these models for emotional support chatbot applications, improving accessibility and usability for users seeking assistance.

4.4 Context Management

Context management techniques in LLMs play a crucial role in enhancing their performance and generating more coherent responses. Two notable techniques are recursive summarization and soft prompt compression.

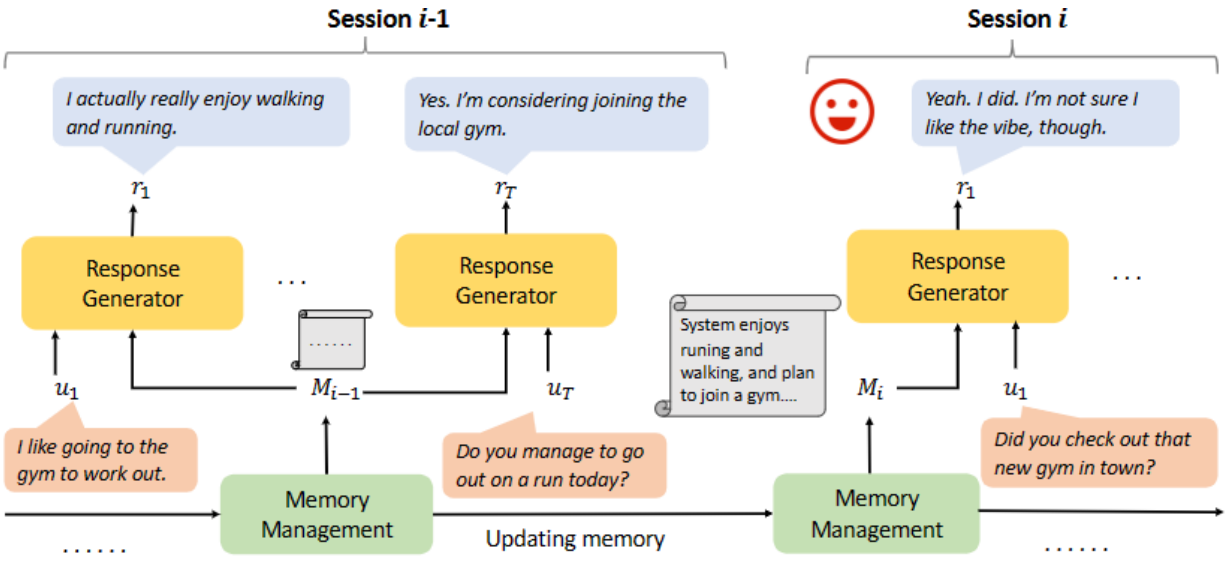


Figure 4.5: The schematic overview of recursive summarization

Recursive summarization [5] involves iteratively summarizing a piece of text to distill its main ideas into a concise form, enabling the model to focus on relevant information while disregarding redundant or less important details 4.5. This technique helps LLMs manage complex input and generate more concise and relevant responses.

Soft prompt compression [6] is another context management technique that involves compressing the input prompt or context into a more condensed form, allowing the model to focus on the most salient aspects of the input. By compressing the prompt, the model can better understand the context and generate responses that are more aligned with the user's intent 4.6.

Additionally, recent research has explored the integration of user-specific knowledge into LLMs for generating responses in longitudinal dialogues (LDs) [7]. By grounding LLMs with

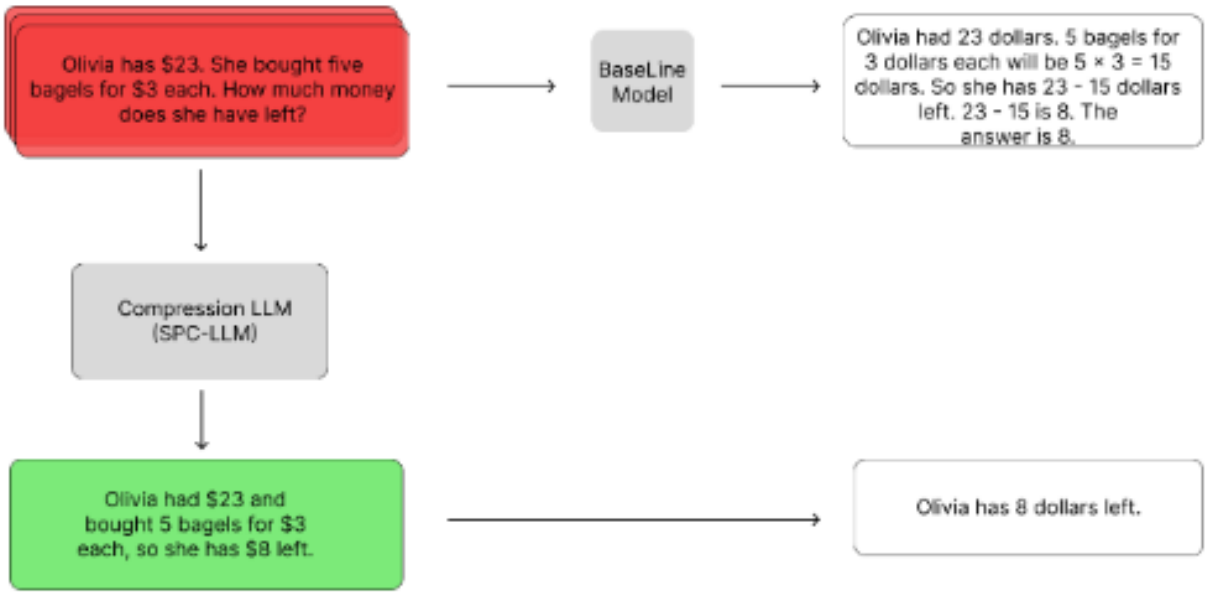


Figure 4.6: Illustration of SPC shows the compressed conversational answer expect with question.

user-specific knowledge 4.7, such as personal experiences and preferences, these models can generate more coherent and personalized responses. Figure 4.7 includes techniques such as representing user knowledge through syntactic or graph-based representations have shown promising results in improving the quality of generated responses in LDs.

Influencing from previous approaches, we have also explored techniques such as removing stopwords to increase the context length effectively. By removing stopwords, which are common words that carry little semantic meaning, the model can focus on more informative content, thereby potentially improving its understanding of the context and generating more relevant responses.

In the context of emotional support conversations, where real-time interaction and responsiveness are crucial, the computational demands of recursive summarization and soft prompt compression could hinder the efficiency of LLMs. Processing and summarizing large amounts of text recursively or compressing prompts into concise forms may introduce latency, which is undesirable in scenarios where prompt responses are needed to provide emotional support effectively.

