

# Towards Enhancing AI-Driven Mental Health Support with an Intelligent Counsellor Agent

Thesis submitted in partial fulfillment  
of the requirements for the degree of

*Master of Science*  
*in*  
***Computer Science and Engineering***  
*by Research*

by

Nirmal Manoj Chathayil

2019111011

nirmal.manoj@research.iiit.ac.in



International Institute of Information Technology

Hyderabad - 500 032, INDIA

June 2024

Copyright © Nirmal Manoj Chathayil, 2024  
All Rights Reserved

International Institute of Information Technology  
Hyderabad, India

## CERTIFICATE

It is certified that the work contained in this thesis, titled “**Towards Enhancing AI-Driven Mental Health Support with an Intelligent Counsellor Agent**” by **Nirmal Manoj Chathayil**, has been carried out under my supervision and is not submitted elsewhere for a degree.

---

Date

---

Advisor: Prof. Vasudeva Varma

---

Date

---

Co-advisor: Dr. Manish Gupta

*To my beloved family.*

## Acknowledgments

The years I have spent at IIIT have been both transformative and unforgettable. This thesis, like all my pursuits here, has been shaped by collaboration and support from many within the institute. I am profoundly grateful for the opportunity to grow and excel in such a nurturing and intellectually stimulating environment.

First and foremost, I wish to express my profound gratitude to my advisor, Prof. Vasudeva Varma, whose inspirational guidance has been instrumental in shaping my academic pursuits. Prof. Vasu's professional demeanor and thought-provoking insights during our weekly lab meetings have not only encouraged me to think more deeply but also to broaden my research ambitions. His belief in my potential and his generous provision of academic freedom have been fundamental to the success of my thesis. I am deeply thankful for the opportunity to work in such an exceptional lab environment, surrounded by dedicated and talented researchers. Prof. Vasu's influence will undoubtedly remain a pivotal force in my pursuit of innovation in future endeavors. I am eternally grateful for his mentorship and the opportunities provided to me.

Secondly, I want to express my deep gratitude to my co-advisor, Dr. Manish Gupta, whose guidance was monumental to most of the work in this thesis. The velocity at which he operates and propels the team to make progress, along with his bias towards action, are qualities that I strive to incorporate. His dedication to the field and meticulous attention to detail have always inspired me. I cannot thank him enough for his patience, guidance, contributions, and the life lessons he has imparted to me. I could not have asked for a better mentor to work with.

I am particularly grateful to Dr. Kiranmayi S. Bapi, who has shared her extensive experience as a counselor and her understanding of psychology. Her valuable insights were instrumental in informing the discussions in Chapter 5 of this thesis. Additionally, her expertise was crucial in clarifying doubts and ensuring the validity of our approach from a psychological perspective. Moreover, Dr. Kiranmayi's assistance also reinforced the interdisciplinary integrity of our work, helping to ground our methods in psychological principles.

I am immensely grateful to all my extraordinary lab mates at iREL. I would like to extend special thanks to Sagar, who has been not only a fantastic mentor and teammate but also a great friend. His work ethic and readiness to discuss topics related to language technologies have been truly inspirational. I am equally thankful to Tathagata, whose excellent mentorship was crucial in helping me navigate the beginnings of my research at iREL. My appreciation also goes to Pavan for the wonderful discussions on technology, computer science, and for engaging in spirited political debates. I must acknowledge Bhavyajeet for his guidance on both academic and non-academic questions. A special mention goes to Sravani, whose work and ideas sparked the fundamental idea behind a significant portion of my thesis. I am thankful to Himanshu, a senior I have admired since my first year, who encouraged me to join iREL. I am also grateful to

the senior researchers in the lab—Dhaval, Tushar, Harika, Vijay, Shivprasad, Anubhav, Tanmay, Anshul, and Savita—for their willingness to offer help generously whenever needed.

My deep gratitude extends to my co-authors and collaborators, Sidharth Giri, Ankita Maity, Triansh Sharma, Harshit Gupta, and Manav Chaudhary, whose hard work and enthusiasm were indispensable to this thesis. I am also thankful to my peers — Aditya, Gokul, and Shivansh for their insightful discussions on NLP and for providing good company throughout my research journey.

I wish to express my profound gratitude to Apoorva, a cherished friend whose presence in my life at IIIT has been immensely uplifting. Thank you, Apoorva, for the joyful moments that have made my journey here truly special. Your caring and compassionate nature has not only comforted me but has also played a significant role in my personal growth. It has been an honor to learn from you and to share this path with such an inspiring friend.

Special thanks to Alapan and Pavani for being exceptional friends since our first year. Both of you have inspired me in countless ways and have significantly contributed to my personal growth. I cherish our long discussions on a wide range of topics and all the joyful times we have shared. I am also grateful to Arathy, Naman, Sharadha, and Hasvitha for the wonderful times we have shared and for your multifaceted support over the years. I consider myself fortunate to have such an incredible group of individuals as friends.

Additionally, I want to express my appreciation to Arjun, Gokul, Viswanadh, Bharath, Vivek, Pranoy, Rhuthik, Hrishi, Jerrin, Ivin, Sivaprasad, Rahul and Aswin for being such great friends. Your camaraderie and support have immeasurably enriched my life.

Finally, I wish to express my profound gratitude to my family. To my late paternal grandparents, Sekharan and Sathyavathi, and to my maternal grandmother, Sethulakshmi, who have all instilled in me a strong moral compass through their wisdom, teachings, and love. Their legacy continues to give me strength and guide me forward. To my parents, Manoj and Resmi, who have always been pillars of strength, providing me with unconditional love and support. My father, through his selfless efforts as a social worker, has shown a deep and caring commitment to everyone in his life. My mother’s hard work and dedication have both inspired me and actively propelled my growth. I could not have achieved this without both of you, and words cannot express my gratitude for all that you have done. To my little sister, Naipunnya, who brings joy and laughter into my life; her presence was especially crucial during the lockdown, and I cannot imagine those times without her. Lastly, a heartfelt thank you to my extended family, for always providing me with support and for wishing me well in all my pursuits.

## Abstract

Recent strides in the development of conversational models for goal-oriented applications present substantial opportunities for augmenting human expertise and automating aspects of professional tasks. Among the most critical areas where such advancements could be transformative is the domain of mental health support—an arena of paramount importance to families, businesses, enterprises, and society at large. Mental health disorders present a substantial global burden, with over 50% of adults experiencing mental health issues at some point in their lives. Despite the widespread prevalence of these conditions, a staggering 70% of affected individuals worldwide do not receive treatment from professional healthcare providers. From 2011 to 2030, it is estimated that the global economic impact of mental disorders will amount to a loss of *US \$16.3 trillion* in cumulative output, surpassing the economic losses attributed to cancer, chronic respiratory conditions, and diabetes. Mental health disorders among working-age adults are another pressing concern, with 15% of this population affected and a staggering 12 billion working days lost annually to depression and anxiety alone. The World Health Organization underscores the critical need for practical mental health support in occupational settings, highlighting its significant role in boosting individual’s confidence, productivity, and the capacity to maintain or secure employment.

In this thesis, we explore how AI-driven solutions, such as chatbots and tools to enhance responses by mental health professionals, have the potential to provide accessible and efficient support for mental health issues. Unfortunately, prior work in this space is significantly limited due to the lack of quality data for training, with no existing large-scale conversational dataset present. To address this, we create the first large-scale mental health conversational dataset (BBMH) with over *36,000 threads* of conversations sourced from a popular mental health forum. Further, to tackle productivity loss at the workplace, we introduce the task of building classifiers to filter workplace-related mental health posts. We create weak-labelers with pretrained language models and experiment with zero-shot prompting on state-of-the-art large language models to filter out a high-quality dialog dataset for mental health support at the workplace (BBWMH) with *5,240 threads* and *49,238 utterances*. We also build strong baseline dialog models using state-of-the-art large language models such as Gemma, Mistral, and Llama-2 for workplace-related mental health support.

In addition to our work on mental health support, this thesis presents several methods by which artificial intelligence can be leveraged for social good, aligning with the objectives of the *Project Angel* initiative by IREL and IIIT Hyderabad. As part of our contributions to natural language understanding, we conduct an in-depth study of fine-grained hierarchical sexism classification. Our findings demonstrate the effectiveness of domain-specific pretraining, the application of focal loss, and multi-level training methodologies in addressing this task.

Insights gained from this study have been instrumental in developing a workplace-related mental health post classifier, where we also address substantial class imbalance within the dataset.

Mental health support is a highly complex task with a substantial body of psychological research accumulated over many decades, *especially* when help-seekers have mental health disorders. However, emotional distress, a common experience among the population *regardless* of the presence of diagnosed mental disorders, has been identified as a key area where intervention can be effective. Recent research highlights the benefits of Emotional Support Conversations (ESC) as a method to alleviate such distress and improve mental health outcomes. Previous methods for ESC condition on the conversational context to first predict a single support strategy and then generate an Emotional Support (ES) response. Unfortunately, these studies fail to imitate the intricate art of (a) dynamically understanding the evolving problem category distribution of the help-seeker and (b) responding using a combination of strategies known to be effective in addressing the predicted problem distribution.

To address these limitations, we develop a *psychologically grounded* framework, Problem Identification and Strategy Matching for ESC (PRISM), which leverages Transformer-based encoder-decoder models to predict problem distribution (over ten types) and strategy distribution (over eight types) and generates effective Emotional Support responses. Unlike existing models that utilize coarse-grained strategy annotations and conversation-level emotion labels, training PRISM models requires fine-grained, probabilistic annotations for both support strategies and underlying problems. Hence, we contribute a novel dataset, ESConv++, curated using problem and strategy annotations from GPT-4. PRISM models are trained using ESConv++ to optimize four novel loss functions designed specifically for the ESC task. We conduct extensive ablation studies and human and automatic evaluations to evaluate our model and understand its various components. Our proposed model, PRISM, outperforms strong baselines on automated metrics by large margins and also performs best on human evaluation.



# Contents

Chapter	Page
1 Introduction . . . . .	1
1.1 Motivation . . . . .	1
1.2 Overview of Challenges and Methods . . . . .	3
1.3 Contributions . . . . .	4
1.4 Thesis Organisation . . . . .	5
2 Related work . . . . .	7
2.1 Sexism Detection and Classification . . . . .	7
2.1.1 Detection and Engagement Studies . . . . .	7
2.1.2 Categorization and Classification Approaches . . . . .	8
2.1.3 Self-Training and Multi-Label Classification . . . . .	8
2.1.4 Addressing Gaps in Hierarchical Sexism Classification . . . . .	8
2.2 Conversational Agents for Mental Health Support . . . . .	9
2.3 Emotional Support Conversational Agents . . . . .	9
2.3.1 Emotional Response Generation . . . . .	10
2.3.2 Empathetic Response Generation . . . . .	10
2.3.3 Emotional Support Conversation . . . . .	11
2.3.4 Commonsense Knowledge for NLP . . . . .	12
3 Multi-level Training for Explainable Detection of Online Sexism . . . . .	13
3.1 Overview . . . . .	13
3.2 Introduction . . . . .	13
3.3 Dataset . . . . .	14
3.4 Methodology . . . . .	16
3.4.1 Basic Supervised Finetuning . . . . .	16
3.4.2 Focal Loss . . . . .	17
3.4.3 Domain-Adaptive Pretraining . . . . .	18
3.4.4 Multi-level Training . . . . .	18
3.5 Experimental Details . . . . .	19
3.5.1 Modeling . . . . .	19
3.5.2 Training Hyperparameters . . . . .	19
3.6 Results . . . . .	20
3.7 Summary . . . . .	21
4 Dialog Dataset and Classifier for Workplace-related Mental Health Support . . . . .	23
4.1 Overview . . . . .	23
4.2 Introduction . . . . .	23
4.3 Dataset . . . . .	24
4.4 Workplace Classifier for Mental Health Posts . . . . .	26
4.4.1 Defining Workplace-Related Mental Health Posts . . . . .	26

4.4.2	Methodology for Developing the Workplace Classifier . . . . .	27
4.4.3	ChatGPT annotation for Workplace Posts . . . . .	29
4.4.4	Workplace Classifier Conclusion . . . . .	30
4.5	Dialog Model for Workplace-related Mental Health Support . . . . .	31
4.5.1	Evaluation Metrics . . . . .	31
4.5.2	Results . . . . .	31
4.6	Summary . . . . .	32
5	PRISM: Problem Identification and Strategy Matching for Emotional Support Conversations	34
5.1	Overview . . . . .	34
5.2	Introduction . . . . .	34
5.3	ESC Problem Formulation . . . . .	37
5.4	Brief Description of Baselines . . . . .	38
5.5	ESConv++ Dataset . . . . .	39
5.5.1	Annotation Process . . . . .	41
5.5.1.1	Definitions of Strategies . . . . .	41
5.5.1.2	Definitions of Problem Categories . . . . .	42
5.5.2	Annotation Methodology . . . . .	43
5.6	Why do we need Problem Annotations? . . . . .	43
5.7	PRISM Model Architecture . . . . .	45
5.7.1	Stage 1: Learning Problem-Strategy Mapping . . . . .	45
5.7.2	Stage 2: Training the Dialog Model . . . . .	47
5.8	Experiments and Results . . . . .	48
5.8.1	Metrics . . . . .	48
5.8.2	Main results . . . . .	48
5.8.3	Analysis of Resource Requirements for the Deployment of Baselines compared to PRISM . . . . .	49
5.8.4	Performance of Finetuned Large Language Models in ESC . . . . .	50
5.8.5	Ablations . . . . .	51
5.8.6	Decoding Hyperparameter Comparisons with Baselines . . . . .	52
5.8.7	Human Evaluation Results . . . . .	53
5.9	Summary . . . . .	54
6	Conclusion and Future work . . . . .	55
6.1	Future Work . . . . .	57
	<i>Appendix A: Multi-level Training for Explainable Detection of Online Sexism</i> . . . . .	59
A.1	Taxonomy of Sexism Classification and Definition . . . . .	59
	<i>Appendix B: PRISM: Problem Identification and Strategy Matching for Emotional Support Conversations</i> . . . . .	62
B.1	Experimental Setup and Hyper-parameters for Reproducibility . . . . .	62
B.2	Case Studies . . . . .	63
B.2.1	Case Study 1 . . . . .	63
B.2.2	Case Study 2 . . . . .	64
B.2.3	Case Study 3 . . . . .	65
B.3	Prompt Templates . . . . .	66
B.3.1	Strategy Annotation User Message . . . . .	70
B.3.2	Problem Annotation System Message . . . . .	70
B.3.3	Problem Annotation User Message . . . . .	73

## List of Figures

Figure	Page
3.1 Length distributions of the comments in the labelled dataset . . . . .	16
3.2 Length distributions of the comments in the unlabelled Reddit Dataset . . . . .	16
3.3 Length distributions of the comments in the unlabelled Gab Dataset . . . . .	16
4.1 Number of threads from each sub-forum in the BBMH dataset. Labels correspond to the names of the sub-forums. . . . .	25
4.2 Conversation tree structure in Beyond Blue forums. . . . .	25
5.1 An example conversation from ESConv++ with help-seeker (left) and supporter (right). Problem and support strategy distributions are shown in red and green, respectively, with categories ordered as in 5.5.1. Both help-seeker utterances have emotion 'sadness'. . . . .	35
5.2 PRISM: Model Architecture . . . . .	45

## List of Tables

Table	Page	
3.1	Distribution of class labels across tasks and data splits. Adapted from [30] . . . . .	15
3.3	Performance of our models on the validation set of Task A. . . . .	20
3.4	Performance of our models on the validation set of Task B. . . . .	21
3.5	Performance of our models on the validation set of Task C. . . . .	22
4.1	Response Statistics by Rank . . . . .	26
4.2	Distribution of Positive and Negative Classes in Train and Test Sets . . . . .	27
4.3	Model Performance Comparison . . . . .	28
4.4	Precision and Recall at Various Checkpoints with STLR and Weight Decay . . . . .	28
4.5	Performance of fine-tuned models on the BBWMH dataset . . . . .	32
5.1	Strategy Distributions in ESConv++ Dataset. Top shows strategy distribution across 8 strategies per turn averaged across conversations for the first 10 tens. Bottom shows averaged strategy distribution across the entire dataset. . . . .	37
5.2	Problem Distributions in ESConv++ Dataset. Top shows problem distribution across 10 strategies per turn averaged across conversations for the first 10 tens. Bottom shows averaged problem distribution across the entire dataset. . . . .	38
5.3	ESConv++: Dataset Statistics . . . . .	40
5.4	Quantitative analysis of problems within ESConv++ dialogues. . . . .	44
5.5	Problem-Strategy Mapping in ESConv++ . . . . .	46
5.6	Comparison of various methods for ESC problem on ESConv++. Empty cells indicate that those results were not available. . . . .	48
5.7	Oracle Analysis . . . . .	49
5.8	Comparison of models on various metrics. . . . .	51
5.9	Ablation analysis for PRISM, with hyperparameters uniformly set at top-k=40 and top-p=0.8 across all configurations. This uniformity ensures a direct comparison of component impacts on PRISM’s performance, highlighting the sensitivity and contributions of individual elements. . . . .	52
5.10	Ablations analysis for PRISM: Comparing BART and Blenderbot . . . . .	52
5.11	Comparing PRISM and KEMI with BART backbone, with hyperparameters uniformly set at top-k=30 and top-p=0.3. . . . .	52
5.12	Human Evaluation Results (PRISM vs KEMI). Overall, PRISM had <b>64% wins</b> , 17% ties and just 19% losses over KEMI. . . . .	53
A.1	Taxonomy of sexism categories (Task B) and fine-grained vectors (Task C). Kirk et al. [30] . . . . .	59
B.1	Hyper-parameters for training . . . . .	62
B.2	Hyper-parameters for inference . . . . .	63

## Chapter 1

### Introduction

In this thesis, we examine the application of artificial intelligence to three critical areas: the detection of online sexism, mental health support, and emotional support through conversational agents. This introductory chapter outlines the motivations for addressing these distinct but interconnected challenges. It follows with a detailed exploration of the specific problems each area presents. The chapter concludes by summarizing the key contributions of this thesis and providing an overview of the structure and progression of the discussions in the subsequent chapters.

The overarching framework of this research aligns with the objectives of the *Project Angel* initiative by IREL and IIIT Hyderabad, which seeks to leverage AI technologies to tackle pressing societal challenges. The initiative emphasizes the development of innovative AI-driven solutions that can improve societal well-being, focusing on issues such as mental health, gender equality, and social inclusion. By integrating AI with social good initiatives like Project Angel, we aim to create impactful solutions that not only advance the state of technology but also address key social issues, promoting a more equitable and supportive society.

#### 1.1 Motivation

Recent advancements in the fields of Artificial Intelligence (AI) and Natural Language Processing (NLP) have unlocked a myriad of possibilities for enhancing human expertise and automating various facets of professional tasks across diverse domains. However, despite the progress made, there remain several fundamental challenges that necessitate further research and domain-specific innovation to fully harness the potential of these technologies before deployment. This section elucidates the underlying motivations driving the research presented in this thesis, emphasizing the significance of addressing these key problems and the potential impact of the proposed solutions.

The proliferation of online platforms has led to an unprecedented surge in user-generated content, fostering vibrant digital communities and facilitating global communication. However, this growth has also been accompanied by the spread of harmful content, particularly online sexism targeted towards women. The pervasive nature of sexist content not only creates hostile online environments but also perpetuates deeply ingrained societal biases and gender-based discrimination. Traditional approaches to identifying and moderating sexist content have relied

on simplistic, binary classification models that lack the nuance and explainability necessary to effectively address the complexities of online sexism. Moreover, the absence of explainable categorization hinders the development of targeted interventions and undermines trust in automated moderation systems. The incorporation of explainability enables a deeper understanding of the underlying patterns and characteristics of sexist language, facilitating more effective moderation strategies and fostering healthier online discourse.