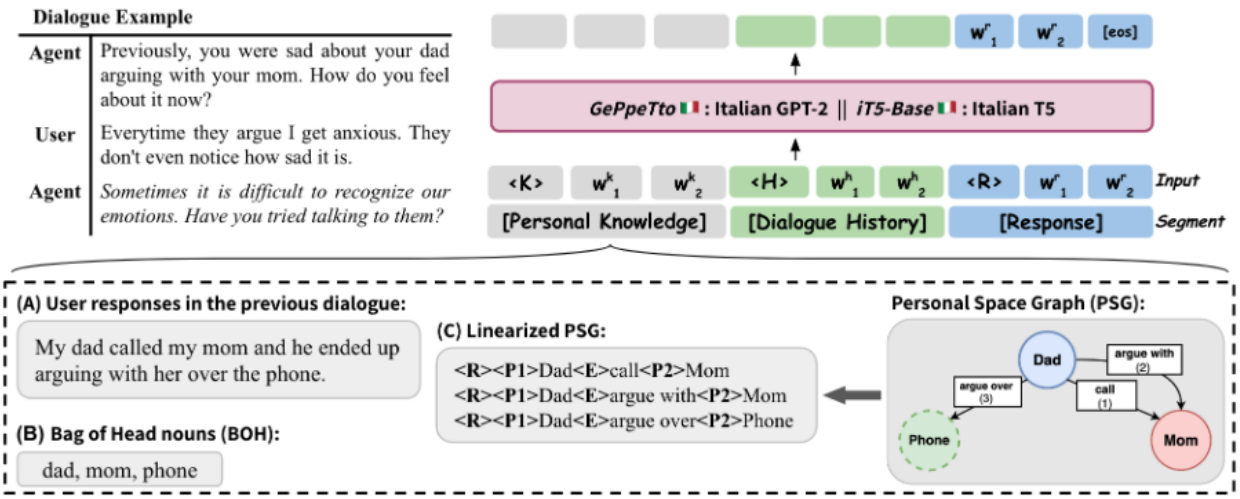


Figure 4.7: Representations of personal knowledge using Bag of Head nouns (BOH), Unprocessed text (RAW), and Personal Space Graph (PSG).

Alternatively, user-specific knowledge and strategies like removing stopwords can offer more computationally efficient approaches to managing context in emotional support conversations. LDs, which encompass multi-session interactions with users, provide a natural context that can be leveraged by LLMs without the need for extensive preprocessing or summarization. By maintaining context across multiple interactions, LLMs can better understand users' emotions and tailor responses accordingly.

Similarly, techniques like removing stopwords from user input can help increase the effective context length without significantly increasing computational complexity. By focusing on the most informative words and phrases while discarding redundant or less meaningful ones, LLMs can efficiently utilize available context to generate empathetic and supportive responses in emotional support conversations.

Chapter 5

Design and Development

5.1 Fine Tuning Architecture

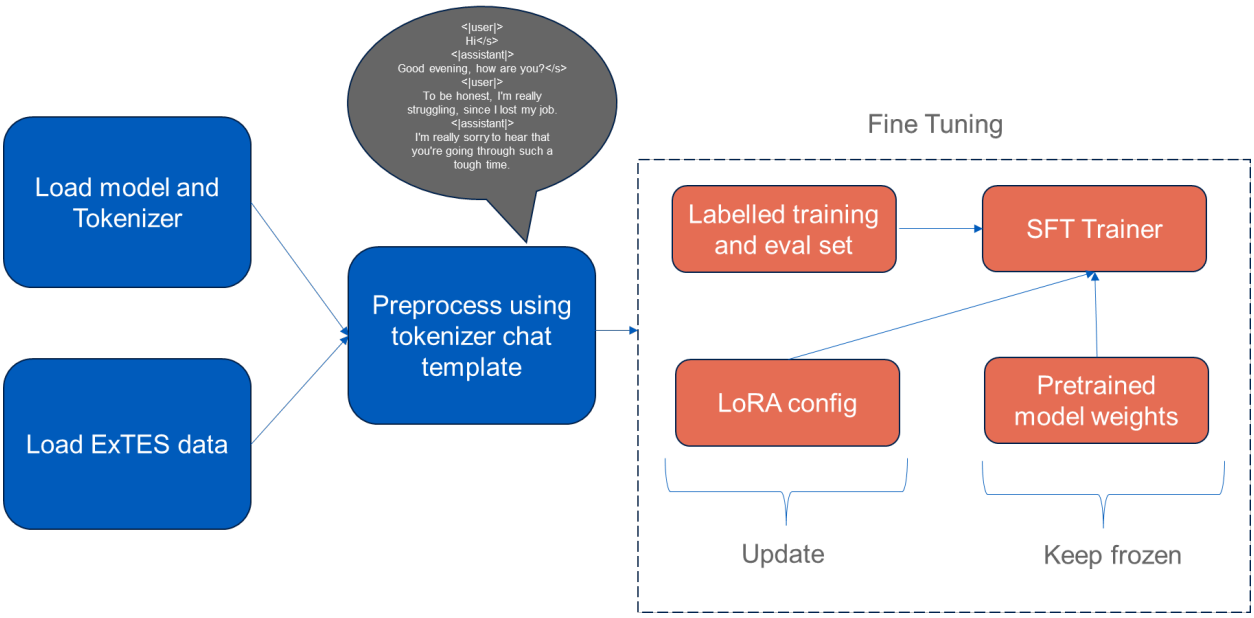


Figure 5.1: Finetuning design

The architecture depicted in Figure 5.1 illustrates the systematic approach to fine-tune a pretrained language model using the Low-Rank adaptation Technique for emotional support conversations. Initially, the pretrained language model and its tokenizer are loaded, typically at half precision except for Llama2 due to compatibility reasons. These components form the core infrastructure for generating responses, with the tokenizer responsible for processing input text into a format interpretable by the model. Ahead of loading, the ExTES dataset undergoes preprocessing to adhere to the user-followed-by-assistant conversational structure,

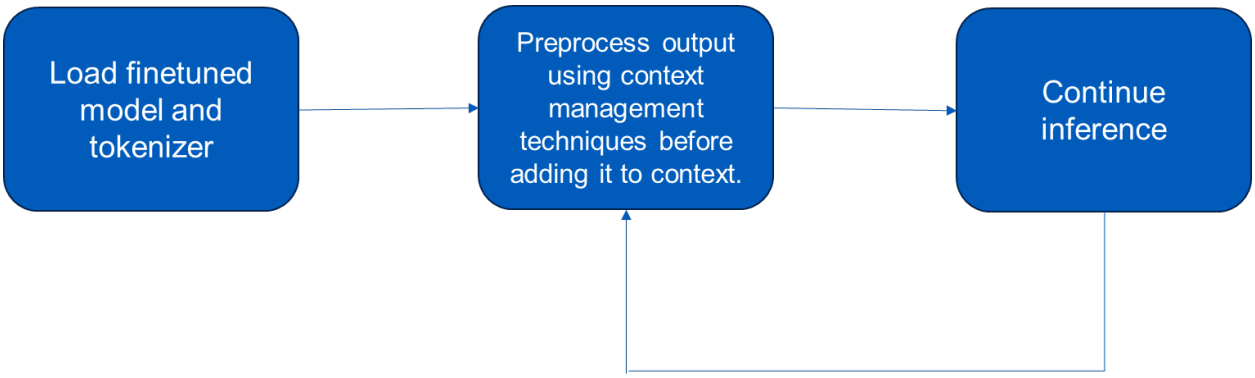


Figure 5.2: Inference

a prerequisite for chat model fine-tuning. Utilizing a specific chat template unique to each model, this preprocessing involves embedding special tokens and formatting conventions to demarcate conversational turns and session boundaries effectively.

Subsequently, the preprocessed ExTES dataset is partitioned into labeled training and evaluation subsets. The former fuels the fine-tuning process, with a specialized trainer, like the Supervised Fine-Tuning (SFT) trainer, orchestrating the adjustment of language model parameters based on the annotated training data. This iterative process optimizes the model's performance for the nuances of emotional support conversations. Crucially, the fine-tuning procedure incorporates the LoRA configuration, wherein only the LoRA weights undergo updates, preserving the integrity of the pretrained model weights. LoRA introduces trainable rank decomposition matrices to each layer, effectively reducing the number of trainable parameters while retaining crucial information gleaned during pre-training. This strategic weight update strategy streamlines the fine-tuning process, enhancing its efficiency and resource utilization.

During the inference process (see Figure 5.2), the output generated by the model undergoes preprocessing, notably involving the removal of stopwords. This step is crucial for effectively managing the limited context memory available to the model during inference. By eliminating stopwords—common words that carry minimal semantic meaning—the model can focus on more informative content within the given context, thus enhancing the relevance

Parameter	Value
learning rate	2.0e-04
LoRA rank	16
LoRA alpha	32
LoRA dropout	0.1
bias	None
task type	CAUSAL LM
target modules	q_proj, k_proj, v_proj, o_proj

Table 5.1: Training parameters

and coherence of the generated responses.

The user-specific knowledge method, which typically involves incorporating personalized information into the model's understanding, was not utilized in this particular scenario. Despite attempts to integrate this approach, challenges arose when certain utterances lacked nouns or verbs necessary for constructing meaningful knowledge representations. Consequently, the absence of such linguistic elements hindered the model's ability to generate coherent responses, prompting the decision to forego the user-specific knowledge method in favor of other preprocessing techniques aimed at improving response quality.

5.2 Training

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. These parameters shown in Table 5.1 were fine-tuned in alignment with the model's learning dynamics, ensuring optimal adaptation to the emotional support conversation task. Throughout the training process, the model iteratively adjusted its weights, leveraging the LoRA technique to selectively update trainable rank decomposition matrices while keeping the pretrained model weights frozen. This approach facilitated the reduction of trainable parameters, thereby enhancing computational efficiency and memory utilization without compromising the valuable information gleaned during pre-training. By adhering to the prescribed LoRA configuration and adapting the model's parameters accordingly,

the training procedure aimed to strike a delicate balance between model complexity and performance, ultimately striving to optimize the model's ability to generate coherent and contextually relevant responses in emotional support conversations.