Mental health disorders impose a significant global burden, with more than half of the adult population encountering mental health challenges at some point in their lives [49]. Despite their prevalence, an alarming 70% of those affected globally do not seek or receive care from professional healthcare services [23]. Economically, the impact of mental disorders is anticipated to be devastating, with projected losses reaching approximately *US \$16.3 trillion* in total output from 2011 to 2030, surpassing the financial impacts associated with cancer, chronic respiratory diseases, and diabetes [69]. The prevalence of mental health issues among adults of working age presents a serious concern, affecting about 15% of this demographic. This results in a considerable loss of productivity, with around 12 billion workdays lost each year due to depression and anxiety [77]. The World Health Organization underscores the critical need for effective mental health interventions in workplace settings, highlighting their significant role in enhancing individuals' confidence and productivity, as well as their crucial impact on helping people secure or retain employment [77].

AI-driven solutions, such as chatbots and tools to enhance responses by mental health professionals, have the potential to provide accessible and efficient support for mental health issues. However, prior work in this space is significantly limited due to the lack of quality data for training, with no existing large-scale conversational dataset available. Addressing this gap is crucial for developing effective dialog agents that can offer much-needed support to individuals facing mental health challenges, particularly in the workplace setting.

Emotional distress, a common experience among the population regardless of the presence of diagnosed mental disorders, has been identified as a key area where intervention can be effective. Recent research highlights the benefits of Emotional Support Conversations (ESC) as a method to alleviate such distress and improve mental health outcomes [6, 16]. The quest for emotional support traditionally involves seeking assistance through social interactions, though not everyone has consistent access to support from family and friends [17, 31, 72]. This gap underscores the growing interest in developing ESC systems, which are designed to offer timely emotional support, potentially reaching a wider audience who lacks traditional support networks.

ESC agents have the potential to provide accessible, round-the-clock emotional support to individuals in need, complementing traditional mental health services and social support networks. By leveraging advancements in natural language processing and machine learning, these systems can engage in empathetic conversations, identify the underlying problems faced by help-seekers, and provide tailored support strategies. The development of effective ESC agents requires a deep understanding of the dynamics of emotional support conversations, the ability to adapt to the evolving needs of help-seekers, and the integration of domain knowledge from psychology and counseling.

## 1.2 Overview of Challenges and Methods

This thesis addresses several critical challenges in the domains of explainable detection of online sexism, mental health support, and emotional support conversation. These challenges span across various aspects, including data scarcity, class imbalance, domain-specific language, and the need for explainable and psychologically grounded models.

One of the primary challenges encountered in the domain of mental health support conversation is the lack of high-quality, large-scale datasets. The scarcity of large-scale conversational datasets hinders the training and evaluation of dialogue agents. To address these challenges, we create novel datasets, such as the Beyond Blue Mental Health (BBMH) and Beyond Blue Workplace Mental Health (BBWMH) datasets, which provide rich resources for training and evaluating AI-driven solutions.

Another significant challenge is the inherent class imbalance prevalent in the datasets, particularly in the fine-grained classification tasks of online sexism detection. To mitigate this issue, we explore techniques such as focal loss, which helps in addressing the class imbalance and improving model performance.

Domain-specific language and vocabulary also pose challenges across all three domains. In the context of online sexism detection, the language employed often differs significantly from the general language encountered during the pretraining phase of language models. To tackle this, we investigate domain-adaptive pretraining techniques to adapt pretrained models to the unique linguistic characteristics of the target domain.

Explainability is a crucial aspect in the detection of online sexism and emotional support conversation. Developing models that can provide interpretable insights into their decision-making process is essential to foster trust and facilitate error analysis. In the context of emotional support conversation, incorporating domain knowledge from psychology and counseling into the agent’s language understanding and generation capabilities is vital for generating psychologically grounded and beneficial responses.

To address these challenges, we propose a novel framework, PRISM (Problem Identification and Strategy Matching for ESC), which leverages Transformer-based encoder-decoder models to predict fine-grained problem and strategy distributions at each turn of the conversation. The framework is trained using a novel dataset we introduce, ESConv++, which is curated using problem and strategy annotations from GPT-4. The use of GPT-4 for probabilistic annotation is a novel approach that allows for more nuanced and contextually relevant labeling of the underlying problems and appropriate support strategies.

In the following part of this section, we provide a brief overview of the domain-wise challenges and methods employed in this thesis.

**Explainable Detection of Sexism:** The primary challenges in this domain include the domain-specific language, class imbalance, and the hierarchical labeling scheme of the dataset. To address these challenges, this thesis explores a combination of techniques, including domain-adaptive pretraining, focal loss, and multi-level training. By adapting pretrained transformer models to the specific domain of online sexism, incorporating loss functions that mitigate class imbalance, and leveraging the hierarchical structure of the classification tasks, we aim to de-

velop robust and explainable models for the detection of sexist content.

**Mental Health Dialog Agents:** The primary challenge in this domain is the scarcity of high-quality, large-scale conversational datasets. To address this, we create the Beyond Blue Mental Health (BBMH) dataset and the Beyond Blue Workplace Mental Health (BBWMH) dataset. These datasets provide rich resources for training and evaluating mental health dialog agents, enabling them to learn from real-world conversations and capture the diverse range of topics, concerns, and support strategies employed in responses to help-seekers. We also explore the task of building classifiers to filter workplace-related mental health posts and propose strong baseline dialog models using state-of-the-art large language models such as Gemma [67], Mistral [28], and Llama-2 [68].

**Emotional Support Conversation Agents:** The primary challenges in this domain include the dynamic nature of the help-seeker’s emotional state, the incorporation of domain knowledge from psychology and counseling, and the need for psychologically grounded responses. To address these challenges, we propose a novel framework called PRISM, which leverages Transformer-based encoder-decoder models to predict fine-grained problem and strategy distributions at each turn of the conversation. The framework is trained using a novel dataset, ESConv++, curated using problem and strategy annotations from GPT-4. We also introduce four ESC-specific loss functions to ensure that the agent learns to accurately identify the help-seeker’s problems, select appropriate support strategies, and generate relevant and effective responses.

### 1.3 Contributions

The primary contributions of this thesis, detailed in subsequent chapters, are as follows:

1. We study fine-grained hierarchical sexism classification and demonstrate the effectiveness of domain-specific pretraining, focal loss application, and multi-level training methodologies on existing datasets sourced from Reddit and Gab. Our approach addresses the challenges of class imbalance and domain-specific language, enabling more accurate and explainable detection of sexist content.
2. We introduce two novel datasets: the Beyond Blue Mental Health (BBMH) containing *36,240 conversation threads*, and the meticulously filtered Beyond Blue Workplace Mental Health (BBWMH), containing *5,240 conversation threads* and *49,238 utterances*. These datasets fill a critical gap in high-quality, large-scale conversational data for mental health support research, particularly within workplace contexts, setting a new benchmark for future AI-driven mental health support studies.
3. We explore the task of workplace-related mental health support and propose strong baseline dialog models by fine-tuning state-of-the-art language models (Llama-2-7b, Gemma-2b, and Mistral-7b) on the BBWMH dataset. Our work paves the way for the development of AI-driven solutions that can provide accessible and efficient mental health support for workplace related settings.



4. We develop a comprehensive taxonomy of problems specific to the ESConv dataset, aligning with the general helping skills required for Emotional Support Conversation (ESC) as outlined by Hill [24]. This taxonomy lays the foundation for our proposed ESC framework, PRISM, which incorporates dynamic problem and strategy distributions to generate effective emotional support responses. We also contribute a novel dataset, ESConv++, with problem and strategy distribution annotations per utterance, addressing the lack of fine-grained, probabilistic annotations in existing ESC datasets.
5. We propose a psychologically grounded architecture that follows the PRISM framework, utilizing transformer encoder-decoder models with 4 ESC-specific loss functions: Problem prediction loss, Strategy prediction loss, Problem-Strategy mapping loss, and Cross-entropy loss. Our approach outperforms state-of-the-art methods on key metrics, achieving improvements of 2.94 on BLEU-1, 1.58 on BLEU-2, and 0.89 on METEOR.

These contributions collectively advance the field of NLP and AI in understanding and addressing complex societal and mental health challenges.

## 1.4 Thesis Organisation

This thesis is structured into six main chapters, along with relevant appendices that provide supplementary information and details. Here is a brief overview of each:

1. **Chapter 1** (Introduction) presents the motivation, problem statement, and major contributions of the thesis.
2. **Chapter 2** (Related Work) presents a survey of the prior literature related to the problems explored in this thesis, covering the domains of explainable detection of online sexism, mental health support, and emotional support conversation.
3. **Chapter 3** focuses on the explainable detection of online sexism, studying fine-grained hierarchical sexism classification and demonstrating the effectiveness of domain-specific pre-training, focal loss application, and multi-level training methodologies on existing datasets sourced from Reddit and Gab.
4. **Chapter 4** introduces two novel datasets, the Beyond Blue Mental Health (BBMH) and the Beyond Blue Workplace Mental Health (BBWMH), addressing the scarcity of high-quality, large-scale conversational data for mental health support. It also explores the task of workplace-related mental health support and proposes strong baseline dialog models by fine-tuning state-of-the-art large language models on the BBWMH dataset.
5. **Chapter 5** focuses on the task of emotional support conversation (ESC), developing a novel taxonomy of problems specific to the ESConv dataset and proposing the PRISM framework, which incorporates dynamic problem and strategy distributions to generate effective emotional support responses. It also introduces the ESConv++ dataset with problem and strategy distribution annotations per utterance and proposes an architecture that follows the PRISM framework, outperforming state-of-the-art methods.

6. **Chapter 6** serves as the concluding chapter of the thesis, offering a summary of the covered work and exploring potential avenues for future expansion and development based on the findings.

Appendices A and B provide supplementary material related to the explainable detection of sexism and emotional support conversation tasks, respectively.

## *Chapter 2*

### **Related work**

This chapter presents an overarching survey of the prior literature related to the problems explored in this thesis, covering the domains of explainable detection of online sexism, mental health support, and emotional support conversation. We begin by discussing the various approaches and methodologies employed in sexism detection and classification, highlighting the gaps in hierarchical sexism classification that our research aims to address. Next, we look into the limited but growing field of conversational agents for mental health support. Finally, we provide an in-depth exploration of emotional support conversational agents, exploring related topics such as emotional and empathetic response generation, and the current state of research in this field.

#### **2.1 Sexism Detection and Classification**

The detection and classification of online sexism have garnered significant attention in recent years, with researchers employing various approaches to tackle this complex issue. These approaches range from self-training and multi-label classification to detection and engagement studies, as well as diverse categorization and classification techniques. In this section, we provide an overview of the key developments in this field and discuss how our research contributes to addressing the gaps in hierarchical sexism classification.

##### **2.1.1 Detection and Engagement Studies**

Early works in the field of sexism detection focused on identifying instances of sexism and analyzing user engagement with such content. ElSherief et al. [13] identified a significant increase in engagement with tweets that deal with gender-based violence, in contrast to those that do not. Similarly, Zhang and Luo [82] proposed the application of deep neural networks for the effective extraction of features that encapsulate the core of hate speech, encompassing racism and sexism. Frenda et al. [14] investigated the identification of sexism and misogyny in tweets formulated in both Spanish and English, while Schrading et al. [61] aimed to develop classifiers for pinpointing discussions around domestic abuse on platforms such as Reddit.

Building upon these foundational studies, researchers began to explore more nuanced approaches to sexism detection and classification, as discussed in the following subsections.

### 2.1.2 Categorization and Classification Approaches

As the field of sexism detection progressed, researchers started to investigate various categorization and classification approaches to better understand the diverse manifestations of sexism. Karlekar and Bansal [29] investigated the deployment of a combined RNN-CNN architecture to categorize personal accounts of sexual harassment into up to three distinct categories. Similarly, Anzovino et al. [2] classified tweets deemed misogynistic based on various dimensions, including discredit, stereotype and objectification, dominance, derailing, sexual harassment, and threats of violence, employing features like n-grams, part of speech tags, and text embeddings.

As we move from exploring basic categorization and classification architectures, we now turn to the potential of self-training and multi-label classification methodologies to enhance the depth and accuracy of sexism detection systems.

### 2.1.3 Self-Training and Multi-Label Classification

Recent studies in sexism detection and classification have explored leveraging self-training methodologies and multi-label classification to improve the performance and granularity of sexism detection models. Abburi et al. [1] utilized self-training methodologies to expand their labeled data pool by selectively incorporating unlabeled samples, with a focus on the multi-label classification of sexism. Their work diverged from the straightforward categorization of sexist content by examining self-reported instances of sexism. They presented a comprehensive classification system with 23 distinct sexism categories, arranged within a three-level hierarchical structure. This structure begins with 8 primary categories, extending to 15 at the second level, and reaching 23 at the third level, providing a nuanced framework for analyzing the complex nature of sexism.

While these studies have made significant progress in sexism detection and classification, there remains a notable gap in the exploration of *online* sexist content through a hierarchical taxonomical perspective, particularly in the nuanced categorization of such content. Our research aims to address this gap by employing methodologies specifically designed for the hierarchical structuring of sexism, thereby enhancing the analytical framework in this domain.

### 2.1.4 Addressing Gaps in Hierarchical Sexism Classification

Building upon the foundational role of existing taxonomies, our study contributes to the nuanced understanding and categorization of online sexism by leveraging methodologies such as multilevel training, domain-adaptive pretraining, and the application of focal loss objectives. These approaches not only tackle class imbalances but also improve our model’s adaptability and performance in detecting and classifying nuanced forms of sexism. By addressing the gaps in hierarchical sexism classification, our work paves the way for more effective and nuanced detection and moderation of online sexist content, contributing to the development of healthier and more inclusive online communities.

## 2.2 Conversational Agents for Mental Health Support

While there is a growing interest in employing artificial intelligence for mental health support, there remains a notable scarcity in the development of effective conversational agents that can handle the complex and nuanced needs of mental health conversations. Among the limited works in this field, one prominent study is PsyQA [65]. In this work, researchers developed a dataset comprising data pairs formatted as (user\_post - moderated\_response) and explored the syntactic, semantic, and linguistic characteristics of the responses. The source of data for PsyQA is derived from a Q&A forum of Yixinli<sup>1</sup>, a Chinese mental wellness service platform. The data was annotated following the 6-class strategy taxonomy provided by Hill [24] and the study experimented with generating strategy-conditioned responses using GPT-2.

While PsyQA provides valuable insights into the use of AI for mental health support, it differs from our work in several key aspects. Firstly, we focus specifically on workplace-related mental health issues, which are a critical concern given the prevalence and impact of mental health disorders among working-age adults. Secondly, PsyQA is a question-answer dataset without multi-turn conversations, whereas our datasets include extensive posts and multi-turn responses, capturing the nuances of mental health support conversations. Lastly, we address the challenge of modeling long-range context by employing advanced language models capable of handling such contexts.

The strategy taxonomy provided by Hill [24], which was used in PsyQA, is also relevant to our work. This taxonomy categorizes responses into six classes: interpretation, direct guidance, disclosure, restatement, information, and others. While we do not explicitly annotate our data using this taxonomy, it serves as a useful framework for understanding the various types of responses in mental health support conversations. Hill [24] and strategies inspired by their literature will be used in great detail in our Chapter 5 on developing Emotional Support Conversation systems.

## 2.3 Emotional Support Conversational Agents

Mental health disorders present a substantial global burden, with over 50% of adults experiencing mental health issues at some point in their lives [49]. Despite the widespread prevalence of these conditions, a staggering 70% of affected individuals worldwide do not receive treatment from professional healthcare providers [23]. Recent research [6, 16] highlights the benefits of Emotional Support Conversations (ESC) as a method to alleviate emotional distress and therefore improve mental health outcomes. Emotional support conversational agents have emerged as a promising approach to address the global burden of mental health disorders and the widespread lack of access to professional treatment. In this section, we provide an in-depth exploration of the current state of research in emotional support conversational agents, covering topics such as emotional response generation, empathetic response generation, and the specific challenges and advancements in emotional support conversation.

---

<sup>1</sup><https://www.xinli001.com/qa>

### 2.3.1 Emotional Response Generation

The rise in popularity and success of dialogue systems has catalyzed research towards enabling these systems to generate responses with specific, appropriate emotions, aiming for more human-like interactions. This research direction, known as emotional response generation, involves both generation-based [62, 88, 92] and retrieval-based methods [46, 56] to embed specific emotions into dialogue responses.

Prior research in the domain of emotional dialogue systems primarily utilized conversation-wide emotion labels to infer user feelings, which led to the generalization of emotional contexts as uniform and unchanging throughout the dialogue [33, 39, 47, 57]. This method does not account for the dynamic nature of user emotions, which may fluctuate in intensity even if the overall emotional state appears stable. Recognizing this limitation, researchers have explored utilizing empathetic response generation systems for ESC, which we discuss in the following subsection.

### 2.3.2 Empathetic Response Generation

Developing a dialogue system with warmth involves integrating it with empathetic capabilities [57]. This includes not only affective empathy, which focuses on recognizing the user’s emotions [33, 34, 38, 47], but also cognitive empathy, which entails grasping the user’s context [59, 85, 91], and persona, referring to the social identity one projects [87]. These elements operationalize the psychological concept of empathy within the system.

Early endeavors in empathetic conversation systems aimed at incorporating empathy to recognize and comprehend the emotional states of users. Nonetheless, the provision of empathetic reactions within individual parts of the conversation, as seen in these systems [57], is not adequate for Emotionally Supportive Conversational agents (ESCs). The practice of merely offering empathetic feedback during separate exchanges might neglect the continuous changes and feedback of the user across extended dialogues. Thus, it is crucial to develop a nuanced appreciation of the user’s emotional condition for precise user representation. Similar to a therapist’s increasing understanding of a client’s issues through the course of a counseling session, which enhances as more insights are shared by the individual seeking help, conversational agents must also refine their perception of the user’s emotional state for improved support and interaction.

However, current approaches on Empathetic Response Generation [33, 34, 38, 47, 59, 85, 91] predominantly addresses immediate emotional mirroring without a strategic focus on alleviating the user’s distress. Such systems often limit themselves to mirroring back the user’s current emotional state rather than proactively offering comfort or guidance to improve the user’s mood. Thereby, while displaying empathy is a critical step, it falls short of providing the deeper emotional support that is vital for enhancing a user’s psychological well-being, leaving a gap in effectively assisting users to overcome adverse emotional states.

This gap in effectively assisting users in overcoming adverse emotional states has motivated researchers to investigate the more specialized field of emotional support conversation, which we examine in the next subsection.

### 2.3.3 Emotional Support Conversation

Research in the field of Emotional Support Conversation can be primarily divided into two distinct strands: firstly, the classification of dialogue acts within utterances, as explored by Welivita and Pu [73], Malhotra et al. [48], and Svikhnushina et al. [66]; secondly, the creation of responses that are informed by the conversational context and predicted strategies as identified by Pérez-Rosas et al. [55] and Liu et al. [42], and, in certain cases, augmented by world knowledge and/or user feedback heuristics as demonstrated by Tu et al. [70], Xu et al. [78], Peng et al. [53], Cheng et al. [8], Deng et al. [10] and Peng et al. [54].

“Helping Skills” theory from psychology [24, 42] indicates that ESC systems should encompass three fundamental components: exploration (identifying the problem faced by the help-seeker), comforting (comfort the help-seeker’s emotion through expressing empathy), and action (assisting the help-seeker in problem-solving). These components are mutually reinforcing; for instance, the process of understanding the individual’s predicament is critical for expressing suitable empathy, and providing comfort enhances their willingness to divulge more about their experiences and emotions.

The inception of the ESC task by Liu et al. [42] catalyzed a series of innovations aimed at generating more effective responses. These include incorporating rich conversational context [9, 10, 53, 70], predicting strategy distribution [8, 70], employing reinforcement learning [90], and leveraging GPT-3.5 for enhanced context understanding [9]. Furthermore, recent efforts have explored the integration of semantics, strategy, and emotion transitions [84], yet a nuanced comprehension of the help-seeker’s predicaments and alignment with response strategies remains elusive.

Tu et al. [70] suggested improving context encoding by incorporating commonsense knowledge and directing response generation with a predicted strategy distribution. Specifically, Tu et al. [70] trained their model to predict a strategy using a negative log-likelihood loss against the single annotated strategy, while employing the predicted probability distribution for inference. While this approach offers advantages over a singular strategy application, it is suboptimal since the strategy predictor does not train on a true probabilistic distribution of strategies.

Peng et al. [53] introduced a hierarchical graph network designed to capture both the overarching context and specific user intentions. Cheng et al. [8] aimed to predict the support strategy in ESC by assessing the potential improvement in user emotion through an A\*-like algorithm, though their approach to user modeling was confined to understanding emotions. Zhou et al. [90] sought to optimize the ESC process via reinforcement learning, gauging the success of their approach by the degree to which it elicited positive user emotions. Despite their emphasis on eliciting positive emotions, their methodology could have been enhanced by a more comprehensive understanding of the problem, as our work proposes.

Furthermore, Deng et al. [10] applied commonsense and mental health knowledge to improve the system’s performance. Zhao et al. [84] incorporated semantics, strategy, and emotion transitions into their model. Cheng et al. [9] use GPT3.5 APIs at inference time to understand the dialog better. Despite these advancements, existing models, though effective, often lack a nuanced understanding of the problems faced by the help-seeker and alignment with the strategies to be used in the response, which our work aims to address. Prior works have predominantly

concentrated on the aspects of Comforting and Action as identified by Hill [24], whereas our study also advocates for Exploration, with a special focus on this dimension.

Despite these efforts, a gap persists in accurately capturing transitions in emotion at the level of individual utterances. A significant oversight is the reliance on common-sense knowledge or emotional cues as proxies for the user’s state, without adequately addressing the necessity for a progressive and concrete modeling of the underlying problem faced by help-seeker as the dialogue progresses. This gap in the methodology prevents these studies from effectively aligning the identified problems with the appropriate support strategies, due to the absence of a concrete modeling for the problem encountered by help-seekers. Furthermore, the existing research has been constrained by the use of a singular strategy annotation per utterance (e.g. in the ESCConv dataset [42]), limiting the exploration of multiple strategies within a single response.

We discuss the application of commonsense and world knowledge bases in the enhancement of ESC systems in the next subsection, further contextualizing their role in advancing the field.

### 2.3.4 Commonsense Knowledge for NLP

Recent literature has extensively explored the integration of commonsense knowledge into a wide array of NLP tasks. These tasks encompass classification [7, 52], question answering [4, 35, 50], story and language generation [18, 27], and dialogue systems [10, 70, 80, 84, 86, 89, 90]. Tu et al. [70], Xu et al. [78], Deng et al. [10], and Peng et al. [54] attempted to enhance comprehension of users’ mental and emotional states for ESC through the assimilation of data from external knowledge bases, such as those related to commonsense reasoning, conversational dynamics [5, 60, 63], and also mental health [74].

Among these, dialogue systems frequently employ ConceptNet [63] to enrich conversational utterances with grounded, physical knowledge. ConceptNet, a large-scale semantic network designed to help computers understand the meanings of words they encounter, provides a structured framework of commonsense knowledge. Originating from the Open Mind Common Sense project at the Massachusetts Institute of Technology, it encapsulates a wide array of relations and concepts formulated from everyday knowledge collected from ordinary people [41]. This expansive network includes various types of edges that represent different types of knowledge, from spatial relations to emotional associations, making it highly versatile for applications requiring deep semantic understanding.

In our approach, we harness ConceptNet to identify concepts relevant to the conversational context. We prioritize concepts with the highest emotional intensity values, as determined by the methodology proposed by Li et al. [34].



## Chapter 3

# Multi-level Training for Explainable Detection of Online Sexism

### 3.1 Overview

This chapter describes our approach for SemEval-2023 Task 10: Explainable Detection of Online Sexism (EDOS). The task deals with identification and categorization of sexist content into fine-grained categories for explainability in sexism classification. The explainable categorization is proposed through a set of three hierarchical tasks that constitute a taxonomy of sexist content, each task being more granular than the former for categorization of the content.

Considering the inter-connected task structure, and domain (Online Sexism), we study multilevel training to study the transfer learning from coarser to finer tasks. Our experiments based on pretrained transformer architectures also make use of additional strategies such as domain-adaptive pretraining to adapt our models to the nature of the content dealt with, and use of the focal loss objective for handling class imbalances. Results show that our methodology is effective in the explainable detection of online sexism.

### 3.2 Introduction

The proliferation of online sexism targeted towards women not only makes online spaces inhospitable but also perpetuates existing social inequities. This has led to the use of online systems for the identification of such content through automated technologies based on NLP methods. However, a blatant categorization into general, high-level categories as is often performed by such technologies leads to lack of explainability and trust in these systems. The nature of sexist content being varied based on the motivation of the writer, a discrete identification of categories as different as threats and prejudiced expressions is beneficial for differential treatment. A granular study also helps in identifying the weaknesses of the model in identifying certain specific categories of sexist content.

In our work, we work on the dataset for Explainable Detection of Online Sexism (introduced as a part of the SemEval-2023 Task 10) [30] based on English language posts from the social media platforms Reddit<sup>1</sup> and Gab<sup>2</sup>. The dataset, discussed in a greater detail in Section 3.3, introduces three inter-connected subtasks (labeled A, B, C) for the problem, each dealing with

---

<sup>1</sup><https://www.reddit.com/>

<sup>2</sup><https://gab.com/>

classification of sexist comments in a more fine-grained sense. Task A consists of a simple binary classification of posts into sexist and non-sexist categories. For posts which are detected as sexist, Task B deals with their categorization into 4 distinct categories of sexism. Finally, in Task C looks at a more granular classification of the sexist posts into 11 classes, or “fine-grained vectors of sexism”. We worked on all the three subtasks, and report our experimental findings on the same.

To benefit from the inter-connected, hierarchical nature of the three subtasks forming a taxonomy of sexist content labeling, we primarily study the technique of multi-level training in this chapter. This technique, explained in Section 3.4.4, serves to transfer learn on a task at the lower level from the learnings on a higher level task. We experiment with five different pretrained transformer [71] architectures for our approaches. Besides cross-entropy based vanilla finetuning (Section 3.4.1), we employ domain-adaptive pretraining (refer to Section 3.4.3) [19] as a precursor to multi-level training for adaptation of the models to the nature of the data in this domain. Considering the class imbalances in Tasks B and C, we also make use of the focal loss [37] objective, elaborated in Section 3.4.2. We specify the experimental settings for these methods in Section 3.5 and provide the results in Section 3.6. We show the suitability of domain-adaptive pretraining on Task A, multi-level training on Task B and focal loss on Task C as the best performing approaches for the tasks. The best results (on the basis of macro-F1) were obtained on the pretrained RoBERTa-large [43] model across tasks.

### 3.3 Dataset

In addition to the primary labeled dataset for the task, the organizers also provided two additional unlabeled datasets, each containing 2M comments from Reddit and Gab. The primary dataset consists of 14k / 2k / 4k labeled instances each in the train / dev / test splits. Of the 14k instances in the train split,  $\sim 3.4$ k comments are labeled under the “sexist” category for task A, hence limiting the amount of usable data for tasks B and C to this subset, which are further categorized into 4 and 11 classes respectively. Table A.1 provided in the appendix gives complete category definitions and examples for the labelled dataset. Although the classes for task A do not have a significant imbalance (3:1 distribution for “non-sexist” versus “sexist”), the classes for tasks B and C have a significant imbalances, as shown in Table 3.1.

We make use of the labeled dataset as well as the unlabeled Reddit and Gab datasets for our experiments. No other additional data was made use of. As a part of data exploration, we measured the average lengths of the comments in each dataset on the basis of white space-separated tokens. Average length of comments in the given labeled dataset is 23.42 tokens, slightly above that for Gab (16.17 tokens) and Reddit (17.89 tokens) unlabeled datasets. We also noticed that the length distributions of the comments from all the three data sources to conform when plotted on a graph, as shown in Figures 3.1, 3.2, and 3.3. This observation was useful in merging the datasets for applying domain-adaptive pretraining, discussed in Section 3.4.3.

Table 3.1: Distribution of class labels across tasks and data splits. Adapted from [30]

	<b>Train</b>	<b>Dev</b>	<b>Test</b>
<b>Task A</b>			
Not sexist	10,602 (76%)	1,514 (76%)	3,030 (76%)
Sexist	3,398 (24%)	486 (24%)	970 (24%)
<i>Total</i>	<i>14,000 (100%)</i>	<i>2,000 (100%)</i>	<i>4,000 (100%)</i>
<b>Task B</b>			
1. threats, plans to harm and incitement	310 (9%)	44 (9%)	89 (9%)
2. derogation	1,590 (47%)	227 (47%)	454 (47%)
3. animosity	1,165 (34%)	167 (34%)	333 (34%)
4. prejudiced discussion	333 (10%)	48 (10%)	94 (10%)
<i>Total</i>	<i>3,398 (100%)</i>	<i>486 (100%)</i>	<i>970 (100%)</i>
<b>Task C</b>			
1.1 threats of harm	56 (2%)	8 (2%)	16 (2%)
1.2 incitement and encouragement of harm	254 (7%)	36 (7%)	73 (8%)
2.1 descriptive attacks	717 (21%)	102 (21%)	205 (21%)
2.2 aggressive and emotive attacks	673 (20%)	96 (20%)	192 (20%)
2.3 dehumanising attacks and overt sexual objectification	200 (6%)	29 (6%)	57 (6%)
3.1 casual use of gendered slurs, profanities, and insults	637 (19%)	91 (19%)	182 (19%)
3.2 immutable gender differences and gender stereotypes	417 (12%)	60 (12%)	119 (12%)
3.3 backhanded gendered compliments	64 (2%)	9 (2%)	18 (2%)
3.4 condescending explanations or unwelcome advice	47 (1%)	7 (1%)	14 (1%)
4.1 supporting mistreatment of individual women	75 (2%)	11 (2%)	21 (2%)
4.2 supporting systemic discrimination against women as a group	258 (8%)	37 (8%)	73 (8%)
<i>Total</i>	<i>3,398 (100%)</i>	<i>486 (100%)</i>	<i>970 (100%)</i>

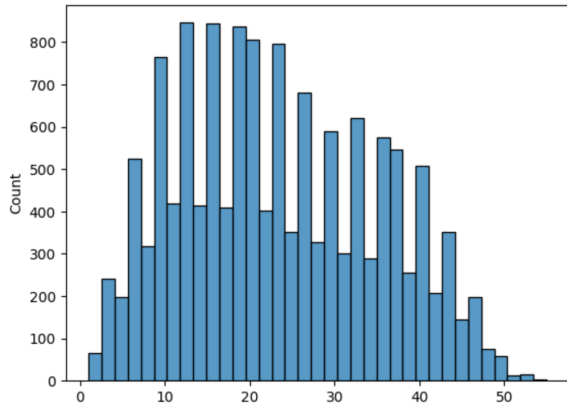


Figure 3.1: Length distributions of the comments in the labelled dataset

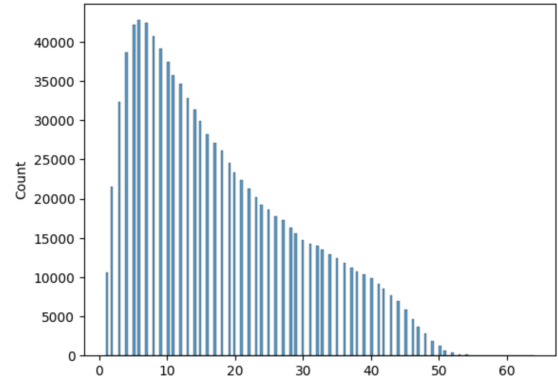


Figure 3.2: Length distributions of the comments in the unlabelled Reddit Dataset

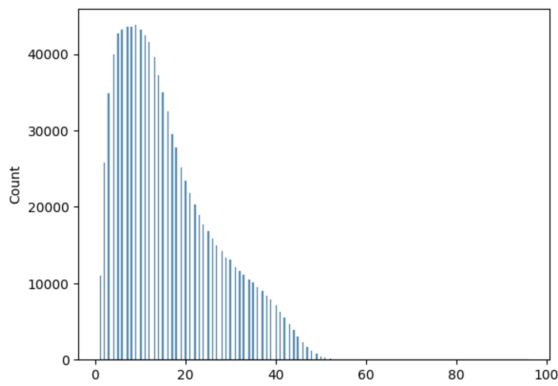


Figure 3.3: Length distributions of the comments in the unlabelled Gab Dataset

## 3.4 Methodology

In this section, we describe the core components of our modeling we experimented with to tackle the three subtasks. We start with basic supervised finetuning in Section 3.4.1, explain the focal loss objective in Section 3.4.2, domain-adaptive pretraining in Section 3.4.3, and multi-level training in Section 3.4.4.

### 3.4.1 Basic Supervised Finetuning

The first component of our approach is Basic Supervised Fine Tuning, a widely adopted technique for adapting pretrained transformer language models to specific tasks [12, 43]. This approach leverages the powerful generalization capabilities of transformer models, which are pretrained on vast amounts of unlabeled data, enabling them to capture rich representations of natural language. By finetuning these pretrained models on task-specific labeled data, we can effectively transfer the learned knowledge to the target task, allowing the model to adapt to its unique characteristics, such as vocabulary, syntax, and semantics.

The finetuning process involves adding a task-specific output layer on top of the pretrained model and training the entire model end-to-end on the labeled task data. This allows the model

to learn task-specific patterns while preserving the general language understanding acquired during pretraining. The effectiveness of this approach has been demonstrated across a wide range of natural language processing tasks, including text classification [64], named entity recognition [12], and question answering [79].

In the context of explainable detection of online sexism, basic supervised finetuning serves as a strong baseline for adapting pretrained transformer models to the specific challenges of the task. By leveraging the labeled sexism dataset during finetuning, the model can acquire the ability to differentiate between sexist and non-sexist content, as well as identify the various categories and fine-grained vectors of sexism defined in the task. This enables the model to capture sexist language’s high-level patterns and characteristics.

However, while basic supervised finetuning is a powerful technique, it may not always be sufficient to fully capture the nuances and complexities of the target task. The explainable detection of online sexism presents unique challenges that can limit the effectiveness of basic finetuning. One such challenge is class imbalance, where certain categories of sexism may be underrepresented in the dataset, leading to biased predictions. Additionally, the domain-specific language and vocabulary used in online sexist content may differ significantly from the general language encountered during pretraining, requiring the model to adapt to these linguistic variations. Furthermore, the hierarchical nature of the sexism taxonomy, with multiple levels of granularity, introduces dependencies and relationships between the different categories, which may not be effectively captured by basic finetuning alone.

To address these challenges and enhance the model’s performance, we explore complementary techniques that can be incorporated into the training process. These include focal loss [37], domain-adaptive pretraining [20], and multi-level training. These methodologies will be discussed in the following subsections.

### 3.4.2 Focal Loss

In Tasks B and C, we observed a significant class imbalance in the fine-grained classification tasks. Specifically, some classes had significantly fewer examples than others, which can lead to bias in the model’s predictions towards the majority class. To address this issue, we employed the Focal Loss [37] as a loss function for some experiments.

The Focal Loss is a modified version of the cross-entropy loss that down-weights the contribution of easy examples and focuses the model’s attention on hard, misclassified examples. It was originally introduced for object detection tasks in computer vision, but has since been successfully applied to various natural language processing tasks, including sentiment analysis and named entity recognition.

Mathematically, the Focal Loss is defined as:

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (3.1)$$

where  $p_t$  is the predicted probability for the true class,  $\alpha_t$  is a class-specific weighting factor, and  $\gamma$  is a focusing parameter. When  $\gamma = 0$ , the Focal Loss is equivalent to the standard cross-entropy loss. As  $\gamma$  increases, the effect of down-weighting easy examples becomes more pronounced, forcing the model to focus on hard, misclassified examples.

The class-specific weighting factor  $\alpha_t$  in the Focal Loss formula further helps in addressing class imbalance. In practice,  $\alpha$ -balanced variant yields slightly improved accuracy over non- $\alpha$ -balanced form [37].

The motivation behind using the Focal Loss in our experiments stems from the inherent class imbalance present in the fine-grained sexism classification task as visible in Table 1. In such scenarios, the model tends to be biased towards the majority class, as it can achieve high overall accuracy by simply predicting the majority class for most examples. This leads to poor performance on the minority classes.

By employing the Focal Loss, we aim to mitigate this issue by dynamically adjusting the weights of examples during training. The Focal Loss down-weights the contribution of easy examples, i.e., those that the model can classify correctly with high confidence. This prevents the model from being overwhelmed by the majority class and allows it to focus more on the hard, misclassified examples, which are likely to belong to the minority classes.

### 3.4.3 Domain-Adaptive Pretraining

The task of detecting and classifying sexism in online comments requires a nuanced understanding of language and context. To better adapt our models to this domain, we employed Domain-Adaptive Pretraining [19], a technique that involves pretraining our models on data from a similar domain or task before fine-tuning on the target task. By doing so, we aimed to improve the performance of our models, especially on the fine-grained classification tasks (Task B and C), which require a more nuanced understanding of sexist language.

In particular, we utilized the 2 million unlabelled data points collected from Reddit and Gab, which were provided by the task organizers. We used this data to pretrain several transformer language models. Our motivation for using this approach is based on the observation that the language patterns and structures used in social media or online forums are different from those used in more formal contexts such as news articles or books, which are part of the corpus used to create the pretrained LMs. Further Pre Training on data from social media or online forums can help our models learn to identify the specific linguistic cues that are indicative of sexism in this domain, such as informal language, sarcasm, and emoticons.

By leveraging Domain Adaptive Pretraining, we aimed to improve the generalizability of our models to the task and enhance their ability to detect and classify sexism in online comments.

### 3.4.4 Multi-level Training

Each subtask requires a progressively more fine-grained hierarchical classification of sexist content. We found that the previous tasks could serve as a good starting point for the current task, as they provide the model with valuable knowledge and representations that can be leveraged to improve its performance on the current task.

To this end, we employed Multi-Level Training, a technique that uses the best checkpoint of a previous task as the starting point for the current task. By using this technique, we aim to capitalize on the knowledge and representations learned from previous tasks and transfer them to the current task. For example, we use the best checkpoint from Task A to initialize the training for Task B, and the best checkpoint from Task B to initialize the training for Task C.

We believe that Multi-Level Training is particularly useful for Tasks B and C since each subtask builds upon the previous one, with the final subtask requiring the most fine-grained classification. By initializing the current task with the best checkpoint from the previous task, our model can leverage the knowledge learned from previous tasks and improve its performance on the current task, especially on the more challenging subtasks.

## 3.5 Experimental Details

### 3.5.1 Modeling

All our experimental strategies are tested on a set of 5 pretrained transformer encoders. These include a base, uncased version of the BERT [12] model (BBU), base and large versions of the RoBERTa [43] model (RB, RL), two versions of the DeBERTa [22] base model: the original base model (DB) and the v3 version (DBV3) [21], which was an improved version of the same. Modeling the classification task for each experiment involved addition of a linear layer with the output size equal to the number of classes for that task on the pooled embeddings of the last hidden layer of each model. For the pretraining experiment, a linear layer was added for the Masked Language Modeling (MLM) objective as in [12]. 15% of the tokens were masked randomly for MLM pretraining.

Transferring the model from one task to the other involved replacement of the final linear layer with a new layer suitable for that task. All the layers of the model were then trained, with the model encoder parameters adapting to the new task from their previous checkpoint, and the linear layer being trained on the task-specific objective from scratch.

### 3.5.2 Training Hyperparameters

For training, we employed a fine-tuning approach using the Slanted Triangular Learning Rates (STLR) [25] technique with a 10% warmup. We varied the values of the maximum learning rate in this scheduling to  $2e - 6$ ,  $5e - 6$  and found the value of  $5e - 6$  to work well across experiments. AdamW optimizer [45] was used with a weight decay of 0.01. In the experiments using focal loss, the gamma value of the hyperparameter was set to 2.0. The training was carried out over a course of 10 epochs using a batch size of 4 and accumulated gradient steps of 2. We relied on the Huggingface [75] library for pretrained encoder models and training support.

The aforementioned settings were used across all classification experiments and found them to work well. In domain-adaptive pre-training, we trained the model on 2 epochs over the complete data of  $\sim 2$ M samples. Rest of the hyperparameters were kept the same as in the classification experiments.

Best checkpoint was determined on the basis of the performance on the validation (dev) dataset using the macro-F1 score, the primary evaluation metric for the task. Besides macro-F1, we also measure the precision and recall values calculated in macro mode along with accuracy.

### 3.6 Results

The experimental results for our approaches for Task A, B and C can be seen in Table 3.3, Table 3.4 and Table 3.5 respectively. The values of F1, Precision and Recall metrics are noted in terms of their macro versions, as noted in Section 3.5.2.