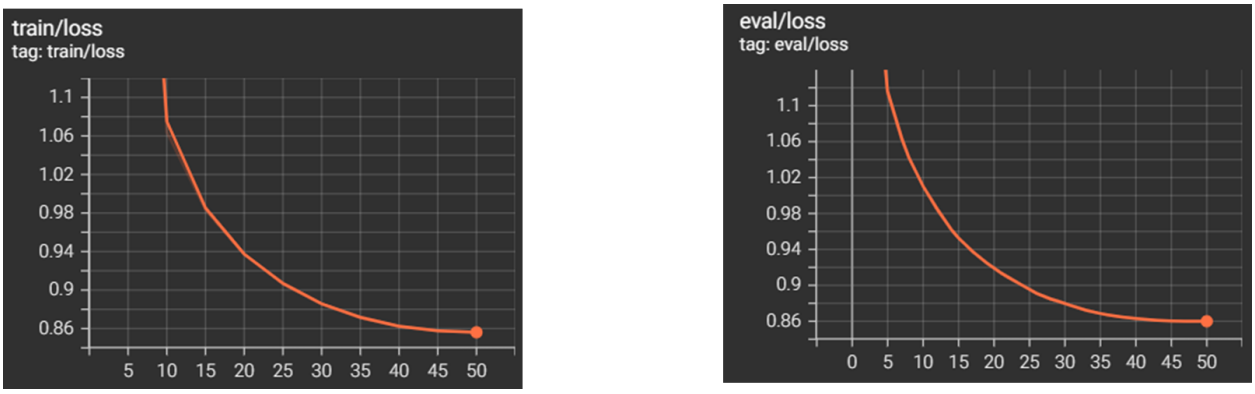


Figure 5.3: Mistral Training and Evaluation loss

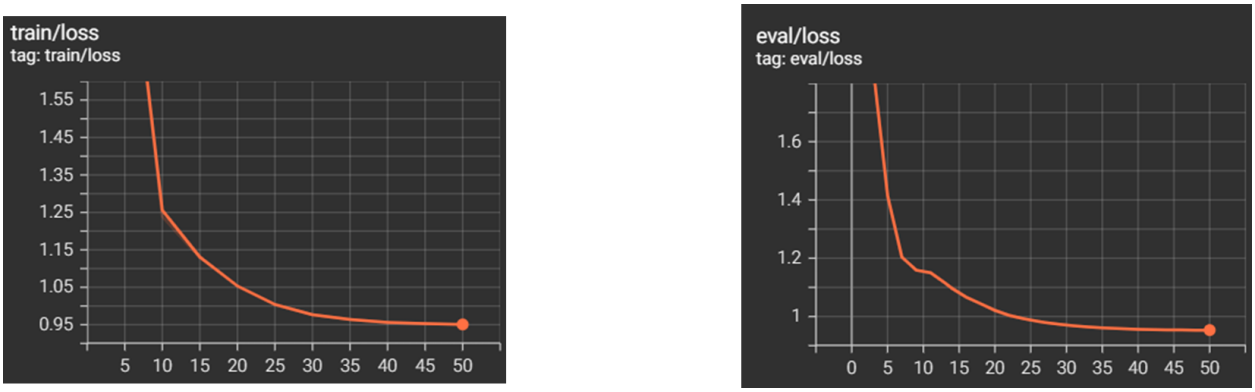


Figure 5.4: Llama Training and Evaluation loss

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 without excessively memorizing it. Conversely, the training curve for the Phi 3 model displayed some inconsistencies. However, despite the fluctuations in the training loss, the evaluation loss exhibited a steady decrease over time. Consequently, it was decided to proceed with running inference on the Phi 3

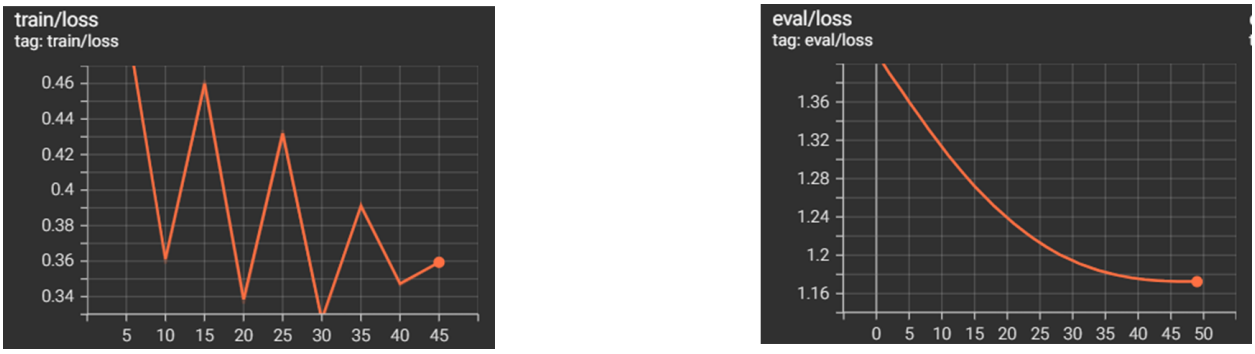


Figure 5.5: Phi3 Training and Evaluation loss

model as well, based on the promising trend of decreasing evaluation loss.

Chapter 6

Results

During the inference phase, a standardized set of parameters shown in 6.1 was utilized for text generation across all models to ensure consistency in the generation process and enable fair comparisons between different models. This approach facilitated an unbiased evaluation of the models' performance. The inference was conducted on both the ExTES and ESConv datasets, allowing for cross-dataset validation and assessment of the models' generalization ability across different conversation corpora. The evaluated responses are compared with the base models and the finetuned ones.

Evaluation of the generated responses involved a combination of automatic metrics and using an LLM as a judge which was demonstrated in [12], [10] and [11]. Given the constraints and limitations associated with human evaluations, leveraging LLMs 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.

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

Table 6.1: Generation parameters

Model	PPL	B-2	R-2	R-L	Bert P	Bert R	Bert F1
Base	44.57	1.2	7.7	17	49.4	61.4	54.5
Base w/o sw	679.725	8.3	4.2	12.7	45.7	55.2	49.2
Finetuned	3.696	21.1	17.7	30.9	61.8	62.7	62.1
Finetuned w/o sw	41.99	19.02	14.2	25.6	60.6	57.7	58.9

Table 6.2: Mistral on ExTES dataset

6.1 Performance on ExTES

6.1.1 Mistral 7B

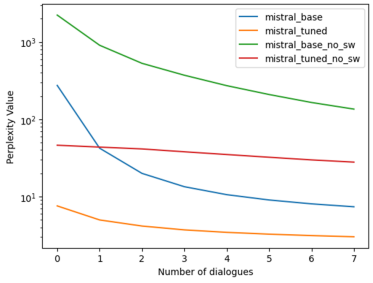


Figure 6.1: PPL range with number of utterances

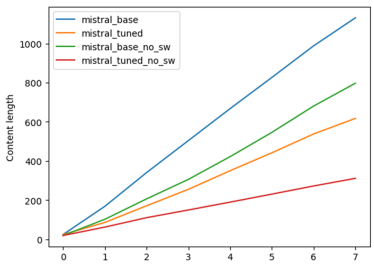


Figure 6.2: Content length with increase in utterances

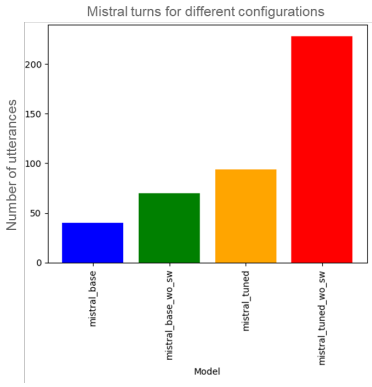


Figure 6.3: Stopwords impact on no of utterances

In automatic evaluations (see Figure 6.2), Mistral demonstrates superior performance compared to the other two models, exhibiting better perplexity, BLEU, ROUGE, and BERT scores. The significant disparity in perplexity scores between the base and fine-tuned models indicates that the fine-tuned model exhibits much higher confidence in predicting the next token. However, there is a slight decrease in performance when compared to the model trained with different context representations achieved by removing stopwords. The trend of perplexity values over increasing utterances, as depicted in Figure 6.1, showcases a consistent decrease across all model versions, as expected. The fine-tuned model exhibits the best performance, followed by the base model without stopwords. Surprisingly, removing stopwords leads to an increase in perplexity values for both models.

Further analysis in Figure 6.2 illustrates the relationship between the length of context and utterances. The base model comprises over 1000 tokens in just seven utterances, while the fine-tuned model contains 600 tokens, and the fine-tuned version with stopwords removed contains slightly above 200 tokens, representing almost a one-third decrease in context length. Additionally, Figure 6.3 demonstrates the impact of this trend on the number of utterances accommodated. Removing stopwords allows for the accommodation of almost 250 utterances, whereas including stopwords limits the accommodation to less than 100 utterances. This discrepancy highlights the trade-off between metric performance and the amount of context accommodated, emphasizing the potential benefits of removing stopwords, particularly in longitudinal dialogues.

6.1.2 Llama 7B Chat

In Llama models, a similar trend to Mistral was observed, although the base models exhibited much lower perplexity scores comparatively. This difference may be attributed to the chat version of the model, which has been trained on a substantial amount of conversational data, unlike Mistral. The number of utterances for the Llama model was significantly lower than Mistral (see Figure 6.6), primarily due to the shorter context length of 4k tokens compared to Mistral's 8k. Despite these differences, the overall trends remained consistent with Mistral. Details and values are provided in Table 6.3 and Figures 6.5, 6.6, and 6.4.

In automatic evaluations, the scores for the Llama model are generally lower compared to the Mistral model across almost all categories. This discrepancy suggests that the Mistral model outperforms the Llama model in terms of various evaluation metrics such as perplexity, BLEU, ROUGE, and BERT scores. The differences in performance may be attributed to several factors, including the architecture design, training data, and fine-tuning process. Despite these variations, both models demonstrate consistent trends in their performance across different evaluation criteria.

Model	PPL	B-2	R-2	R-L	Bert P	Bert R	Bert F1
Base	20.13	13.4	9.3	18.6	49.9	63.2	55.5
Base w/o sw	232.90	6.2	6.9	16.4	48.2	60.4	53.4
Finetuned	6.47	19.4	9.6	21.6	55.1	55.2	54.7
Finetuned w/o sw	44.60	11.4	3.2	11.4	49.5	46.6	47.6

Table 6.3: Llama on ExTES dataset

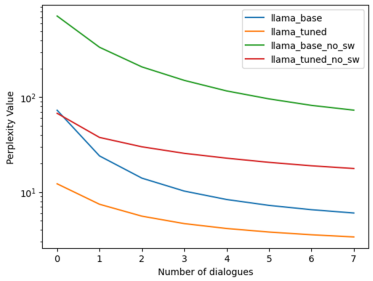


Figure 6.4: PPL range with number of utterances

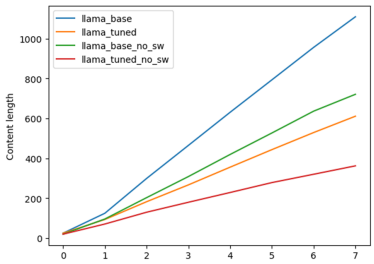


Figure 6.5: Content length with increase in utterances

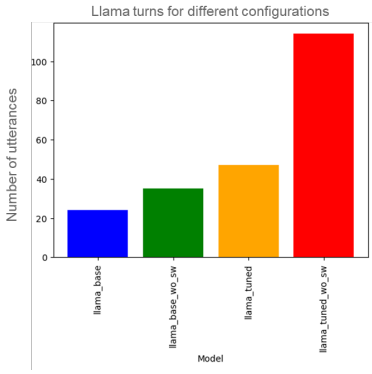


Figure 6.6: Stopwords impact on no of utterances

6.1.3 Phi 3 mini

The Phi3 model, despite its smaller size compared to Mistral and Llama, exhibited poor performance in terms of automatic metrics 6.4. Although the perplexity values were lower, indicating better prediction confidence, the other evaluation scores were significantly lower compared to the other two models. Upon inspecting the model's outputs, it was observed that they did not correlate well with the fine-tuned datasets, suggesting a lack of coherence and relevance in the generated responses. Moreover, the training curve exhibited erratic behavior, oscillating between values, which further indicated instability in the training process. Due to these unsatisfactory results from both automatic evaluations and qualitative assessments

Model	PPL	B-2	R-2	R-L	Bert P	Bert R	Bert F1
Base	6.57	0.8	3.3	6.6	40.3	59.1	47.9
Base w/o sw	113.73	0.7	2.9	6.7	34.5	50.4	40.9
Finetuned	6.32	1.45	3.8	7.2	40.5	60.0	48.3
Finetuned w/o sw	108.04	0.86	3.5	7.5	34.7	49.3	40.7

Table 6.4: Phi3 on ExTES dataset

Model	PPL	B-2	R-2	R-L	Bert P	Bert R	Bert F1
Base	3045.5	1.1	1.0	7.5	35.8	46.0	40.0
Base w/o sw	4155.9	0.5	0.6	5.0	34.1	42.7	37.6
Finetuned	12.05	34.0	30.5	34.56	39.5	46.5	42.4
Finetuned w/o sw	64.96	32.77	20.33	29.45	39.4	39.7	39.1

Table 6.5: Mistral on ESConv dataset

of the model's outputs, further experiments with the Phi3 model were not pursued.

6.2 Performance on ESConv

The evaluations were conducted on the ESConversation dataset to assess the models' ability to generalize to similar conversational data. This cross-dataset evaluation provided insights into how well the models could adapt to different conversation corpora and handle diverse conversational styles and topics. By evaluating the models on ESConversation, which shares similarities with ExTES but contains distinct conversation dynamics and themes, we could gain a better understanding of their generalization capabilities. This approach helped validate the robustness and versatility of the models beyond the specific dataset they were fine-tuned on, providing valuable insights into their real-world applicability for emotional support conversations across various contexts.

6.2.1 Mistral 7B

On the ESConv dataset, Mistral exhibited similar trends to those observed on the ExTES dataset, albeit with some variations in certain metrics. While the general trends remained consistent, there were notable differences in specific evaluation scores. For instance, metrics such as Rouge and Bleu scores showed considerable improvement, indicating Mistral's good generalizability across datasets (6.5). However, perplexity values were noticeably higher for both the base and fine-tuned models on the ESConv dataset compared to ExTES (6.8). This difference can be attributed to the structured nature of the ExTES data, where it was