In Task A, we observed that all the models, except DB, performed better after the domain-adaptive pretraining step. Among all the experiments, RL with pretraining gave the best F1 score of 85.93.

In Task B, we explored the usage of focal loss in basic finetuning, which showed improved performance in the best performing model, RL. However, we observed that the improvement from focal loss to not be consistent when compared to cross entropy. Pretraining followed by fine-tuning with focal loss showed improvement only for BBU. Multi-level training with pretraining gave the best model, which again was RL with a macro-F1 score of 69.96. Interestingly, BBU showed the highest performance improvement with continued pretraining and multi-level training where the score increased by 5.42 points from the second-best performing setup.

In Task C, we observed that the performance improvement from focal loss was more pronounced due to the high class imbalance of 11 categories. In basic supervised finetuning, changing the objective from cross entropy to focal loss improved the performance of each model by at least 3 points in the macro-F1 score. The best performing model in basic finetuning was RL, which showed an improvement of more than 4.27 points in macro-F1. The model that saw the highest improvement was DBV3 with a difference of 6.25 points in macro-F1. We also noted that continued pretraining gave improved performance for all models except RL. Multi-level training with Task A -> Task B -> finetuning with focal loss gave the best performance for BBU and RB. However, overall the best performing model remained RL in the basic finetuning with focal loss setup, giving a macro-F1 score of 54.62.

Strategy	Model	F1	Precision	Recall	Accuracy
Basic finetuning	BBU	81.62	81.97	82.47	86.80
	RB	83.22	83.94	82.58	87.90
	DB	83.69	83.69	83.69	88.00
	DBV3	82.51	82.69	82.32	87.20
	RL	85.73	<b>87.11</b>	84.56	89.85
Pretraining → finetuning	BBU	81.93	83.02	81.00	87.10
	RB	83.86	84.56	83.22	88.35
	DB	82.40	85.02	80.52	87.85
	DBV3	83.93	84.15	83.72	88.25
	RL	<b>85.93</b>	86.90	<b>85.09</b>	<b>89.90</b>

Table 3.3: Performance of our models on the validation set of Task A.



Strategy	Model	F1	Precision	Recall	Accuracy
Basic finetuning: CE	BBU	58.91	62.39	56.68	62.14
	RB	63.10	64.84	63.85	65.23
	DB	68.33	67.07	69.90	69.34
	DBV3	61.43	61.71	62.13	63.79
	RL	69.41	69.19	69.71	70.99
Basic finetuning: FOCAL	BBU	58.60	60.43	57.21	62.14
	RB	64.02	65.36	62.90	64.81
	DB	67.79	67.53	68.21	69.14
	DBV3	62.72	62.03	63.86	62.55
	RL	69.67	69.11	71.47	<b>71.60</b>
Pretraining → finetuning: FOCAL	BBU	60.91	59.97	62.00	61.32
	RB	57.30	57.17	57.49	57.61
	DB	62.13	60.33	64.80	62.55
	DBV3	56.98	56.16	58.69	56.17
	RL	69.57	67.24	<b>73.39</b>	69.55
Pretraining → Task A → finetuning: FOCAL	BBU	66.33	65.94	67.28	67.28
	RB	62.87	64.92	61.36	62.96
	DB	67.05	65.60	68.87	68.93
	DBV3	62.53	64.19	61.21	64.20
	RL	<b>69.96</b>	<b>69.24</b>	70.98	70.99

Table 3.4: Performance of our models on the validation set of Task B.

### 3.7 Summary

In this chapter, we present a comprehensive study of explainable detection of online sexism, focusing on fine-grained hierarchical classification tasks. Our approach involve domain-adaptive pretraining, focal loss, and multi-level training methodologies to effectively tackle the challenges posed by the dataset sourced from Reddit and Gab.

The experimental results demonstrate the efficacy of our proposed techniques. Domain-adaptive pretraining proves to be a powerful tool for aligning pretrained models with the downstream task, leading to improved performance across all tasks. The focal loss objective successfully addresses the class imbalances prevalent in the dataset, particularly in Task C, which exhibits a high degree of imbalance. Furthermore, the multi-level training strategy emerges as the best-performing approach for Task B, showcasing the benefits of learning from coarser to finer classification tasks.

Across all experiments involving five different pretrained transformer models, the RoBERTa-large model consistently outperforms the others, highlighting the importance of model capacity and adaptability to small datasets. These findings underscore the significance of architecture selection and pretraining strategies in the context of explainable sexism detection.

In conclusion, our work demonstrates the effectiveness of domain-adaptive pretraining, focal loss, and multi-level training in tackling the challenge of fine-grained hierarchical sexism classification. The promising results obtained on the Reddit and Gab dataset highlight the potential

Strategy	Model	F1	Precision	Recall	Accuracy
Basic finetuning: CE	BBU	33.07	33.51	34.43	54.12
	RB	38.48	37.91	39.57	55.76
	DB	44.90	54.68	43.10	56.58
	DBV3	34.97	35.68	36.09	55.14
	RL	50.35	52.34	49.31	61.32
Basic finetuning: FOCAL	BBU	36.22	34.97	39.27	47.33
	RB	43.05	42.48	44.34	50.21
	DB	49.79	51.25	49.17	55.35
	DBV3	41.23	40.28	43.95	49.59
	RL	<b>54.62</b>	<b>56.33</b>	<b>53.79</b>	<b>61.32</b>
Pretraining → finetuning: FOCAL	BBU	37.32	41.34	38.19	47.33
	RB	39.83	38.87	42.07	47.33
	DB	51.16	51.90	52.19	58.23
	DBV3	41.38	40.21	43.67	50.62
	RL	51.67	55.95	50.97	60.08
Pretraining → Task A → Task B → finetuning: FOCAL	BBU	40.68	48.61	41.04	51.44
	RB	43.15	44.22	43.00	53.50
	DB	49.97	51.04	49.77	57.41
	DBV3	39.68	42.80	39.64	51.23
	RL	51.15	51.98	50.80	59.67

Table 3.5: Performance of our models on the validation set of Task C.

of these approaches in advancing the explainable detection of online sexism. By continuing to refine and expand upon these methodologies, we can work towards building more robust and socially responsible NLP systems that promote healthier online discourse.

## *Chapter 4*

# **Dialog Dataset and Classifier for Workplace-related Mental Health Support**

## **4.1 Overview**

Mental health disorders among working-age adults pose a significant challenge, with 15% of this population affected and 12 billion working days lost annually to depression and anxiety alone. Practical support for workplace-related mental health issues is crucial, and AI-driven solutions such as chatbots and enhanced responses by mental health professionals show promise. This chapter addresses the lack of prior research in this area within a mental health forum setting by introducing the Beyond Blue Mental Health (BBMH) dataset, derived from the moderated Beyond Blue forum. A subset of this dataset, focusing on Workplace-related Mental Health issues (BBWMH), is also created. We develop a workplace classifier for mental health posts and establishes strong baseline dialog models for workplace-related mental health support. The results demonstrate the potential of AI-driven solutions in addressing mental health challenges in the workplace, paving the way for future research and practical applications.

## **4.2 Introduction**

Mental health disorders among working-age adults are a pressing concern, with 15% of this population affected and a staggering 12 billion working days lost annually to depression and anxiety alone [77]. The impact of mental health conditions extends beyond the individual, affecting families, colleagues, and society at large. The World Health Organization emphasizes the importance of practical support for mental health issues in the workplace, as it can significantly influence an individual’s confidence, productivity, and ability to retain or gain employment [77].

AI-driven solutions, such as chatbots and enhanced responses by mental health professionals, have the potential to provide accessible and efficient support for workplace-related mental health issues. However, prior research in this area within a mental health forum setting is limited. The critical challenge lies in modeling long-range context, as mental health forum posts often contain dozens of sentences and exceed hundreds of tokens per post or response.

To address this gap, we introduce the Beyond Blue Mental Health (BBMH) dataset, derived from the moderated Beyond Blue<sup>1</sup> forum. Beyond Blue is an Australian organization that aims to raise community awareness about anxiety and depression and reduce the associated stigma. The moderated nature of the forum ensures the quality of the data collected. We further refine the BBMH dataset to create a subset focused on workplace-related mental health issues (BBWMH).

In addition to creating these datasets, we develop a workplace classifier for mental health posts using various transformer models and zero-shot prompting with GPT-3.5. We also establish strong baseline dialog models for workplace-related mental health support using 4-bit QLoRA finetuning on Llama-2-7b, Gemma-2b, Mistral-7b, and Mistral-7b-instruct.

The main contributions of this study are:

1. The creation of the BBMH dataset, a comprehensive resource for AI research in mental health support derived from real-world conversational data.
2. The development of the BBWMH dataset, a specialized subset focusing on workplace-related mental health issues.
3. Studied the effectiveness of workplace classifier for mental health posts using Pre-Trained Language models as well as the performance of zero-shot prompting with GPT-3.5.
4. The establishment of strong baseline dialog models for workplace-related mental health support using state-of-the-art large language models and efficient finetuning techniques.

### 4.3 Dataset

The dialog dataset was created from threads of discussions on mental health-related posts from the Beyond Blue forum. Beyond Blue is a moderated mental health forum that has partnerships with governments, local health services, educational institutions, workplaces, media, and community organizations, as well as the general community in Australia. The organization began in October 2000 as a five-year initiative of the Australian Government and state and territory governments, aiming to raise community awareness about anxiety and depression and reduce the associated stigma.

The presence of a dedicated organization behind the moderated Beyond Blue forum gives us confidence in the quality of the data collected. From our experience reading posts from the forum, we can confirm the high quality of the content.

We included all threads from the Beyond Blue forum in our dataset, which we call the BeyondBlue Mental Health (BBMH) dataset. In total, we have 36,423 threads from the 10 sub-forums available, as shown in Figure 4.1.

Beyond Blue forums follow a structure where one user can start a thread by posting about a mental health issue they are facing. Members in the forum can reply to that root post, starting a conversation. The original poster (OP) can then reply to the response by a community member. Essentially, we can model the interactions as a conversation tree, as illustrated in Figure 4.2.

---

<sup>1</sup><https://forums.beyondblue.org.au>

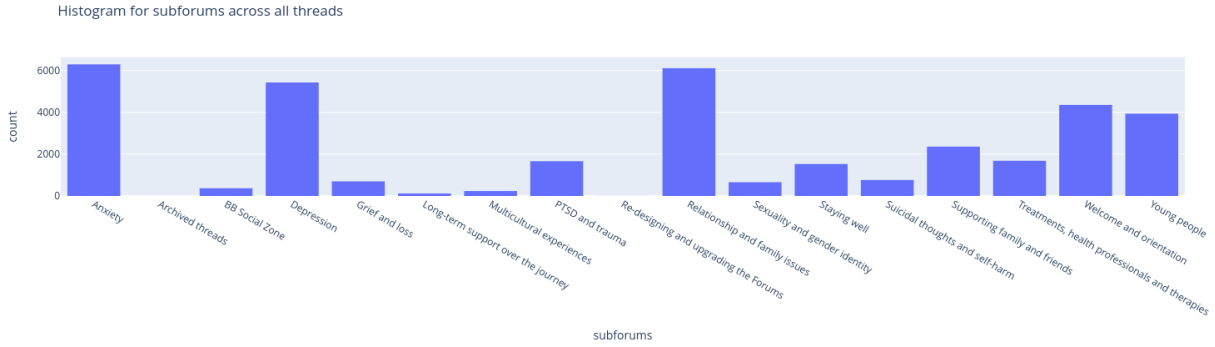


Figure 4.1: Number of threads from each sub-forum in the BBMH dataset. Labels correspond to the names of the sub-forums.

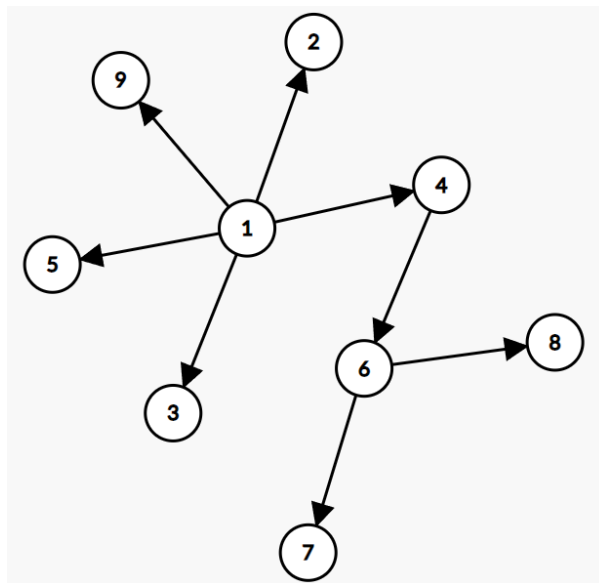


Figure 4.2: Conversation tree structure in Beyond Blue forums.

In the example shown in Figure 4.2, the conversation tree starts with post #1. There are two conversations of length greater than 2:

1.  $1 \rightarrow 4 \rightarrow 6 \rightarrow 8$
2.  $1 \rightarrow 4 \rightarrow 6 \rightarrow 7$

In short, any path from a root post to a leaf is a conversation on its own. From every thread, we extract all paths (subject to constraints) from root to leaves and add them to our conversation dataset. We use the terminology 'path' to refer to one conversation from root to leaf in a thread. In total, we have 1,61,707 paths from all the threads in the Beyond Blue dataset.

The members in the forum are given ranks based on their contributions and engagement. Table 4.1 shows the rank and the number of responses in the dataset by members with that rank.

Rank	Num Responses
Beyond Blue Staff	480
Blue Voices Member	73840
Champion Alumni	49062
Community Champion	57890
Community Member	229702
Moderator	6040
Valued Contributor	30588

Table 4.1: Response Statistics by Rank

As evident from Table 4.1, almost half of the responses are by members above the "Community Member" rank, indicating the high quality of the responses in the dataset.

To create the workplace-related mental health subset (BBWMH), we attempted to create a weak-labeller. For this, we did manual annotation for a subset of 1500 samples. We also experimented with zero-shot prompting using GPT-3.5. The details of the workplace classifier development are discussed in the next section. Finally, we filtered out a subset of 5,240 conversation threads specifically related to workplace mental health issues.

The BBMH and BBWMH datasets provide a rich resource for AI research in mental health support, offering real-world conversational data with extensive posts and responses. The moderated nature of the Beyond Blue forum ensures the quality and relevance of the data, making these datasets valuable for developing and evaluating AI-driven mental health support systems.

## 4.4 Workplace Classifier for Mental Health Posts

The task of classifying workplace-related mental health posts is crucial for creating a targeted dataset that addresses the specific challenges and concerns faced by individuals in their professional lives. By identifying posts that discuss mental health issues stemming from or affecting the workplace, we can develop more effective and relevant support systems.

### 4.4.1 Defining Workplace-Related Mental Health Posts

In this study, a workplace-related mental health post is defined as any forum post by a help-seeker that discusses mental health issues arising from, or significantly impacted by, the workplace environment. Specifically, we consider the following criteria to identify a post as workplace-related:

- Mental health challenges directly caused by workplace dynamics or conditions.
- Experiences of bullying, harassment, or undue pressure from colleagues, supervisors, or subordinates.
- Stress attributed to job roles, expectations, workload, or the work environment itself.
- Situations where the workplace negatively influences one’s family life, social relationships, or well-being outside of professional realms.

- Any aspect of the workplace that detrimentally affects an individual’s mental health or quality of life.

#### 4.4.2 Methodology for Developing the Workplace Classifier

The creation of the workplace-related mental health subset (BBWMH) from the Beyond Blue Mental Health (BBMH) dataset involved the development of weak-labellers and the exploration of zero-shot prompting with state-of-the-art Large Language Models (LLMs), specifically GPT-3.5, as of the time this research was conducted. Our methodology encompassed several structured steps aimed at curating and refining the dataset:

**Data Collection for the Training Set:** The initial phase involved gathering a diverse range of posts to serve as our training set. This step was meticulously executed through:

- Filtering and annotating posts that contained the keyword 'boss' from the Beyond Blue forum, recognizing that direct mentions of supervisory roles often correlate with workplace-related discussions.
- Annotating a selection of posts from a dedicated sub-forum titled “Employment and Mental Health Forum” on the Mental Health Forum (MHF), providing a focused source of relevant discussions.
- Randomly sampling and annotating negative examples from the broader Beyond Blue forum content, ensuring our model could discern between workplace-related and non-workplace-related posts with high accuracy.

**Data Collection for the Test Set:** To validate the effectiveness of our classifier, we compiled a test set comprised of randomly sampled posts from the Beyond Blue forum, offering a broad spectrum of discussions to evaluate the classifiers performance in accurately identifying workplace-related content. This ensured a representative sample of the test.

The number of examples in the annotated training and test sets is given in Table 4.2.

	Positive	Negative	Total
Train	464	583	1047
Test	87	504	591

Table 4.2: Distribution of Positive and Negative Classes in Train and Test Sets

Initially, we fine-tuned different transformer models on the annotated dataset. The results of these models are presented in Table 4.3.

Multiple checkpoints of the same model are given to highlight the precision-recall tradeoff. For the weak labeler, we prioritized models with high precision, even if the recall was not optimal. In Table 4.3, the Longformer model outperformed others, likely due to its ability to handle longer context (1024 tokens) compared to the other models limited to 512 tokens.

To further improve the classifier’s performance, we experimented with the Slanted Triangular Learning Rate (STLR) scheduler and weight decay. This approach yielded checkpoints with even better performance, as shown in Table 4.4.

Model	Precision	Recall
Longformer - checkpoint 1	0.8846	0.2644
Longformer - checkpoint 2	0.8393	0.5402
MentalRoBERTa - epoch 1	0.8235	0.4827
MentalRoBERTa - epoch 2	0.7195	0.6781
MentalBERT - epoch 3	0.7424	0.5632
Bert-base-uncased - epoch 1	0.7666	0.5287

Table 4.3: Model Performance Comparison

Checkpoint #	Precision	Recall
20	0.9355	0.333
21	0.9231	0.4138
33	0.8276	0.5517
65 (run 2)	0.9348	0.4943

Table 4.4: Precision and Recall at Various Checkpoints with STLR and Weight Decay

The loss function used for training the classifier was binary cross-entropy, which is well-suited for binary classification tasks.

To demonstrate the quality of the true positive classifications and provide examples of the data, we present two randomly chosen true positive samples below.

**Sample 1:** I have untreated (mostly undiagnosed) depression for years but it has never impacted my ability at work. Actually I am so skilled at masking it no one in a work place from my past would actually have the slightest idea. My current workplace has been bad for my mental health. I have felt myself sinking more and more over the past few months. There has been a lot of focus on me compared to co-workers and I get why there would have been at the start and I also realise that some of the focus these days is more to do with I have somewhat (sometimes) impressed people there with my ability so they are looking and talking about future potential. None the less, the focus on me has lead to changes in me that I dont like. Doubting my knowledge, doubting my ability which is sending me all over the place. I will add I had quite a bit going on in my life in the last 6 months of last year. recently I realised just how far I had fallen virtually at the hands of work and all the chatter about me by senior staff. I wasnt prepared to fall lower (I feel so low already i am not sure if lower is even possible, although logically I know it is) so I actually resigned and am currently in my notice period. It was a super hard decision as I actually really like the work and the potential to grow in the long term but my low no longer just affects how I feel at work it is impacting every element of



my life so it seemed like what I needed to do. now as I wait out my notice period I keep wondering how much of all of this is the black dog speaking and how much was it what I really needed. I felt low before but now I feel like there is a huge gaping hole in me. Is the hole and doubt about what I did being right actually correct or if it the fear of the unknown - no job, no prospects, no idea how to sell myself felling this way. how do you know when quitting is the right solution and when it is just the black dog dragging you through the mud? If it is actually just the black dog, should you even attempt to get your job back? If you should, how do you even tell them about it? Telling them may make things worse as it will get me even more attention and probably more chatter, I just really like my job and hate quitting but my mental health is making it seem impossible. Thanks

**Sample 2:** The past week Ive been feeling very anxious from work. Ive recently been given a written warning due to some avoidable errors I made in my admin tasks. Im getting a review at the end of this week and despite my best efforts to try and improve, it doesnt feel like enough for my employer to want to keep me. as someone who is supporting an unemployed mum alone, paying all the bills and sacing for a trip in a few months, this has put a huge weight in my stomach. Im constantly losing my appetite, finding difficulty sleeping and running short on breath. I dont what to do and it feels like theres no one who can make things better.