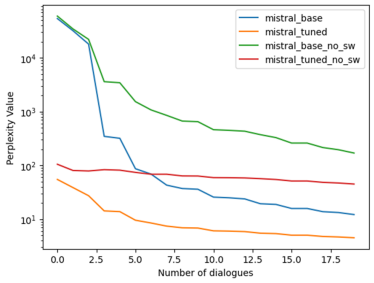


Figure 6.7: PPL range with number of utterances

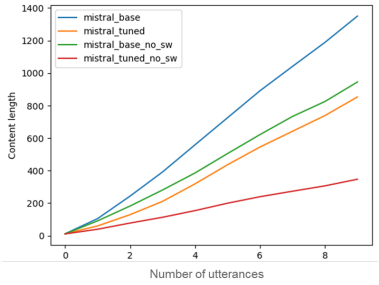


Figure 6.8: Content length with increase in utterances

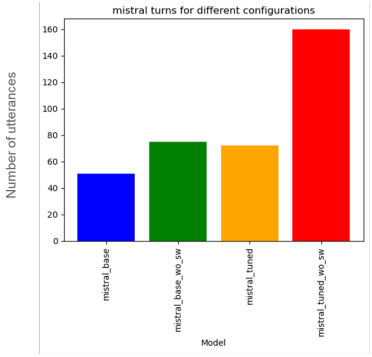


Figure 6.9: Stopwords impact on no of utterances

Model	PPL	B-2	R-2	R-L	Bert P	Bert R	Bert F1
Base	280.31	1.3	1.0	7.7	35.7	46.0	40.0
Base w/o sw	694.99	1.2	0.9	7.6	35.5	46.2	39.9
Finetuned	14.93	26.8	21.6	14.5	38.5	46.1	41.6
Finetuned w/o sw	55.56	18.6	9.4	7.4	35.1	37.3	35.9

Table 6.6: Llama on ESConv dataset

relatively easier to predict the next token, resulting in lower perplexity values. Despite these variations, the overall trends observed on ESConv, including context length and utterance accommodation, were consistent with those seen on ExTES (6.7, 6.9). This indicates Mistral's robust performance across different conversation datasets, reaffirming its potential for diverse conversational contexts.

6.2.2 Llama 7B Chat

On the ESConv dataset, Llama exhibited a trend similar to that observed on ExTES, with some notable differences in evaluation metrics. While the overall trend remained consistent, Llama's automatic evaluation metrics were notably lower compared to Mistral, indicating comparatively lower performance (6.6). However, perplexity scores were considerably lower for both the base and fine-tuned models on the ESConv dataset compared to ExTES (6.11). This difference could be attributed to the chat data on which the initial base model was trained, which may have provided a more conversational context conducive to lower per-

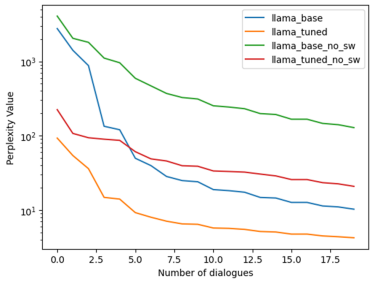


Figure 6.10: PPL range with number of utterances

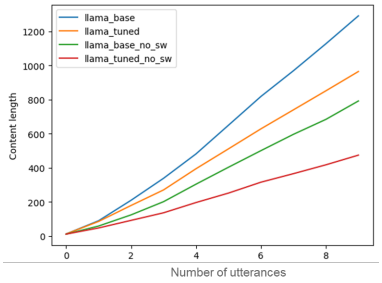


Figure 6.11: Content length with increase in utterances

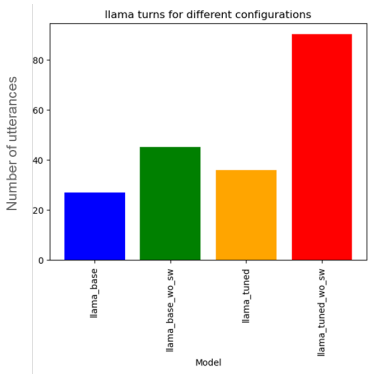


Figure 6.12: Stopwords impact on no of utterances

Model	PPL	B-2	R-2	R-L	Bert P	Bert R	Bert F1
Base	9.66	0.71	0.3	0.6	36.0	39.1	38.4
Base w/o sw	119.79	0.81	0.6	4.3	31.2	47.4	42.3
Finetuned	14.975	0.1	0.84	0.7	31.5	46.0	37.3
Finetuned w/o sw	176.79	0.16	0.4	0.6	25.7	37.9	30.4

Table 6.7: Phi3 on ESConv dataset

plexity scores. Despite these variations, the general trends observed on ESConv, including context length and utterance accommodation, were consistent with those seen on ExTES (6.10, 6.12). Overall, Llama exhibited a similar trend across different conversation datasets, indicating its potential for various conversational contexts despite the differences in automatic evaluation metrics.

6.2.3 Phi 3 mini

On the ESConv dataset, Phi3 demonstrated poor performance across various automatic evaluation metrics, mirroring the trends observed on the ExTES dataset although the perplexity values were notably lower compared to the other models(6.7). However, other evaluation metrics such as BLEU, ROUGE, and BERT scores were considerably lower, suggesting subpar performance with respect to reference sentences. Due to these unfavorable results, further experiments were not conducted with Phi3 on the ESConv dataset. The specific values for each metric can be referred to in the corresponding table (6.7).

6.3 GPT for evaluations

GPT-4 has exhibited promising capabilities in understanding and effectively serving as a judge for various conversational tasks, as highlighted by previous research studies. Leveraging its advanced language understanding capabilities and extensive pre-training, GPT-4 has demonstrated the ability to comprehend nuanced conversational contexts, discern user intents, and provide relevant and coherent responses. This proficiency makes GPT-4 a valuable tool for evaluating the performance of language models in tasks such as emotional support conversation, where context comprehension and response quality are paramount. By employing GPT-4 as a judge in evaluation processes, researchers can leverage its robust capabilities to assess the effectiveness and appropriateness of language model responses, contributing to the refinement and optimization of conversational AI systems. The prompt for evaluation is shown in 6.13 which has been curated based on the prompts used in [12].

Evaluate the conversation on a score of 0 to 10 on the basis how emotionally supportive the assistant is to the user. The assistant should be able to offer good suggestion, provide comfort and must be engaging and relevant to the content. The criteria for evaluation is Engagement, Fluency, Comforting, Appropriateness and Suggestion. The output should be in the form of json.

Figure 6.13: Prompt for evaluation

The evaluations of the language models 6.8 reveal that the llama2 chat fine-tuned model performs the best overall, followed by the fine-tuned model without stopwords, which contrasts with the results obtained from automatic evaluations. Specifically, the base model llama2 demonstrates excellence in engagement, fluency, and suggestion categories, likely at-

Model	**Engagement**	**Fluency**	**Comforting**	**Appropriateness**	**Suggestion**
Mistral (base)	7	8.5	7	<u>8.5</u>	8
Mistral (finetuned)	<u>9.25</u>	9	**9.5**	**10**	9
Mistral (finetuned w/o sw)	9	7	8.5	<u>8.5</u>	8
Llama2 chat (base)	9	**10**	7.45	<u>8.15</u>	8.75
Llama2 chat (finetuned)	**9.5**	**10**	**9.5**	**10**	**9.5**
Llama2 chat (finetuned w/o sw)	**9.5**	<u>9</u>	<u>9</u>	10	<u>9</u>

Table 6.8: GPT4 for evaluations

tributed to its extensive training on chat data. Mistral, while not far behind, achieves the highest scores in the comforting category, which is particularly crucial in emotional support conversations. It's noteworthy that the evaluation was conducted solely on ExTES data, focusing on limited dialogues due to cost constraints. Despite these limitations, the findings underscore the nuanced performance of different models across various evaluation metrics and shed light on their respective strengths and weaknesses in supporting emotional support conversations. The best scores are highlighted in bold and the second best are underlined.

Model	ExTES	ESConv
Mistral Base	9.2	7.8
Mistral Base w/o sw	9.7	8.2
Mistral Fine-tuned	**3.4**	**3.0**
Mistral Fine-tuned w/o sw	3.7	3.2
Llama2 Base	15.5	11.3
Llama2 Base w/o sw	16.2	12.1
Llama2 Fine-tuned	7.4	6.80
Llama2 Fine-tuned w/o sw	7.9	7.1
Phi3 Base	5.6	7.2
Phi3 Base w/o sw	5.5	7.4
Phi3 Fine-tuned	4.3	6.91
Phi3 Fine-tuned w/o sw	4.6	7.1

Table 6.9: Inference times

6.4 Inference times

From Table 6.9, it's evident that the Mistral fine-tuned model exhibited the shortest inference time, clocking in at just 3.4 seconds. This impressive speed was nearly 300% faster than the base model, which took 9.2 seconds. The notable reduction in inference time can be attributed to the substantially lower perplexity values observed during inference from both datasets. Fine-tuning, coupled with nucleus sampling, likely contributed to the decreased number of tokens sampled during inference, thus leading to quicker processing times. The reduced time in mistral models can be attributed to the sliding window attention discussed in Chapter 4. This trend of fine-tuned models boasting shorter inference times was consistently observed across all models, further highlighting the efficiency gains achieved through fine-tuning.

6.5 Examples

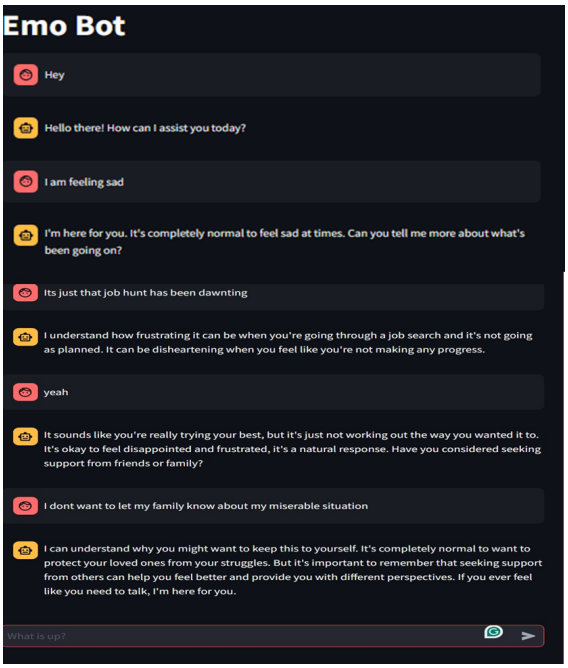


Figure 6.14: COnversation with Fine-tuned model

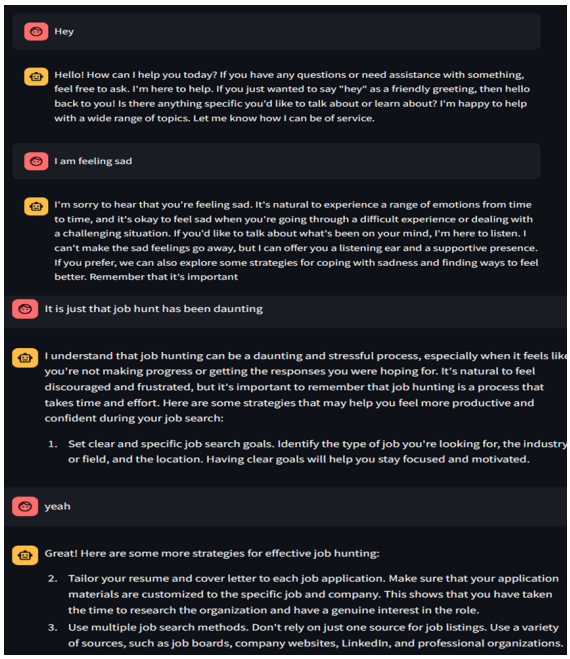


Figure 6.15: Conversation with base model

Figure 6.15 depicts an interaction with the base model, while Figure 6.14 illustrates the interaction with the fine-tuned model aimed at providing emotional support. In the base model interaction, the responses primarily consist of suggestions and advice, with minimal engagement or attempt to initiate conversation with the user. Conversely, in the interaction with the fine-tuned model, the responses are characterized by shorter, more coherent messages. The bot actively engages with the user by asking questions about their situation and offering support through phrases like "I am here for you." Overall, the fine-tuned model demonstrates a more engaging and supportive interaction compared to the base model.

Chapter 7

Conclusion and Future work

In conclusion, the evaluation results shed light on the strengths and performance of the Mistral and Llama models in the context of emotional support conversations. Mistral demonstrated superior performance in automatic evaluations, boasting commendable scores across various metrics. On the other hand, Llama garnered positive feedback in GPT-4 evaluations, particularly excelling in engagement, fluency, and suggestion aspects. Moreover, Mistral's reduced inference time positions it as a promising candidate for real-time deployment, offering timely responses in emotionally sensitive interactions. This comparative analysis underscores the importance of considering multiple evaluation criteria to comprehensively assess the capabilities of language models in different contexts, thereby guiding informed decisions regarding model selection and deployment strategies.