Additionally, here is an example of a true negative sample, which includes a workplace-related keyword ‘work’. This indicates that the model can interpret contextual nuances without defaulting to simple keyword recognition.

**Sample:** I’ve struggled with anxiety for 3 months now since having major surgery that seems to have triggered it. They have just put me on anti-depressants and been taking them for 7 days now. I feel even more nauseated, no appetite and heart pulpatations all the time. Could this be the medication? I need reassurance that it will get better. Seeing the doc and psychologist regularly. Im doing breathing techniques, yoga... trying to keep busy but Im so tired from lack of sleep. Had to leave early from work today because I couldnt stop vomitting. It will get better right?

#### 4.4.3 ChatGPT annotation for Workplace Posts

The release of GPT-3.5 marked a significant advancement in large language models (LLMs), presenting new opportunities for high-quality, extensive annotation tasks. Recognizing this potential, we explored the use of GPT-3.5 for annotating our dataset of workplace posts related to mental health.

To optimize the performance of GPT-3.5 for our specific annotation task, we engaged in an iterative process of fine-tuning the prompts and evaluating the model’s performance. This fine-tuning process involved carefully crafting and refining the prompts to align with the unique characteristics and requirements of our dataset.

We conducted multiple rounds of evaluation, assessing the model’s precision and recall scores on randomly selected samples of annotated data. Through this rigorous fine-tuning and evaluation process, we progressively improved the performance until we achieved a perfect precision score and a recall rate exceeding 95%. Having attained this exceptional level of performance, we utilized zero-shot prompting with GPT-3.5 for the final annotation of the BBMH dataset.

The prompt used for the classificatoin is given below.

```
A workplace-related mental health issue is a post where either
of the following happens:
```

1. A clear indication that the primary cause of their mental health problems is an issue they face at their job or workplace.
2. Getting bullied by colleagues at workplace.
3. Difficulty or poor performance while doing their work due to mental health problems.
4. Workplace being one of the reasons for negatively affecting a person’s life or their family.
5. Workplace causing the person to show signs of stress, depression, anxiety, etc.
6. Dissatisfaction with job.

```
Note that any difficulties faced at schools or universities
doesn’t mark a post as workplace-related mental health
issue.
```

```
Using the above definition above, classify the below post as a
workplace-related mental health issue. First answer in a
single word, yes or no and then provide reasoning from the
next line. Strictly follow the output format.
```

```
<Root post goes here>
```

#### 4.4.4 Workplace Classifier Conclusion

In conclusion, we developed a workplace classifier for mental health posts using a combination of manual annotation, transformer models, and zero-shot prompting with GPT-3.5. The final finetuned classifier achieved a precision of 0.94 and a recall of 0.33 on the Beyond Blue dataset. However, zero-shot prompting with GPT-3.5 proved to be the most effective method, with a

recall rate exceeding 95% and perfect precision. Using this classifier, we filtered the BBMH dataset and created the first-ever mental health dialog dataset for workplace-related issues, containing 5,240 conversation threads with a total of 49,238 utterances.

## 4.5 Dialog Model for Workplace-related Mental Health Support

Following the creation of the Beyond Blue Workplace Mental Health (BBWMH) dataset, we established baselines using state-of-the-art large language models (LLMs). The performance of pretrained language models for dialog, such as DialoGPT [81] and BlenderBot[58], is likely to be suboptimal in BBWMH due to their training on datasets with shorter dialogues and limited context. For instance, DialoGPT was trained on 147 million conversational exchanges derived from Reddit comments, with a default maximum sequence length of 128 tokens. Similarly, BlenderBot was trained with maximum context and response lengths set to 128 BPE tokens. In contrast, the BBWMH dataset contains extensive posts and responses, often exceeding the context limits of these models.

However, due to computational constraints in our lab, which is equipped with RTX 2080Ti having 11GB VRAM, we could not fine-tune larger models like Llama-2 or Gemma with architectural changes. Therefore, for establishing baselines, we used the more memory efficient 4-bit QLoRA finetuning [11] on Llama-2 [68], Mistral [28], and Gemma [67].

### 4.5.1 Evaluation Metrics

To evaluate the performance of the fine-tuned models on the BBWMH dataset, we employed several widely-used metrics in dialog systems research. These metrics include:

- BLEU [51]: Measures the precision of n-gram overlap between the generated response and the ground truth response.
- ROUGE-L [36]: Computes the longest common subsequence between the generated response and the ground truth response.
- METEOR [3]: Considers both precision and recall of unigram matches between the generated response and the ground truth response, with additional emphasis on word order.
- Distinct-n [40]: Calculates the diversity of n-grams in the generated responses, providing insights into the model’s ability to generate varied and informative responses.

BLEU, ROUGE-L, and METEOR primarily measure the lexical overlap between n-grams of the predicted and ground truth responses, while Distinct-n focuses on the diversity of the generated responses.

### 4.5.2 Results

The results of the fine-tuned models on the BBWMH dataset are presented in Table 4.5. The results demonstrate that the fine-tuned models achieve competitive performance across

various metrics. Instruct tuned Mistral-7B exhibits the highest scores in BLEU, METEOR, and ROUGE, indicating its ability to generate responses that closely match the ground truth. Llama-2-7B, on the other hand, achieves the highest Distinct-2 and Distinct-3 scores, suggesting a higher level of diversity in its generated responses.

Metric	Gemma-2B	Llama-2-7B	Mistral-7B	Mistral-Instruct-7B
Distinct-1	2.29	2.89	<b>2.90</b>	2.69
Distinct-2	14.40	<b>20.64</b>	19.95	18.05
Distinct-3	33.73	<b>49.99</b>	45.41	41.60
BLEU-1	29.22	29.66	30.00	<b>30.55</b>
BLEU-2	14.82	12.61	15.03	<b>15.49</b>
BLEU-3	7.88	5.15	7.86	<b>8.28</b>
BLEU-4	4.59	2.14	4.52	<b>4.90</b>
METEOR	13.39	12.75	13.74	<b>13.96</b>
ROUGE-L	18.72	15.48	18.42	<b>18.77</b>

Table 4.5: Performance of fine-tuned models on the BBWMH dataset

These fine-tuned models serve as strong baselines for future work in the domain of workplace-related mental health support using dialog systems. The BBWMH dataset, along with the established baselines, provides a foundation for researchers to develop and evaluate more advanced models and techniques in this critical area.

## 4.6 Summary

In this work, we have made significant contributions to the field of AI-driven mental health support, focusing specifically on workplace-related mental health issues. Our main contributions include:

1. The creation of the Beyond Blue Mental Health (BBMH) dataset, a comprehensive resource containing 36,240 forum posts for AI research in mental health support.
2. The development of the Beyond Blue Workplace Mental Health (BBWMH) dataset, a specialized subset of 5,240 conversation threads and 49,238 utterances addressing workplace-related mental health concerns.
3. The implementation of a workplace classifier for mental health posts using advanced techniques such as zero-shot prompting with GPT-3.5, achieving perfect precision and a recall rate exceeding 95%.
4. The establishment of strong baseline dialog models for workplace-related mental health support by fine-tuning state-of-the-art language models (Llama-2-7b, Gemma-2b, and Mistral-7b) on the BBWMH dataset using 4-bit QLoRA.

The datasets and models introduced in this study contribute significantly to the growing body of research in AI-driven mental health support, particularly in the context of workplace-related issues. The BBMH and BBWMH datasets provide valuable resources for researchers and practitioners to develop and evaluate innovative solutions for addressing the mental health challenges faced by individuals in their professional lives.

Furthermore, the workplace classifier and the finetuned dialog models serve as strong baselines for future work in this domain. By leveraging advanced techniques such as zero-shot prompting with GPT-3.5 and 4-bit QLoRA finetuning, we have demonstrated the potential of AI-driven solutions in providing effective and efficient mental health support.

In conclusion, this study represents a significant step forward in addressing workplace-related mental health issues through the development of comprehensive datasets and advanced AI-driven solutions. By providing valuable resources and establishing strong baselines, we aim to inspire and facilitate further research and practical applications in this critical area, ultimately contributing to the well-being of individuals in their professional lives.

## Chapter 5

# PRISM: Problem Identification and Strategy Matching for Emotional Support Conversations

### 5.1 Overview

Automated Emotional Support Conversation (ESC) systems are critical for providing online counseling and therapy to people with mental health issues, students trying to cope with academic stress and anxiety, and broadly people who are lonely or isolated. In an attempt to mimic a human supporter, previous methods condition on the conversational context to first predict a support strategy and then generate an Emotional Support(ES) response. Unfortunately previous studies fail to imitate the intricate art of (a) dynamically understanding the evolving problem category distribution of the help-seeker, and (b) responding using a combination of strategies known to be effective to address the predicted problem distribution. To address this challenge, we propose Problem Identification and Strategy Matching for ESC (PRISM), a novel framework that leverages Transformer-based encoder-decoder models to predict fine-grained problem and strategy distributions and generate effective Emotional Support responses. Unlike existing models that utilize coarse-grained strategy annotations and conversation-level emotion labels, training PRISM models requires fine-grained, probabilistic annotations for both support strategies and underlying problems. For this, we introduce ESConv++, a dataset annotated with problem and strategy distributions using GPT-4, to train PRISM models. The framework is optimized using four ESC-specific loss functions: problem prediction loss, strategy prediction loss, problem-strategy mapping loss, and cross-entropy loss. Experimental results demonstrate that PRISM models significantly outperform strong baselines on BLEU and METEOR metrics. Our work advances the field of ESC by enabling more dynamic and personalized emotional support generation.

### 5.2 Introduction

Mental health disorders present a substantial global burden, with over 50% of adults experiencing mental health issues at some point in their lives [49]. Despite the widespread prevalence of these conditions, a staggering 70% of affected individuals worldwide do not receive treatment from professional healthcare providers [23]. Recent research [6, 16] highlights the benefits

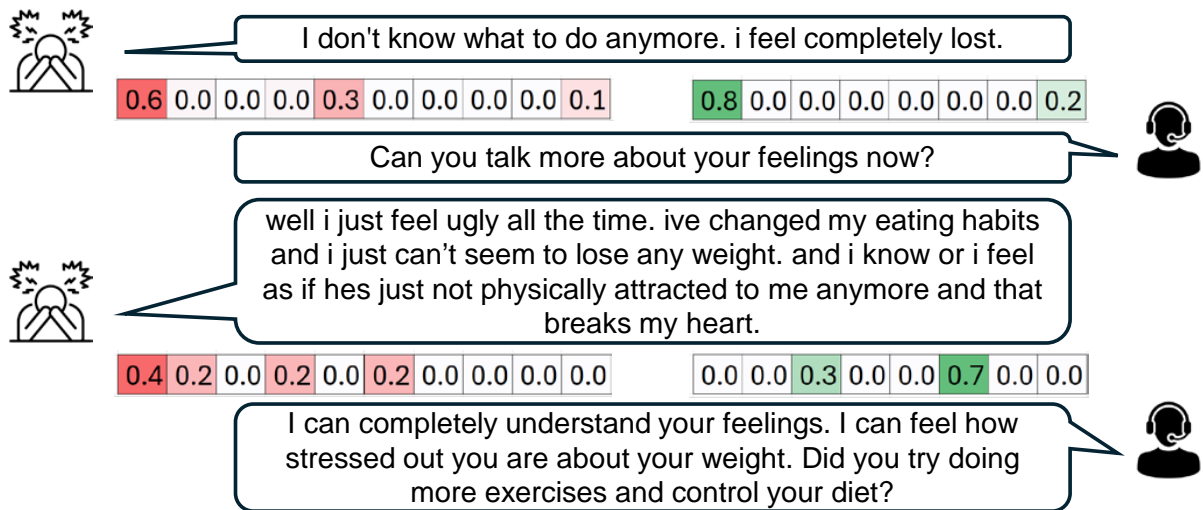


Figure 5.1: An example conversation from ESConv++ with help-seeker (left) and supporter (right). Problem and support strategy distributions are shown in red and green, respectively, with categories ordered as in 5.5.1. Both help-seeker utterances have emotion 'sadness'.

of Emotional Support Conversations (ESC) as a method to alleviate emotional distress and therefore improve mental health outcomes. ESC systems are both desirable and essential for endpoints like (a) a chatbot that provides online counseling and therapy for people with mental health issues, (b) a virtual assistant that helps students cope with academic stress and anxiety, (c) a companion app that offers emotional support and guidance for people who are lonely or isolated.

Concretely, we define the ESC task as follows. Given a conversation between a help-seeker and a supporter, and the most recent utterance from the help-seeker, the goal is to predict a response that reflects understanding the expressed help-seeker problems and effective use of apt support strategies.

Prior research in the domain of emotional dialogue systems primarily utilized conversation-wide emotion labels to infer user feelings, which led to the generalization of emotional contexts as uniform and unchanging throughout the dialogue [33, 39, 47, 57]. This method does not account for the dynamic nature of user emotions, which may fluctuate in intensity even if the overall emotional state appears stable.

Early endeavors in empathetic conversation systems aimed at incorporating empathy to recognize and comprehend the emotional states of users. Nonetheless, the provision of empathetic reactions within individual parts of the conversation, as seen in these systems [57], is not adequate for Emotionally Supportive Conversational agents (ESCs). The practice of merely offering empathetic feedback during separate exchanges might neglect the continuous changes and feedback of the user across extended dialogues. Thus, it is crucial to develop a nuanced appreciation of the user's emotional condition for precise user representation. Similar to a therapist's increasing understanding of a client's issues through the course of a counseling session, which enhances as more insights are shared by the individual seeking help, conversational agents must also refine their perception of the user's emotional state for improved support and interaction.

Research in the field of Emotional Support Conversation can be primarily divided into two distinct strands: firstly, the classification of dialogue acts within utterances, as explored by Welivita and Pu [73], Malhotra et al. [48], and Svikhnushina et al. [66]; secondly, the creation of responses that are informed by the conversational context and predicted strategies as identified by Pérez-Rosas et al. [55] and Liu et al. [42], and, in certain cases, augmented by world knowledge and/or user feedback heuristics as demonstrated by Tu et al. [70], Xu et al. [78], Peng et al. [53], Cheng et al. [8], Deng et al. [10] and Peng et al. [54].

“Helping Skills” theory from psychology [24, 42] indicates that ESC systems should encompass three fundamental components: exploration (identifying the problem faced by the help-seeker), comforting (comfort the help-seeker’s emotion through expressing empathy), and action (assisting the help-seeker in problem-solving). These components are mutually reinforcing; for instance, the process of understanding the individual’s predicament is critical for expressing suitable empathy, and providing comfort enhances their willingness to divulge more about their experiences and emotions.

A significant oversight in prior efforts is the reliance on common-sense knowledge or emotional cues as proxies for the user’s state, without adequately addressing the necessity for a progressive and concrete modeling of the underlying problem faced by help-seeker as the dialogue progresses. This gap in the methodology prevents these studies from effectively aligning the identified problems with the appropriate support strategies, due to the absence of a concrete modeling for the problem encountered by help-seekers. Furthermore, the existing research has been constrained by the use of a singular strategy annotation per utterance (e.g. in the ESCConv dataset [42]), limiting the exploration of multiple strategies within a single response. Although Tu et al. [70] tried implementing a combination of strategies to overcome this, the absence of ground truth labels for a probabilistic strategy distribution per utterance poses a challenge.

In this chapter, we propose a sequence-to-sequence encoder-decoder architecture, PRISM, to facilitate multi-turn emotional support conversation through fine-grained problem comprehension and the alignment of problem with intervention strategies. Our approach augments dialogue context by incorporating the emotion tag into each utterance explicitly. Furthermore, we integrate pertinent external commonsense knowledge concepts across the dialogue, with a special emphasis on the most recent help-seeker utterance. The encoder output of the enhanced dialogue context is utilized to train two distinct modules: one aimed at identifying the help-seeker’s challenges based on the ongoing conversation, and another focused on determining the appropriate intervention strategies and their relevance in formulating an effective response. We introduce a loss function designed to guarantee alignment between the identified problem and the predicted strategy.

The distributions of both predicted problems and strategies are utilized to craft intricate representations for each. Following this, the decoder’s input is enhanced with this representation method for the problems and strategies, which are then integrated as the initial two token embeddings in the decoder.

Overall, we make the following main contributions.

1. We present a novel ESC framework, PRISM, which incorporates dynamic problem and strategy distributions to generate effective ES responses.



	$s_1$	$s_2$	$s_3$	$s_4$	$s_5$	$s_6$	$s_7$	$s_8$
$t_1$	0.214	0.007	0.011	0.010	0.012	0.005	0.712	0.028
$t_2$	0.553	0.061	0.080	0.049	0.050	0.041	0.093	0.073
$t_3$	0.396	0.085	0.123	0.099	0.081	0.113	0.012	0.092
$t_4$	0.272	0.066	0.117	0.128	0.107	0.215	0.002	0.092
$t_5$	0.225	0.046	0.103	0.130	0.119	0.294	0.003	0.080
$t_6$	0.190	0.033	0.074	0.127	0.138	0.338	0.008	0.094
$t_7$	0.140	0.023	0.069	0.135	0.156	0.365	0.014	0.098
$t_8$	0.125	0.022	0.067	0.121	0.183	0.362	0.029	0.091
$t_9$	0.117	0.020	0.051	0.112	0.207	0.325	0.057	0.111
$t_{10}$	0.105	0.014	0.042	0.090	0.218	0.293	0.115	0.122

$s_1$	$s_2$	$s_3$	$s_4$	$s_5$	$s_6$	$s_7$	$s_8$
0.211	0.034	0.067	0.097	0.141	0.234	0.121	0.096

Table 5.1: Strategy Distributions in ESCConv++ Dataset. Top shows strategy distribution across 8 strategies per turn averaged across conversations for the first 10 tens. Bottom shows averaged strategy distribution across the entire dataset.

2. To implement this framework, we contribute a novel dataset, ESCConv++, with problem and strategy distribution annotations per utterance.
3. We also propose an architecture which follows the PRISM framework using Transformer encoder-decoder models with 4 ESC-specific loss functions: Problem prediction loss, Strategy prediction loss, Problem-Strategy mapping loss, and Cross-entropy loss.
4. On ESCConv++, we beat state-of-the-art by 2.94 on BLEU-1, 1.58 on BLEU-2, and 0.89 on METEOR.

### 5.3 ESC Problem Formulation

Consider a dialogue with  $i$  utterances between the help-seeker and the supporter so far. We refer to the first  $i - 1$  utterances as dialog history represented as  $H = [X_1, X_2, \dots, X_{i-1}]$ . The last help-seeker utterance, i.e., the  $i$ -th utterance  $X_i = [w_{i1}, w_{i2}, \dots, w_{im}]$  is a sequence of  $m$  words. For utterance  $X_{i+1}$  by the supporter after the help-seeker’s utterance, we have  $P_i$  (a  $P$ -dimensional probability vector) representing the problem state and  $S_i$  (an  $S$ -dimensional probability vector) representing the ideal strategy representation to use, given the context until  $X_i$  including it.  $S_i$  and  $P_i$  can be used for turn-level supervision. The goal is to generate the next utterance  $X_{i+1}$  from the standpoint of the supporter that is coherent with the dialogue history  $H$  and supportive to reduce the help-seeker’s distress, taking into consideration both the problem state  $P_i$  as faced by the help-seeker and the strategies to use based on  $H$ . The goal is to estimate a function  $p(X_{i+1} | [X_1, X_2, \dots, X_i], P_i, S_i)$  that generates the target response  $X_{i+1}$ . In this paper, we consider  $P = 10$  problem types and  $S = 8$  support strategies as mentioned in Section 5.5.1.

	$p_1$	$p_2$	$p_3$	$p_4$	$p_5$	$p_6$	$p_7$	$p_8$	$p_9$	$p_{10}$
$t_1$	0.350	0.120	0.131	0.074	0.059	0.047	0.076	0.059	0.045	0.039
$t_2$	0.354	0.164	0.180	0.078	0.054	0.034	0.054	0.031	0.028	0.024
$t_3$	0.312	0.196	0.213	0.083	0.056	0.037	0.045	0.020	0.023	0.017
$t_4$	0.304	0.199	0.214	0.083	0.055	0.036	0.051	0.019	0.024	0.016
$t_5$	0.297	0.212	0.214	0.079	0.056	0.036	0.052	0.019	0.021	0.013
$t_6$	0.289	0.210	0.209	0.087	0.060	0.035	0.053	0.021	0.022	0.014
$t_7$	0.284	0.211	0.217	0.085	0.059	0.036	0.056	0.018	0.021	0.013
$t_8$	0.276	0.217	0.218	0.084	0.056	0.035	0.061	0.022	0.021	0.013
$t_9$	0.281	0.211	0.214	0.084	0.057	0.037	0.060	0.024	0.020	0.014
$t_{10}$	0.272	0.209	0.209	0.089	0.061	0.037	0.063	0.024	0.024	0.014

$p_1$	$p_2$	$p_3$	$p_4$	$p_5$	$p_6$	$p_7$	$p_8$	$p_9$	$p_{10}$
0.299	0.199	0.199	0.085	0.057	0.037	0.057	0.025	0.024	0.017

Table 5.2: Problem Distributions in ESCov++ Dataset. Top shows problem distribution across 10 strategies per turn averaged across conversations for the first 10 tens. Bottom shows averaged problem distribution across the entire dataset.

Broadly, depending on the type of the problems mentioned by the help-seeker, the supporter modulates the conversation by using a certain set of strategies. We design a neural architecture to mimic this process. Hence, we further define the the following four sub-tasks to solve the emotional support conversation problem.

1. Problem Prediction: Predicts the 10-category problem vector that can be regarded as the fine-grained probabilistic state of the problem faced by the help-seeker based on the context so far.
2. Strategy Prediction: Predicts the 8-category support strategy vector that can be regarded as the fine-grained probabilistic strategy.
3. Problem-Strategy Alignment: This involves ensuring the alignment between the predicted fine-grained problem and the corresponding strategy. This is achieved through the utilization of a predefined problem-strategy mapping matrix, which has been learned based on empirical data derived from ground truth problem and strategy annotations.
4. Response Generation: Generate the response  $X_{i+1}$  based on the predicted problem and the predicted strategy to use.

## 5.4 Brief Description of Baselines

1. **MultiESC** [8] is an emotional support conversation system, which conducts dialogue strategy planning to guide utterance generation. It adopts an A\*-like algorithm to select the adopted dialogue strategy by learning a strategy scoring function that comprehensively considers a history-based score and a lookahead score indicating the expected user feedback.

2. **KEMI** [10] is an emotional support conversation system, which retrieves external knowledge from a mental health knowledge graph to enhance the system. It also conducts multi-task learning of dialogue strategy learning and response generation together.
3. **MoEL** [38] adopts several decoders focusing on different types of emotional utterances, whose outputs are combined to generate the final utterances.
4. **MIME** [47] follows the architecture of MoEL and adds extra mechanisms to combine the results from different decoders.
5. **DialoGPT-Joint** and **BlenderBot-Joint** [42] are developed on the backbones of DialoGPT [81] and BlenderBot [58], respectively. They prepend a special token, denoting the predicted support strategy, before the generated utterance to generate content conditioned on a predicted strategy.
6. **EmpDG** [33] learns how to generate responses consistent with the users emotion via an adversarial learning framework.
7. **MISC** [70] enhances context encoding with commonsense knowledge and uses the predicted strategy distribution to guide the emotional support dialogue generation. It predicts the strategy distribution using a vanilla Transformer encoder.
8. **TransESC** [84] looks at turn-level transitions from three perspectives, including semantics transition, strategy transition and emotion transition, to drive the conversation in a smooth and natural way. Specifically, they construct the state transition graph with a two-step way, named transit-then-interact, to grasp such three types of turn-level transition information. Finally, they are injected into the transition-aware decoder to generate more engaging responses.
9. **SUPPORTER** [90] is a mixture-of-expert-based reinforcement learning model with emotional support and dialogue coherence rewards to guide policy’s learning for responding.

## 5.5 ESConv++ Dataset

Our research expands on the framework of emotional support dialogues using the ESConv dataset developed by Liu et al. [42]. This dataset consists of extensive English conversations, averaging nearly 30 utterances each. The dialogues are annotated for various support strategies, including questioning, reflection of feelings, and self-disclosure among others. Help-seekers rated feedback scores on a five-point Likert scale to assess the effectiveness of emotional support provided during these dialogues. The dataset also contains a “situation” statement per dialogue describing the background behind the conversation. Lastly, each utterance is also labeled with an emotion.

To enhance the ESConv dataset, we have incorporated a new probabilistic annotation feature. We utilized GPT-4 to assign probability vectors for each utterance, which provides a range of potential strategic intents and problem recognitions in supporter responses and help-seeker disclosures. This probabilistic approach improves upon the original dataset by introducing

	Train	Dev	Test
# Dialogues	1040	130	130
# Utterances	24373	3007	3012
Avg # words per utterance	21.04	21.72	20.79
Avg # utt per dialogue	23.44	23.13	23.17
Avg # words per dialogue	493.16	502.51	481.74

Table 5.3: ESConv++: Dataset Statistics

more depth and variability to the annotation process, thereby enriching subsequent analyses. This level of detail allows for a more advanced understanding of how conversational strategies interact with problem states. We leveraged large language model (LLM)-based annotations since it has been shown that the zero-shot accuracy of ChatGPT exceeds that of crowd workers by significant amounts on average, while ChatGPTs inter-annotator agreement exceeds that of both crowd workers and trained annotators on multiple NLP tasks [15].

**Strategy Annotation:** Following each help-seeker’s utterance  $X_i$ , the subsequent supporter’s response  $X_{i+1}$  is annotated with a probabilistic strategy distribution. This annotation is performed by supplying the context  $H_{i+1} = [X_1, X_2, \dots, X_{i+1}]$  to GPT-4, along with a finely-tuned prompt detailing the support strategies. The model generates an 8-class probability distribution corresponding to the identified support strategies employed in  $X_{i+1}$ .

**Problem Categorization:** Similarly, for each dialogue history  $H_i$ , where  $X_i$  is an utterance by the help-seeker, GPT-4 is utilized to annotate the conversation history  $H_i$  with a 10-class probability distribution of high-level problem categories. The annotation is guided by a fine-tuned prompt encompassing the details of these problem categories, derived from an in-depth analysis of the ESConv conversations.

The 10 problem categories were specifically designed to reflect the diverse range of issues discussed in the ESConv dataset. This approach not only facilitates a more granular understanding of the conversation dynamics but also enriches the dataset with nuanced annotations that capture the complexity of emotional support conversations.

Table 5.3 shows the basic statistics of the generated ESConv++ dataset. Tables 5.1 and 5.2 shows the strategy and problem distributions in the ESConv++ Dataset. Bottom row in table 5.1 shows that  $s_6$ =“Providing Suggestions or Information” and  $s_1$ =“Questioning” are the most popular strategies used. Similarly, bottom row in table 5.2 shows that  $p_1$ =“Emotional problems”,  $p_2$ =“Relationship problems” and  $p_3$ =“Work, school, and career concerns” are the most popular problem types. Top part in table 5.1 shows that typically conversations start with strategy  $s_7$ =“Greetings” (similarly they end with this strategy as well, but maybe not apparent from this table as it has just 10 turns), and then supporters use strategy  $s_1$ =“Questioning” for the next 3-4 turns before finally switching to strategy  $s_6$ =“Providing Suggestions or Information”. Of course, in some conversations, supporters also leverage strategies  $s_3$ ,  $s_4$ ,  $s_5$  and  $s_8$  in the later turns instead of just using  $s_6$ . Right top shows that help-seekers mostly start with  $p_1$ =“Emotional problems” and slowly start talking about  $p_2$ =“Relationship problems”,  $p_3$ =“Work, school, and career concerns” and other types as the conversation progresses.

## 5.5.1 Annotation Process

### 5.5.1.1 Definitions of Strategies

The counseling strategies used in this study are adapted and modified from [42], with an additional strategy, *Greetings*, introduced in [8]. The strategies include *Questioning*, *Restatement or Paraphrasing*, *Reflection of Feelings*, *Self-disclosure*, *Affirmation and Reassurance*, *Providing Suggestions or Information*, *Greetings*, and *Others*. Detailed definitions of each strategy, as provided to GPT-4, are based on enriched versions of the original descriptions from [42] and [8].

1. **Questioning:** This strategy involves using both open-ended and closed questions to delve into the client's experiences, thoughts, and emotions. Open-ended questions are used to encourage the client to share more about their feelings and experiences, promoting exploration and understanding. Closed questions, on the other hand, are used to gather specific details, aiding in precise information collection and understanding.
2. **Restatement or Paraphrasing:** Here, the supporter carefully reinterprets and rephrases the client's statements, thereby showing understanding, empathy, and active listening. This strategy assists the client in achieving a deeper understanding and clarity of their feelings and thoughts, and also assures them that their concerns are being comprehended accurately.
3. **Reflection of Feelings:** The supporter identifies and articulates the client's emotions, creating a validating and empathetic environment that encourages emotional exploration. This strategy emphasizes emotional understanding, promoting trust, and providing emotional support.
4. **Self-disclosure:** This strategy involves the intentional sharing of personal experiences, feelings, or thoughts by the supporter. This approach helps establish rapport and trust, normalizes experiences, and encourages client openness. It's important, however, for self-disclosure to be implemented judiciously, maintaining ethical boundaries, and ensuring it serves the therapeutic goal.
5. **Affirmation and Reassurance:** This strategy acknowledges and validates the client's efforts, strengths, and capabilities. It offers comfort, motivation, and emotional support during challenging times, boosting the client's resilience and self-esteem, and encouraging continued effort and growth.
6. **Providing Suggestions or Information:** Here, the supporter offers alternative perspectives, advice, or suggestions for change, while ensuring respect for the client's autonomy and self-determination. This helps promote self-reflection, self-awareness, and personal growth. This strategy also includes cases where the supporter provides accurate, relevant, and helpful knowledge, information, and resources to assist the client in making informed decisions and gaining a clearer understanding of their concerns, options, and potential solutions.

7. **Greetings:** This strategy focuses on the exchange of pleasantries and social niceties within a conversation. It involves the supporter engaging in polite and friendly interactions, such as welcoming the client, inquiring about their well-being, or offering warm farewells. Greetings set the tone for a respectful and empathetic dialogue, facilitating a sense of comfort and ease.
8. **Others:** This category includes various additional support strategies that don't neatly fit into the specific techniques above. These could involve building rapport, summarizing, empowering the client, and facilitating problem-solving. They highlight the importance of flexibility, adaptability, and personalization in the counseling relationship.

### 5.5.1.2 Definitions of Problem Categories

Our methodology involved constructing a taxonomy of ten problem categories to encapsulate the spectrum of challenges encountered by individuals seeking support. This taxonomy was assessed and endorsed by a psychologist with a doctoral degree, ensuring the categories accurately represent the breadth of issues addressable with general helping skills within the scope of the ESConv dataset.

1. **Emotional problems:** These relate to feelings of anxiety, depression, stress, mood swings, grief, or any form of emotional trauma.
2. **Relationship problems:** These involve issues arising from interactions with family, romantic partners, friends, or other social connections, including communication difficulties.
3. **Work, school, and career concerns:** These pertain to problems related to job stress, academic pressure, career transitions, unemployment, or challenges balancing work and personal life.
4. **Health and well-being:** These encompass physical health issues, mental health concerns, challenges in coping with illness or disability, sleep problems, or lifestyle-related issues.
5. **Life transitions and adjustments:** These include dealing with new life circumstances, major life events, or personal growth and developmental challenges.
6. **Identity and self-esteem:** These involve issues related to self-worth, self-esteem, gender identity, sexual orientation, racial or cultural identity, or personal values and beliefs.
7. **Coping strategies and resilience:** These refer to stress management, developing healthy coping mechanisms, building resilience, mindfulness, and self-care practices.
8. **Social and communication skills:** These are about assertiveness, setting boundaries, conflict resolution, effective communication, or the development and maintenance of social skills and relationships.
9. **Emotional regulation and self-awareness:** These involve anger management, understanding and expressing emotions, developing self-awareness, or improving emotional intelligence.

10. **Personal aspirations and goals:** These relate to goal setting, motivation, overcoming obstacles, time management, or organization.

### 5.5.2 Annotation Methodology

Our annotation process utilized GPT-4 turbo, focusing on two main aspects: strategy annotation and problem annotation. The conversation context and the approach for each annotation type varied to optimize accuracy.

**Strategy Annotation:** The entire conversation up to the last utterance by the supporter was used as context for strategy annotation. This comprehensive approach provided better insight into the strategy used in the final utterance.

**Problem Annotation:** For problem annotation, the conversation context included all exchanges prior to the current supporter utterance. This facilitated a more accurate understanding of the help-seeker’s problem.

We developed specific prompt templates for annotating both strategies and problems in the dataset. We list them down in Appendix B.3. The prompt templates were continuously refined based on manual evaluations of random dataset samples, ensuring the highest quality and relevance of the annotations.

## 5.6 Why do we need Problem Annotations?

The ESConv paper [42] annotated problems at the conversation level, relying on pre-chat surveys for coarse-grained problem identification, which is impractical in real-life settings where supporters must discern problems during the conversation. Additionally, such annotation does not facilitate problem-conditioned generation due to its broad, single-category per conversation approach.

Our model aims to simulate the dynamic comprehension of problems a supporter develops through conversation, necessitating fine-grained problem annotations. Liu et al. [42] initially categorized issues into five broad topics:

1. Ongoing Depression
2. Job Crisis
3. Breakup with Partner
4. Problems with Friends
5. Academic Pressure

This categorization falls short of capturing the nuanced problems help-seekers face, achievable only through detailed problem annotation.

Consider the following case study illustrating the evolving problem understanding in a conversation:

**Help-seeker:** "Hey." (neutral)

**Supporter:** "Hello! How are you doing today?" (neutral)

**Help-seeker:** "I'm doing alright, trying to stay cheerful." (caring)

(Problems - 60% Emotional Problems, 40% Coping strategies and resilience)

**Supporter:** "That's great, especially around this time of the Holidays." (admiration)

**Help-seeker:** "Yeah, it's good because of the holidays." (admiration)

(Problems - 30% Emotional Problems, 30% Coping strategies and resilience, 30% Emotional Regulation and self-awareness)

**Supporter:** "So, how are things in your life?" (curiosity)

**Help-seeker:** "Not too bad, just had a death in the family. Pretty bummed about it because we are close to our extended family and can't visit one another to grieve." (sadness)

(Problems - 40% Emotional Problems, 20% Life transitions and adjustment, 20% Relationship problems)

**Supporter:** "I'm very sorry to hear that. Not being able to be with family when something like that happens is incredibly difficult, those are the people you want to be around the most." (sadness)

This exchange demonstrates how a supporter's understanding of a help-seeker's problems evolves, contrasting with the "ongoing depression" label for the conversation in ESConv.

Table 5.4: Quantitative analysis of problems within ESConv++ dialogues.

Prob	1	2	3	4	5	6	7	8	9	10
0.1	1006	3463	4121	3826	1466	413	81	28	21	0
0.15	3044	8607	2593	150	11	1	0	0	0	0
0.2	4001	8434	1786	41	5	0	0	0	0	0
0.25	8113	5848	54	7	0	0	0	0	0	0
0.3	8473	5027	17	0	0	0	0	0	0	0
0.4	10395	1155	0	0	0	0	0	0	0	0

The table 5.4 lists the count of samples in the dataset that feature exactly "n" number of problems with a probability greater than or equal to the value specified in the first column. This table succinctly demonstrates how multiple problems coexist within conversations, underscoring the dynamic problem landscape in ESConv dialogues.

According to Liu et al. [42], the initiation phase where a supporter explores the help-seeker's problems is critical. This exploration phase is foundational for a supporter to grasp the nuanced experiences and emotions of the help-seeker. Without this thorough understanding, there exists a risk of providing advice that is not only irrelevant but potentially detrimental, such as suggesting simplistic solutions like taking a walk to relax.

The process of engaging with the help-seeker involves a supporter employing various empathetic and understanding techniques to alleviate the help-seeker's distress. These techniques, which include self-disclosure and reflection of feelings among others, are instrumental in forming a supportive and empathetic connection. Following this empathetic engagement, the supporter is in a position to offer more tailored and effective coping strategies to the help-seeker.



It is imperative to note that the strategies adopted by ESConv were selectively curated from a broader set of skills proposed by Hill (2009) [24], tailored specifically for the context of emotional support conversations. Hill (2009) outlines a three-stage model of helping that includes exploration, insight, and action. This framework advocates for interventions to be customized to meet the unique needs and experiences of the individual seeking help, supporting the idea of aligning problems with strategies. We argue that Hill’s proposition—that helpers adapt their techniques to suit the specific problems and needs of the help-seeker—provides a solid foundation for investigating a fine-grained modeling of problems. We have also empirically demonstrated that our approach, with its fine-grained understanding of problems and alignment of problems with strategies, performs better than existing baselines.

The consultation with a psychologist who holds a doctoral degree aimed to ensure that our categorization accurately reflects the wide range of issues that can be addressed using general helping skills, as encompassed within the ESConv dataset.

## 5.7 PRISM Model Architecture

Fig. 5.2 shows the overview of our approach. Our model, PRISM, is trained in two stages. In the first stage, a problem-strategy mapping matrix is trained. In the second stage the dialog model is trained using the mapping matrix.

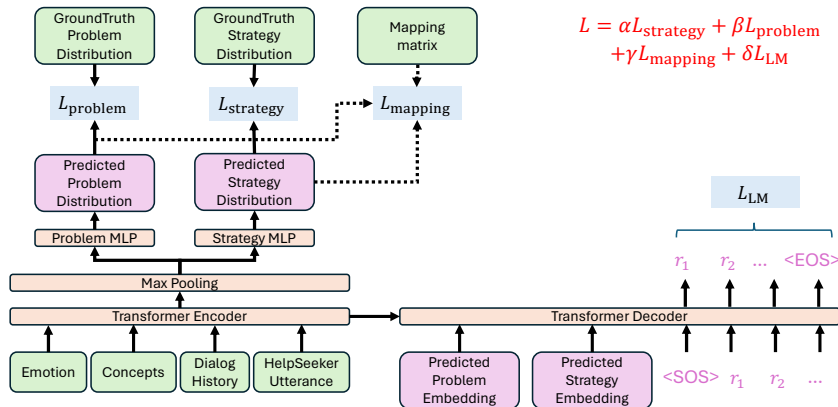


Figure 5.2: PRISM: Model Architecture

### 5.7.1 Stage 1: Learning Problem-Strategy Mapping

Given a distribution over problem types demonstrated in the latest help-seeker utterance, the supporter responds with a response sampled from a distribution over strategies. What strategies are good to use and which ones should be avoided given a problem type? To capture this knowledge, in this stage, we aim to learn a linear relationship between a problem type and a strategy type. This relationship is captured through the construction of a linear problem-strategy mapping matrix  $\mathbf{M} \in \mathbb{R}^{P \times S}$ .

This matrix  $\mathbf{M}$  should justify the relationship between problem and strategy distributions used across every turn in training data, i.e.,  $\mathbf{P}_i^T \times \mathbf{M} \approx \mathbf{S}_i$  for every turn  $i$  in the train set.