Furthermore, the inference time analysis highlighted the consistent superiority of fine-tuned models over their base counterparts, showcasing substantially reduced processing times. This efficiency gain, coupled with the enhanced performance metrics, underscores the efficacy of fine-tuning and LoRA in optimizing language models for emotional support conversations. Additionally, the choice of sampling methods, particularly nucleus sampling, played a pivotal role in further improving inference efficiency by ensuring the selection of relevant tokens, contributing to the overall model effectiveness and responsiveness in real-time interactions.

In the realm of future research, one avenue of exploration involves the pruning of language

models [31], [32], and [33] to alleviate their size and memory footprint, all while mitigating potential performance losses. This approach holds promise for enhancing the scalability and deployability of language models, particularly in resource-constrained environments such as mobile devices or edge computing platforms. Additionally, there's a burgeoning interest in integrating multimodal capabilities, encompassing speech, and vision, into emotion support chatbots. By harnessing the complementary strengths of different modalities, such as text, speech, and visual cues, these advanced chatbots could offer more nuanced and empathetic interactions, thereby fostering deeper connections with users and enhancing the overall quality of emotional support provided.

In summary, the findings of this study underscore the potential of fine-tuning and context management techniques in enhancing the effectiveness and efficiency of language models for emotional support tasks. Continued research and experimentation in this direction hold promise for further advancements in natural language processing and conversational AI, ultimately benefiting individuals seeking emotional support through AI-driven interactions.

Bibliography

[1] Yanran Li et al. "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Greg Kondrak and Taro Watanabe. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 986–995. URL: https://aclanthology.org/I17-1099.

[2] Hannah Rashkin et al. "Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 5370–5381. DOI: 10.18653/v1/P19-1534. URL: https://aclanthology.org/P19-1534.

[3] Siyang Liu et al. *Towards Emotional Support Dialog Systems*. 2021. arXiv: 2106.01144 [cs.CL].

[4] Zhonghua Zheng et al. *Building Emotional Support Chatbots in the Era of LLMs*. 2023. arXiv: 2308.11584 [cs.CL].

[5] Qingyue Wang et al. *Recursively Summarizing Enables Long-Term Dialogue Memory in Large Language Models*. 2024. arXiv: 2308.15022 [cs.CL].

[6] Cangqing Wang et al. *Adapting LLMs for Efficient Context Processing through Soft Prompt Compression*. 2024. arXiv: 2404.04997 [cs.LG].

[7] Seyed Mahed Mousavi, Simone Caldarella, and Giuseppe Riccardi. "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, 2023. DOI: 10.18653/v1/2023.nlp4convai-1.1. URL: http://dx.doi.org/10.18653/v1/2023.nlp4convai-1.1.

[8] Saurav Pawar et al. *The What, Why, and How of Context Length Extension Techniques in Large Language Models – A Detailed Survey*. 2024. arXiv: 2401.07872 [cs.CL].

[9] Yang Liu et al. *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment*. 2023. arXiv: 2303.16634 [cs.CL].

[10] Lianmin Zheng et al. *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. 2023. arXiv: 2306.05685 [cs.CL].

[11] Guiming Hardy Chen et al. *Humans or LLMs as the Judge? A Study on Judgement Biases*. 2024. arXiv: 2402.10669 [cs.CL].

[12] Yen-Ting Lin and Yun-Nung Chen. *LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models*. 2023. arXiv: 2305.13711 [cs.CL].

[13] Jinlan Fu et al. *GPTScore: Evaluate as You Desire*. 2023. arXiv: 2302.04166 [cs.CL].

[14] Ning Ding et al. *Delta Tuning: A Comprehensive Study of Parameter Efficient Methods for Pre-trained Language Models*. 2022. arXiv: 2203.06904 [cs.CL].

[15] Zhaojiang Lin, Andrea Madotto, and Pascale Fung. *Exploring Versatile Generative Language Model Via Parameter-Efficient Transfer Learning*. 2020. arXiv: 2004.03829 [cs.CL].

[16] Zihao Fu et al. *On the Effectiveness of Parameter-Efficient Fine-Tuning*. 2022. arXiv: 2211.15583 [cs.CL].

[17] Lingling Xu et al. *Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment*. 2023. arXiv: 2312.12148 [cs.CL].

[18] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL].

[19] Tim Dettmers et al. *QLoRA: Efficient Finetuning of Quantized LLMs*. 2023. arXiv: 2305.14314 [cs.LG].

[20] Yaqing Wang et al. *AdaMix: Mixture-of-Adaptations for Parameter-efficient Model Tuning*. 2022. arXiv: 2205.12410 [cs.CL].

[21] Qidong Liu et al. *MOELoRA: An MOE-based Parameter Efficient Fine-Tuning Method for Multi-task Medical Applications*. 2023. arXiv: 2310.18339 [cs.CL].

[22] Chengsong Huang et al. *LoraHub: Efficient Cross-Task Generalization via Dynamic LoRA Composition*. 2024. arXiv: 2307.13269 [cs.CL].

[23] Longteng Zhang et al. *LoRA-FA: Memory-efficient Low-rank Adaptation for Large Language Models Fine-tuning*. 2023. arXiv: 2308.03303 [cs.CL].

[24] Clara Hill. *Helping skills: Facilitating exploration, insight, and action (5th ed.)*. Jan. 2020. ISBN: 1433831376. DOI: 10.1037/0000147-000.

[25] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL].

[26] Albert Q. Jiang et al. *Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL].

[27] Joshua Ainslie et al. *GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints*. 2023. arXiv: 2305.13245 [cs.CL].

[28] Ali Hassani et al. *Neighborhood Attention Transformer*. 2023. arXiv: 2204.07143 [cs.CV].

[29] Guangxuan Xiao et al. *Efficient Streaming Language Models with Attention Sinks*. 2024. arXiv: 2309.17453 [cs.CL].

[30] Marah Abdin et al. *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*. 2024. arXiv: 2404.14219 [cs.CL].

[31] Demi Guo, Alexander M. Rush, and Yoon Kim. *Parameter-Efficient Transfer Learning with Diff Pruning*. 2021. arXiv: 2012.07463 [cs.CL].

[32] Pavlo Molchanov et al. *Importance Estimation for Neural Network Pruning*. 2019. arXiv: 1906.10771 [cs.LG].

[33] Victor Sanh, Thomas Wolf, and Alexander M. Rush. *Movement Pruning: Adaptive Sparsity by Fine-Tuning*. 2020. arXiv: 2005.07683 [cs.CL].