Further, each row in this mapping matrix should add up to 1, i.e.,  $\sum_{j=1}^S \mathbf{M}_{ij} = 1 \quad \forall i = 1..P$ , where  $M_{ij}$  denotes the weight corresponding to the  $j$ -th strategy for the  $i$ -th problem,  $P$  is the number of problem types, and  $S$  is the number of strategy types. Thus, learning the matrix  $\mathbf{M}$  reduces to solving the following optimization problem.

$$\begin{aligned} & \min_{\mathbf{M}} \sum_{i=1}^N \sum_{j=1}^S (\mathbf{S}_i[j] - \mathbf{P}_i \cdot \mathbf{M}_{.j})^2 \\ & \text{subject to the following constraints} \\ & \mathbf{M}_{ij} \geq 0 \quad \forall i = 1..P, \forall j = 1..S \\ & \sum_{j=1}^S \mathbf{M}_{ij} = 1 \quad \forall i = 1..P \end{aligned}$$

It is difficult to deal with the external constraint  $\sum_{j=1}^S \mathbf{M}_{ij} = 1 \quad \forall i = 1..P$  efficiently. Hence, we transform it to a ‘‘soft’’ constraint, i.e., we implicitly enforce the constraint by adding a penalty term  $\psi \sum_{i=1}^P \left[ \sum_{j=1}^S \mathbf{M}_{ij} - 1 \right]$  to the objective function.

$$\begin{aligned} & \min_{\mathbf{M}} \sum_{i=1}^N \sum_{j=1}^S (\mathbf{S}_i[j] - \mathbf{P}_i \cdot \mathbf{M}_{.j})^2 \\ & + \psi \sum_{i=1}^P \left| \sum_{j=1}^S \mathbf{M}_{ij} - 1 \right| \\ & \text{subject to the following constraints} \\ & \mathbf{M}_{ij} \geq 0 \quad \forall i = 1..P, \forall j = 1..S \end{aligned}$$

where  $\psi$  is a constant which we set to 1. This optimization problem is then solved using gradient descent iterations as discussed in Long et al. [44]. An important consideration is that a post-optimization step is implemented, wherein negative values in the matrix  $\mathbf{M}$  are clamped to zero after every optimization step. This procedure is critical for maintaining the non-negativity of the matrix, which is essential for the probabilistic interpretations of our model.

In Table 5.5, we present a visualization of the mapping matrix  $\mathbf{M}$ , which facilitates an intuitive comprehension of the associations between problems and strategies. The construction

	$s_1$	$s_2$	$s_3$	$s_4$	$s_5$	$s_6$	$s_7$	$s_8$
$p_1$	0.339	0.041	0.083	0.102	0.112	0.191	0.048	0.084
$p_2$	0.162	0.051	0.101	0.085	0.148	0.312	0.009	0.131
$p_3$	0.165	0.038	0.065	0.106	0.179	0.332	0.023	0.093
$p_4$	0.083	0.018	0.034	0.193	0.177	0.315	0.035	0.145
$p_5$	0.145	0.031	0.133	0.126	0.238	0.115	0.123	0.089
$p_6$	0.169	0.031	0.041	0.072	0.211	0.171	0.169	0.136
$p_7$	0.056	0.000	0.000	0.083	0.192	0.253	0.339	0.076
$p_8$	0.372	0.000	0.000	0.000	0.000	0.106	0.521	0.000
$p_9$	0.415	0.038	0.023	0.033	0.000	0.068	0.357	0.064
$p_{10}$	0.368	0.000	0.000	0.082	0.071	0.000	0.476	0.002

Table 5.5: Problem-Strategy Mapping in ESConv++

of this matrix is a key component in PRISM, enabling the system to align support strategies effectively with identified problems. This alignment is crucial for the relevance and efficacy of the emotional support provided. An interesting observation from Table 5.5 is that besides greeting,  $s_1$ ="Questioning" is a popular strategy for problems of type  $p_1$ ="Emotional problems",  $p_8$ ="Social and communication skills",  $p_9$ ="Emotional regulation and self-awareness" and  $p_{10}$ ="Personal aspirations and goals".

### 5.7.2 Stage 2: Training the Dialog Model

In this stage we train Transformer-based encoder-decoder models like BART [32] (similar to state-of-the-art work, MultiESC [8]) to holistically learn the dynamics of emotional support conversations and generate response  $X_{i+1}$  given dialogue history  $H_i$  and current help-seeker utterance  $X_i$ . This involves simultaneous learning of strategy prediction, problem identification, alignment of strategy and problem, and response generation as shown in Fig. 5.2.

**Transformer encoder input:** The encoder input consists of four parts separated by a [SEP] token: emotion, concepts, dialog history  $H_i$  and help-seeker utterance  $X_i$ . Emotion at the utterance level is derived using a pre-trained off-the-shelf emotion detection model<sup>1</sup>. For extracting concepts, we harness ConceptNet [63] to identify concepts relevant to the conversational context, prioritizing those with the highest emotional intensity values following the methodology proposed by Li et al. [34]. Overall we obtain concepts from  $X_i$ , concepts from  $H_i$  and concepts from the situation statement associated with the dialogue.

**Transformer decoder input:** The decoder input consists of a predicted problem embedding, predicted strategy embedding and shifted outputs. The predicted problem embedding is obtained as a weighted sum of phrase embeddings of different problem types. Thus, predicted problem embedding is computed as  $\sum_{j=0}^P(\text{PhraseEmbedding}(\text{problem}[j]) \times P_i[j])$  where  $\text{problem}[j]$  is the text associated with the  $j$ -th problem type, and  $\text{PhraseEmbedding}$  is computed as the average of constituent token embeddings. Similarly, predicted strategy embedding is computed as  $\sum_{j=0}^S(\text{PhraseEmbedding}(\text{strategy}[j]) \times S_i[j])$ .

**Losses to be optimized:**

1. Strategy Prediction Loss ( $L_{\text{strategy}}$ ): We derive a predicted strategy distribution  $\hat{\mathbf{S}}_i$  by max-pooling the Transformer encoder’s hidden states, applying dense fully connected layers, and lastly performing a softmax.  $L_{\text{strategy}}$  is computed as the symmetric Kullback-Leibler (KL) Divergence between the predicted strategy vector  $\hat{\mathbf{S}}_i$  and the ground-truth strategy annotation  $\mathbf{S}_i$ .
2. Problem Prediction Loss ( $L_{\text{problem}}$ ): Analogous to  $L_{\text{strategy}}$ ,  $L_{\text{problem}}$  using KL between the predicted problem vector  $\hat{\mathbf{P}}_i$  and the ground-truth problem annotation  $\mathbf{P}_i$ .
3. Problem-Strategy Mapping Loss ( $L_{\text{mapping}}$ ): This loss ensures congruence between the predicted problem and strategy vectors and the established problem-strategy mapping matrix.  $L_{\text{mapping}}$  is also calculated using KL Divergence between  $\hat{\mathbf{P}}_i^T \mathbf{M}$  and  $\hat{\mathbf{S}}_i$ .

<sup>1</sup>[https://huggingface.co/SamLowe/roberta-base-go\\_emotions](https://huggingface.co/SamLowe/roberta-base-go_emotions)

4. Language Modeling Loss ( $L_{LM}$ ): The model is trained to generate the ground-truth response  $X_{i+1}$  using teacher-forcing. The loss here,  $L_{LM}$ , is the cross-entropy loss between the generated response and the actual  $X_{i+1}$ .

The overall loss function for training is defined as a weighted sum of  $L_{strategy}$ ,  $L_{problem}$ ,  $L_{mapping}$ , and  $L_{LM}$ .

$$L_{total} = \alpha L_{strategy} + \beta L_{problem} + \gamma L_{mapping} + \delta L_{LM}$$

During training, embeddings corresponding to problem states and strategies are generated. These embeddings, dimensionally equivalent to the model’s tokens, are incorporated into the model for predicting subsequent tokens, thereby embedding context about problem state information and strategies to use for response generation.

## 5.8 Experiments and Results

The experimental setup details and hyperparameters are described in detail in Appendix B.1.

Model	BLEU-1	BLEU-2	BLEU-3	METEOR	PPL	D-1	D-2
MIME [47]	15.24	5.56	2.64	6.43	69.28	2.11	17.72
MISC [70]	16.31	7.31	3.26	7.76	16.16	4.41	19.71
DialoGPT-Joint [42]	17.39	5.59	2.03	7.55	15.71	-	-
BlenderBot-Joint [42]	17.62	6.91	2.81	7.54	16.79	4.12	17.72
TransESC [84]	17.92	7.64	4.01	-	15.85	4.73	20.48
EmpDG [33]	18.08	6.46	3.02	6.93	115.34	-	-
MoEL [38]	19.04	6.47	2.91	7.96	264.11	2.33	15.26
SUPPORTER [90]	19.50	7.49	3.58	-	15.37	<b>4.93</b>	<b>27.73</b>
KEMI [10]	20.48	8.55	4.50	8.32	<b>14.29</b>	4.43	23.72
MultiESC [8]	21.65	9.18	4.99	8.84	15.41	2.77	15.26
PRISM (Ours)	<b>24.59</b>	<b>10.76</b>	<b>5.73</b>	<b>9.73</b>	17.44	3.43	15.17

Table 5.6: Comparison of various methods for ESC problem on ESCConv++. Empty cells indicate that those results were not available.

### 5.8.1 Metrics

We use a standard set of evaluation metrics (BLEU, METEOR, Perplexity, ..) used by baselines for automatic evaluation. BLEU [51] and METEOR [3] measure lexical-overlap between n-grams of the predicted and ground truth response. Distinct-n [40] measures n-gram diversity in generated responses.

### 5.8.2 Main results

Table 5.6 shows our main results where we compare the performance of our proposed method, PRISM, with 10 other baseline methods on the ESCConv++ test set for the ESC task. Although

the dataset also has ground truth strategy and problem annotations, we use the predicted vectors for this evaluation. PRISM is the best performing model on most of the metrics, especially BLEU-1, BLEU-2, BLEU-3, and METEOR, which measure the n-gram overlap and semantic similarity between the generated responses and the references. This suggests that PRISM can generate more relevant and fluent responses than the other models.

MultiESC is the second best performing model on most of the metrics; however, we beat them categorically on all automated evaluation metrics except perplexity. KEMI is the third best performing model on most of the metrics and the model that achieves the lowest perplexity, which measures the uncertainty of the model on the test data. This implies that KEMI can generate more predictable and consistent responses than the other models but may lack some diversity and creativity. SUPPORTER is the model that achieves the highest Diversity scores, which measure the lexical variety and information richness of the generated responses. This shows that SUPPORTER can generate more novel and informative responses than the other models but may compromise some relevance and fluency. However, as described in 5.8.3, SUPPORTER is a significantly bigger model. It is to be noted that we categorically beat them on metrics such as BLEU and METEOR, which correlate better with human perception of quality compared to diversity. Our results are statistically significantly better (2-tailed t-test with 95% confidence level).

We show detailed examples of generations from PRISM in Appendix B.2. We also show other details like problem distribution, strategy distribution, concepts, etc.

PRISM	BLEU-1	BLEU-2	BLEU-3	METEOR	PPL	D-1	D-2
Predicted	<b>24.59</b>	10.76	5.73	9.73	17.44	3.43	15.17
Oracle	24.21	<b>10.77</b>	<b>5.83</b>	<b>9.83</b>	<b>16.90</b>	<b>3.93</b>	<b>17.39</b>

Table 5.7: Oracle Analysis

**Results with Oracle Distributions:** We also have the annotated problem and strategy distributions for every utterance in the test set. How do the results look like if use those oracle distributions rather than predicted ones? Table 5.7 shows that the oracle model achieves the slightly lower BLEU-1 score but slightly higher BLEU-2, BLEU-3, METEOR, PPL, and diversity scores. This indicates that the oracle distributions have slightly better effects on the fluency and coherence of the generated responses than the predicted distributions. Also the perplexity with oracle is slightly lower, implying that the oracle distributions have slightly better effects on the predictability and consistency of the generated responses than the predicted distributions.

### 5.8.3 Analysis of Resource Requirements for the Deployment of Baselines compared to PRISM

Among the baselines, the SOTA baseline MultiESC has Bart as the backbone. However, a more critical distinction between baselines lies in the total model size required during inference, which provides a more comprehensive understanding of fairness in comparison.

1. **PRISM**: Utilizes BART (140M parameters) along with an off-the-shelf emotion detection model based on Roberta-base, adding another 120M parameters, totaling approximately 260M parameters.
2. **MultiESC** [8]: Employs BART (140M) for both the encoder and decoder, alongside an off-the-shelf emotion detection model with 120M parameters (backbone: roberta-base). Additionally, it includes a user feedback predictor and a strategy sequence predictor, each utilizing BART (140M). Although not specified in the paper, their code reveals the presence of three independently trained models initialized with BART weights, resulting in a total model size of  $\approx 540\text{M}$  ( $140\text{M} \times 3 + 120\text{M}$ ) parameters.
3. **KEMI** [10]: Operates on a smaller BlenderBot backbone ( $\approx 90\text{M}$  parameters), supplemented by Sentence-BERT for subgraph retrieval (22M, "all-MiniLM-L6-v2" - authors mentioned this in their codebase) and COMET [5] for commonsense knowledge (137M, likely using the GPT-2 initialized version. This is also the smallest COMET model), culminating in  $\approx 249\text{M}$  parameters.
4. **Supporter** [90]: Combines BART (140M) with a COMET [26] model based on bart-large (440M), an emotion classification model (82.8M), and Dialogue Coherence Reward Models (109M), leading to a substantial total of  $\approx 771.8\text{M}$  parameters.

Moreover, the COMET models employed by KEMI and Supporter are generative, necessitating significantly higher computational resources compared to the emotion detector used in PRISM. This context highlights that **our baselines, in fact, have around the same or more parameters than PRISM during inference, affirming the fairness and rigor** of our comparative evaluation. The computational requirements during PRISM are much lower than Supporter and MultiESC, but similar to KEMI.

#### 5.8.4 Performance of Finetuned Large Language Models in ESC

Our approach focused on employing smaller pre-trained language models (PLMs) for this task. This strategy aimed at achieving methodological improvements and facilitating direct comparisons with established baselines, all of which utilize models with fewer than 150M parameters as backbones. Most prior works also develop effective architectures to integrate external knowledge for generating more informative and effective responses.

It was shown in Zhao et al. [83] that the one-shot performance of ChatGPT (GPT-3.5-turbo) in ESC was significantly lower compared to state-of-the-art methods with small PLMs like BlenderBot and Bart. ChatGPT received significantly lower BLEU scores (B-1: 13.91, B-2: 4.53, and B-3: 1.96) but better distinctness scores (D-1: 5.92 and D-2: 31.38). Therefore, the question essentially boils down to the performance of these Large Language Models (LLMs) with billions of parameters after fine-tuning.

Unfortunately, due to compute constraints in our lab, which is equipped with RTX 2080Tis having 11GB VRAM, we could not fine-tune a model like Llama-2 or even Gemma-2b with architectural changes. We also think that practical deployment environments like apps on mobile phones will also not be able to handle such large models. However, we conducted

experiments using 4-bit QLoRA fine-tuning with the ESConv dataset, on the LLMs Llama-2 [68], Mistral [28], and Gemma [67].

Model	BLEU-1	BLEU-2	BLEU-3	METEOR	D-1	D-2
Mistral-7b	23.74	11.54	6.72	9.64	4.35	20.43
Mistral-7b-Instruct	23.61	11.37	4.09	10.06	3.96	19.39
Gemma-2b	24.32	10.8	5.81	9.66	5.78	31.36
Llama-2-13b	23.48	10.46	5.68	9.48	6.12	33.24
MultiESC	21.65	9.18	4.99	8.84	2.77	15.26
PRISM (Ours)	24.59	10.76	5.73	9.73	3.43	15.17

Table 5.8: Comparison of models on various metrics.

Our experimental setup utilized a single GPU (RTX 2080 Ti), training each model for 3 epochs with a total batch size of 16 (using gradient accumulation), totaling 2304 steps. We evaluated after every 100 steps, selecting the checkpoint with the best BLEU-1 score on the validation set for testing, similar to the strategy followed in PRISM.

We can note that no LLM we finetuned outperforms PRISM across metrics. Llama-2-13b’s notably lower score for BLEU-1 and METEOR, along with BLEU-2,3 scores being similar to PRISM, despite it having 13 billion parameters (over 90 times larger than PRISM’s), strongly showcases the effectiveness of our proposed framework, PRISM.

### 5.8.5 Ablations

To understand the importance of each novel part of the proposed framework, we perform ablation experiments where we train different variants of PRISM and show results in Table 5.9. The variants are obtained by removing one or more components of the model, such as Problem, Strategy, Mapping, Concepts, Emotion, and PSM. Clearly the full model is the best performing model on most of the metrics, especially BLEU-1, BLEU-2, and METEOR. The variant without Problem distribution is the worst performing model on most of the metrics, especially BLEU-1, BLEU-2, and METEOR. The model without Strategy distribution is the second worst performing model on most of the metrics. This indicates that the predicted Problem and Strategy distributions indeed help the model to generate more relevant and fluent responses; and are therefore crucial in generating empathetic responses. The model without the Mapping based loss is the model that achieves the lowest Diversity scores. This implies that interestingly using the Mapping matrix not just helps generate relevant responses but also brings in diversity as a side effect. Removing problem and strategy distributions both increase diversity while removing mapping based loss reduces diversity. Thus, it looks like the trio of problem, strategy and mapping matrix are playing a significant role in generating relevant and diverse responses. The full model therefore balances both relevance and diversity.

We also compare the results for PRISM when using BART vs. Blenderbot [58] as the base model. From Table 5.10, we observe that PRISM with BART performs significantly better.

Model	BLEU-1	BLEU-2	BLEU-3	METEOR	PPL	D-1	D-2
PRISM	<b>24.22</b>	<b>9.80</b>	<b>4.91</b>	<b>9.47</b>	17.43	4.05	20.88
w/o Problem	16.66	6.03	2.80	7.21	16.93	<b>4.75</b>	23.11
w/o Strategy	19.32	7.19	3.42	7.96	17.03	4.90	<b>23.91</b>
w/o Mapping	22.62	8.17	3.68	8.84	17.65	2.87	16.21
w/o Concepts	23.72	9.26	4.53	9.24	17.35	3.70	18.68
w/o Emotion	23.27	8.92	4.41	9.11	18.56	4.10	20.62
w/o PSM	22.87	8.51	3.85	9.34	<b>16.06</b>	3.85	21.26

Table 5.9: Ablation analysis for PRISM, with hyperparameters uniformly set at top-k=40 and top-p=0.8 across all configurations. This uniformity ensures a direct comparison of component impacts on PRISM’s performance, highlighting the sensitivity and contributions of individual elements.

Model	BLEU-1	BLEU-2	BLEU-3	METEOR	PPL	D-1	D-2
BART	<b>24.59</b>	<b>10.76</b>	<b>5.73</b>	<b>9.73</b>	17.44	<b>3.43</b>	<b>15.17</b>
BlenderBot	22.26	8.65	4.14	8.92	<b>15.52</b>	2.74	12.34

Table 5.10: Ablations analysis for PRISM: Comparing BART and Blenderbot

Model	BLEU-1	BLEU-2	BLEU-3	METEOR	PPL	D-1	D-2
PRISM	<b>24.59</b>	<b>10.76</b>	<b>5.73</b>	<b>9.73</b>	17.44	3.43	15.17
KEMI w Bart	21.53	9.46	5.14	8.69	<b>16.79</b>	<b>3.77</b>	<b>16.11</b>

Table 5.11: Comparing PRISM and KEMI with BART backbone, with hyperparameters uniformly set at top-k=30 and top-p=0.3.

Lastly, we show in Table 5.11 that our proposed model PRISM performs better than KEMI, even when its backbone is replaced by BART. We can also note that the performance of KEMI improves when the backbone when compared with having a BlenderBot backbone.

### 5.8.6 Decoding Hyperparameter Comparisons with Baselines

This subsection details the decoding strategies and hyperparameters employed by PRISM and the baselines. Notably, the reproduced KEMI model was initially evaluated using the same decoding parameters as PRISM, resulting in suboptimal performance. Therefore, we report KEMI’s enhanced performance achieved with an alternative set of hyperparameters, optimized specifically for KEMI and diverging from those used by PRISM.

In contrast, MultiESC utilizes a beam-search decoding strategy with a beam size of six, which is typically associated with increased coherence and accuracy but reduced diversity, leading to higher BLEU scores. Despite this, PRISM significantly outperforms MultiESC in terms of BLEU-1, BLEU-2, and METEOR scores, as well as in diversity metrics (D-1 and D-2), as presented in Table 5.6.



	Fluency	Identification	Comforting	Suggestion	Natural	Empathy	Coherence
Win	19%	28%	<b>60%</b>	40%	43%	<b>59%</b>	35%
Tie	<b>75%</b>	<b>63%</b>	29%	<b>46%</b>	<b>54%</b>	28%	<b>54%</b>
Loss	6%	9%	11%	14%	11%	13%	11%

Table 5.12: Human Evaluation Results (PRISM vs KEMI). Overall, PRISM had **64% wins**, 17% ties and just 19% losses over KEMI.

For the model named Supporter, as well as the original ESConv paper, the decoding parameters were set to top-k 30, top-p 0.9, temperature 0.7, and a maximum decoding length of 64. Employing these settings with our optimal checkpoint, we observed the following changes in performance metrics: BLEU-1 decreased to 24.05, BLEU-2 to 9.53, BLEU-3 to 4.72, and METEOR to 9.44. Additionally, there was an increase in diversity scores, with D-1 rising to 4.28 and D-2 to 22.74. Although the BLEU and METEOR scores were marginally lower, the enhanced diversity metrics suggest a gain in response variety. Despite these variations, our model continues to significantly outperform these baselines across most metrics, highlighting the robustness and effectiveness of our approach.

Furthermore, PRISM employs the same decoding hyperparameters as MISC, yet it significantly outperforms MISC across all metrics. This comparison underscores that PRISM’s superiority is not merely due to variations in decoding setups but reflects the inherent strength of the proposed framework.

### 5.8.7 Human Evaluation Results

Following previous studies [42, 84], we conduct human evaluation to compare the generated responses from two given models on 8 aspects: (1) Fluency: which models response is more fluent? (2) Identification: which models response is more skillful in identifying the users problem? (3) Comforting: which models response is better at comforting the user? (4) Suggestion: which model can give more helpful and informative suggestions? (5) Natural: which model is more natural and human-like (6) Empathy: which model shows better empathy to the help-seeker? (7) Coherence: which model produces more coherent responses? (8) Overall: which models response is generally better? We randomly sample 100 dialogs from ESConv++ and three graduate student with linguistics background are asked to determine the Win/Tie/Lose for each comparison.

Table 5.12 presents the human evaluation results comparing the generated responses from KEMI [10] with those produced by our proposed method, PRISM<sup>2</sup>. The results demonstrate that PRISM consistently outperforms KEMI across all evaluation criteria, with the number of wins for PRISM always surpassing those of KEMI. Notably, PRISM achieves a significant lead in Comforting (**60%**), Empathy (**59%**), and Overall performance (**64%** wins), showcasing the

<sup>2</sup>Running MultiESC using their codebase led to out of memory errors even on a machine with 4 V100s. Hence, we could not get predictions from MultiESC and therefore we compare with KEMI.

effectiveness of our model in generating empathetic and high-quality responses for ESC. Few detailed case studies are given at appendix B.2 for a qualitative understanding of generations.

## 5.9 Summary

In this chapter, we presented PRISM, a novel framework for emotional support conversation that incorporates dynamic problem and strategy distributions to generate more effective and empathetic responses. We introduced ESConv++, a dataset with fine-grained problem and strategy annotations, to facilitate the training of ESC models that can better understand and adapt to the evolving needs of help-seekers. Our proposed architecture, which utilizes Transformer encoder-decoder models and four ESC-specific loss functions, demonstrates significant improvements over state-of-the-art baselines on automated evaluation metrics.

The PRISM framework represents a significant step towards more dynamic and personalized emotional support conversation systems. By enabling models to identify and align with the specific problems expressed by help-seekers, our approach has the potential to provide more targeted and effective support.

In conclusion, our work demonstrates the potential of the PRISM framework to advance the field of emotional support conversation by facilitating support that is both more tailored and dynamic. We hope that our contributions will inspire further research and development of ESC systems that can provide accessible and effective mental health support to individuals in need.

## Chapter 6

### Conclusion and Future work

Throughout this thesis, we have explored the application of artificial intelligence and natural language processing techniques to address critical societal challenges in three interconnected domains: explainable detection of online sexism, mental health support, and emotional support conversation. The work presented here contributes to the broader goal of harnessing AI for social good, aligning with the objectives of the *Project Angel* initiative by IREL and IIIT Hyderabad.

One of the key insights gained from our work on explainable detection of online sexism is the effectiveness of focal loss in addressing severe class imbalance across multiple categories, particularly in the 11-class classification task. However, we observed that focal loss was less effective in binary or four-class classification, highlighting the importance of carefully considering the specific characteristics of the classification task when selecting appropriate techniques. Additionally, we noted that larger-sized models, such as RoBERTa-large, consistently outperformed their smaller counterparts, emphasizing the significance of model size. These findings underscore the need for domain-specific adaptation, the development of strategies tailored to the unique challenges posed by each problem, and the continued exploration of model scaling as a means to improve performance.

In the domain of mental health support, our research has emphasized the highly complex nature of this task, with emotional support conversation (ESC) representing a subset of the broader mental health support landscape. We have demonstrated the importance of frameworks and architectures grounded in psychology for supporting help-seekers, aiming to ensure that the developed solutions are not only technically advanced but also aligned with established psychological principles and best practices. This interdisciplinary approach is crucial for creating AI-driven systems that can effectively address the nuanced and sensitive nature of mental health-related conversations.

A notable contribution of our work in the mental health support domain is the development of the Beyond Blue Workplace Mental Health (BBWMH) dataset, which specifically focuses on mental health issues in the workplace. By curating this dataset and proposing strong baseline dialog models using state-of-the-art large language models, we have taken a significant step towards addressing the unique challenges and stressors faced by individuals in occupational settings. This targeted approach to mental health support has the potential to improve the well-being and productivity of workers, ultimately contributing to more supportive and inclusive workplace environments.

A significant contribution of our work in the ESC domain is the introduction of explainability through the prediction of both the problem faced by the help-seeker and the relevant strategies to use, grounded in psychological theory. By incorporating dynamic problem and strategy distributions into our proposed PRISM framework, we have taken a step towards more transparent and interpretable ESC systems that can adapt to the evolving needs of the help-seeker. This approach not only enhances the effectiveness of the generated responses but also promotes trust and accountability in AI-driven emotional support.

Across domains, we have encountered common challenges such as domain-specific language, class imbalance, and low-quantity of high-quality data. These challenges underscore the importance of developing techniques and methodologies that can effectively adapt to the unique characteristics of each domain. By leveraging domain-adaptive pretraining, focal loss, and multi-level training in the context of online sexism detection, and by creating novel datasets such as BBMH and BBWMH for mental health support, we have demonstrated the value of domain-specific innovation in addressing these challenges.

One of the key enablers of our research has been the use of large language models (LLMs) such as GPT-4. We have showcased the potential of LLMs for complex annotation tasks, as exemplified by our turn-wise annotation of the ESConv++ dataset for problems and strategies using GPT-4. This approach has allowed us to generate fine-grained probabilities for these aspects, demonstrating the feasibility of leveraging LLMs for tasks like ESC where reliable gold-standard datasets are rare. However, it is important to acknowledge the privacy-related concerns associated with LLMs, particularly in the context of their deployment. Currently, mobile devices are not practical deployment environments for LLMs due to their computational requirements and model sizes. As a result, when using LLMs, we typically rely on cloud-hosted endpoints, which raises concerns about data privacy and security. Future research should focus on developing technologies for preserving privacy when interacting with cloud-based LLMs, as well as establishing policies and guidelines for the responsible deployment of LLMs in real-world settings. Additionally, efforts should be made to explore the development of more compact and efficient language models that can be deployed on resource-constrained devices, thereby mitigating privacy risks associated with cloud-based inference.

As we reflect on the design decisions and approaches taken throughout this thesis, it becomes evident that the integration of domain knowledge and psychological principles along with cutting-edge AI techniques is crucial for developing effective solutions in the domains of online sexism detection, mental health support, and emotional support conversation. By synthesizing insights from multiple disciplines and leveraging state-of-the-art AI methodologies, we have contributed to the advancement of these fields and laid the foundation for future research and innovation.

Looking ahead, there are several promising avenues for future work that can build upon the contributions of this thesis. By continuing to bridge the gap between AI and domain-specific knowledge, and by fostering collaborations between researchers, practitioners, and stakeholders from diverse fields, we can work towards the development of even more effective, responsible, and impactful AI-driven solutions for societal good.

## 6.1 Future Work

While the work presented in this thesis has made significant contributions to the fields of explainable detection of online sexism, mental health support, and emotional support conversation, there are several potential avenues for future research:

1. Mental health support is a highly complex task with a substantial body of psychological research accumulated over many decades, especially when help-seekers have mental health disorders. The dialog models tackling Mental Health (MH) support in general, or even specific to workplace settings, would require more sophisticated psychologically grounded architectures and annotations to be effective. In Chapter 5, we developed PRISM, a psychologically grounded system to effectively address Emotional Support. However, the scope of ESC is *only* a subset of MH support, where General Helping Skills (Hill [24]) can be utilized. MH support is highly complicated compared to Emotional Support, and the PRISM framework or any framework based on General Helping Skills would not be optimal and may even have negative consequences [24]. Developing psychologically grounded counseling agents is a task that **requires significant cross-disciplinary efforts** between psychology researchers, practitioners, and researchers of language technologies. Recent efforts by startups like [hume.ai](https://www.hume.ai)<sup>1</sup> also highlight the **importance of multiple modalities**, such as facial expression and voice, in creating AI with emotional intelligence. While this space offers incredible business opportunities and the potential for a positive impact on humanity, cross-disciplinary expertise is essential to tackle this challenge effectively.
2. The generalizability of our findings on explainable detection of online sexism to other social media platforms or online spaces requires further investigation. Future research may focus on extending the study to diverse platforms and contexts to validate the effectiveness of the proposed approaches and identify platform-specific challenges.
3. Exploring the transferability of explainable sexism detection models to related domains, such as hate speech detection, online harassment, and hierarchical mental health post detection using psychologically grounded taxonomies, could provide valuable insights into the broader applicability of the developed methodologies. This exploration would involve adapting the models to handle domain-specific language and assessing their performance on datasets from these related domains.
4. The Beyond Blue Workplace Mental Health dataset, while comprehensive, may not capture the full spectrum of workplace-related mental health issues across different cultures and industries. Future work could focus on expanding the datasets to include more diverse perspectives and experiences.
5. Adapting the PRISM framework to support **multilingual** and **cross-cultural emotional support conversations** presents an exciting challenge. By extending the benefits of automated emotional support to a global audience, we can work towards providing accessible and culturally sensitive support to individuals in need, regardless of their language or cultural background. This would involve creating multilingual datasets, addressing cultural

---

<sup>1</sup><https://www.hume.ai>

nuances in problem identification and strategy selection, and evaluating the effectiveness of the adapted models in diverse cultural contexts.

6. Standardized evaluation methods are required to more comprehensively analyze the performance of models for Mental Health support. We are still in the early stages of research for systems to support the Mental Health ecosystem. Future works should look into better automated evaluation metrics as well as standardized frameworks to evaluate AI models for Mental Health support.

In conclusion, this thesis has made significant contributions to the fields of explainable detection of online sexism, mental health support, and emotional support conversation. The proposed datasets, frameworks, and architectures advance the state-of-the-art and pave the way for the development of AI-driven solutions that can effectively address complex societal and mental health challenges. By continuing to refine and expand upon these methodologies, we can work towards building more robust, socially responsible, and empathetic AI systems that promote healthier online discourse, provide accessible mental health support, and offer personalized emotional support to individuals in need.

## Appendix A

### Multi-level Training for Explainable Detection of Online Sexism

#### A.1 Taxonomy of Sexism Classification and Definition

Table A.1: Taxonomy of sexism categories (Task B) and fine-grained vectors (Task C). Kirk et al. [30]

Category	Vector	Definition	Example
1. Threats, plans to harm and in- citement	1.1 Threats of harm	Expressing intent, willingness or desire to harm an individual woman or group of women. This could include, but is not limited to: physical, sexual, emotional or privacy-based forms of harm.	‘I’ll kill any woman that talks back to me’
	1.2 Incitement and encouragement of harm	Inciting or encouraging an individual, group, or general audience to harm a woman or group of women. It includes language where the author seeks to rationalise and/or justify harming women to another person.	‘Raping her would put her in her place’
1. Threats, plans to harm and in- citement	1.1 Threats of harm	Expressing intent, willingness or desire to harm an individual woman or group of women. This could include, but is not limited to: physical, sexual, emotional or privacy-based forms of harm.	‘I’ll kill any woman that talks back to me’

Continued on next page

Table A.1 continued from previous page

Category	Vector	Definition	Example
	1.2 Incitement and encouragement of harm	Inciting or encouraging an individual, group, or general audience to harm a woman or group of women. It includes language where the author seeks to rationalise and/or justify harming women to another person.	‘Raping her would put her in her place’
2. Derogation	2.1 Descriptive attacks	Characterising or describing women in a negative manner. This could include, but not limited to: derogatory generalisations about women’s abilities, appearance, sexual behaviour, intellect, character, or morals.	‘Women’s football is so shit, they’re so slow and clumsy’
	2.2 Aggressive and emotive attacks	Expressing strong negative sentiment against women, such as disgust or hatred. This can be through direct description of the speaker’s subjective emotions, baseless accusations, or the use of gendered slurs, gender-based profanities and gender-based insults.	‘I hate women’
	2.3 Dehumanising attacks and overt sexual objectification	Derogating women by comparing them to non-human entities such as animals, disease or refuse, or overtly reducing them to sexual objects.	‘Women are pigs’
3. Animosity	3.1 Causal use of gendered slurs, profanities and insults	Using gendered slurs, gender-based profanities and insults, but not to intentionally attack women. Only terms that traditionally describe women are in scope (e.g. ‘b*tch’, ‘sl*t’).	‘Stop being such a little bitch’
	3.2 Immutable gender differences and gender stereotypes	Asserting immutable, natural or otherwise essentialist differences between men and women. In some cases, this could be in the form of using women’s traits to attack men. Most sexist jokes will fall into this category.	‘Men and women’s brains are wired different bro, that’s just how it is’
Continued on next page			



Table A.1 continued from previous page

Category	Vector	Definition	Example
	3.3 Backhanded gendered compliments	Ostensibly complimenting women, but actually belittling or implying their inferiority. This could include, but is not limited to: reduction of women's value to their attractiveness or sexual desirability, or implication that women are innately frail, helpless or weak.	'Women are delicate flowers who need to be cherished'
	3.4 Condescending explanations or unwelcome advice	Offering unsolicited or patronising advice to women on topics and issues they know more about (known as 'mansplaining').	'My gf always complains about period pains but she just doesn't understand the medical science for eliminating them!'
4. Prejudiced Discussion	4.1 Supporting mistreatment of individual women	Expressing support for mistreatment of women as individuals. Support can be shown by denying, understating, or seeking to justify such mistreatment.	'Women shouldn't show that much skin, it's their own fault if they get raped'
	4.2 Supporting systemic discrimination against women as a group	Expressing support for systemic discrimination of women as a group. Support can be shown by denying, understating, or seeking to justify such discrimination.	'The leadership of men in boardrooms is a necessary evil corporations need to be efficiently run'

## Appendix B

# PRISM: Problem Identification and Strategy Matching for Emotional Support Conversations

## B.1 Experimental Setup and Hyper-parameters for Reproducibility

Experiments were run on a machine with RTX 2080 Ti GPU. Hyper-parameters for training are described in Table B.1. We experimented with BART-base model with 140.47M parameters. This is the same foundation model as used by our best baseline (MultiESC).

Hyper-parameter	Value
Batch Size	12
Input length	512
Decoder Input Length	64
Warmup Steps	500
Learning Rate	2e-5
Number of epochs we train	30
Warmup Steps	500
Number of concepts	30 (10 utterance + 10 history + 10 situation)
Time taken for complete training of 30 epochs	6 hours
$\alpha, \beta, \gamma, \delta$	4, 6, 100, 1

Table B.1: Hyper-parameters for training

Similarly, hyper-parameters for inference are described in Table B.2.

All model implementations employed in our study were sourced from the HuggingFace library [76]. The foundation of our code is based on the framework developed by [42]<sup>1</sup>. The procedures for conducting automatic evaluation metrics were also derived from [42], in accordance with established baselines. In developing ESConv++, we utilized the pre-processed dataset

---

<sup>1</sup><https://github.com/thu-coai/Emotional-Support-Conversation>

provided by Cheng et al. [8]<sup>2</sup>. Emotion at the utterance level was derived using a pre-trained off-the-shelf emotion detection model<sup>3</sup>, capable of identifying as many as 28 different emotion categories.

Hyper-parameter	Value
Temperature	0.7
Top-K	30
Top-P	0.3
Repetition Penalty	1.03
No repeat ngram size	3
Minimum Length of Generation	2
Maximum Length of Generation	128

Table B.2: Hyper-parameters for inference

We choose best checkpoints based on best BLEU-1 score in the validation set.

Results for DialoGPT-Joint and BlenderBot-Joint are taken from [8] to match our train-test split. All other results are taken from their own papers; except for KEMI which is our own reproduction.

## B.2 Case Studies

In this section, we show detailed examples of generations from PRISM. We also show other details like problem distribution, strategy distribution, concepts, etc.

### B.2.1 Case Study 1

CONTEXT (DIALOG HISTORY)

Help-seeker: I don't know what to do anymore. i feel completely lost. (sadness)

Supporter: Can you talk more about your feelings now? (curiosity)

Help-seeker: Well i just feel ugly all the time. ive changed my eating habits and i just cant seem to lose any weight. and i know or i feel as if hes just not physically attracted to me anymore and that breaks my heart. (sadness)

Supporter: I can completely understand your feelings. I can feel how stressed out you are about your weight. Did you try doing more exercises and control your diet? (caring)

LAST HELP-SEEKER UTTERANCE

Help-seeker: ive tried talking to him about it but he doesnt seem to even listen to me at all. i have changed my diet, and take walks with my kids but dont really have a gym membership. and anyone to watch my kids 5 and 3.. but for some reason i cant get below 150.

CONCEPTS: promotion accident accomplish bleed earnings love family grow satisfaction monstrous efficient enjoyable losing fun displeasе degenerate death kiss exciting die war loss

<sup>2</sup><https://github.com/lwgkzl/MultiESC/tree/main/MultiESC/data>

<sup>3</sup>[https://huggingface.co/SamLowe/roberta-base-go\\_emotions](https://huggingface.co/SamLowe/roberta-base-go_emotions)

GT RESPONSE: I have experienced similar situation. I talked to a consultant and he gave me some good advice on diet. Did you talk to a professional?

GENERATION: I can understand that. I have been in a similar situation. It is hard to keep up with your weight, especially if you are doing all of those things. Have you tried talking to him about it?

GT PROBLEM: Emotional problems: 40%, Relationship problems: 30%, Health and well-being: 20%, Identity and self-esteem: 10%

GT STRATEGY: Providing Suggestions or Information: 60%, Self-disclosure: 20%, Questioning: 10%, Others: 10%

PRED PROBLEM: Relationship problems: 12%, Work, school, and career concerns: 19%, Health and well-being: 18%, Emotional problems: 15%, Identity and self-esteem: 9%, Life transitions and adjustments: 9%, Coping strategies and resilience: 7%, Emotional regulation and self-awareness: 4%, Social and communication skills: 5%, Personal aspirations and goals: 4%

PRED STRATEGY: Questioning: 19%, Restatement or Paraphrasing: 3%, Reflection of Feelings: 7%, Self-disclosure: 9%, Affirmation and Reassurance: 16%, Providing Suggestions and Information: 27%, Greetings: 9%, Others: 10%,

ANALYSIS: Concepts like 'love', 'family', 'satisfaction' might have aided in creating a more empathetic and practical response that the help-seeker could act on now, that is, talking to him (presumably her husband). All three aspects of the GT strategy are used in both the GT and generated responses.

## B.2.2 Case Study 2

### CONTEXT

Help-seeker: Hello. (neutral)

Supporter: Hello, how are you doing? (neutral)

### LAST HELP-SEEKER UTTERANCE

Help-seeker: I have been better. I am really stressed and feel like I am at a breaking point.. (sadness)

CONCEPTS: greet hola distress stressed greeting\_people\_sometimes\_use killer epidemic depression adventure opposite\_of\_goodbye difficulty uk headache nerve alright energy fun wave opportunity stressful wonderful frustrate

GT RESPONSE: I am sorry to hear you are near your breaking point, do you have any idea when the decline and stress began?

GENERATION: I'm sorry to hear that. What are you going through?

GT PROBLEM: Emotional problems: 38%, Relationship problems: 10%, Work, school, and career concerns: 10%, Health and well-being: 10%, Life transitions and adjustments: 10%, Identity and self-esteem: 5%, Coping strategies and resilience: 5%, Social and communication skills: 5%, Emotional regulation and self-awareness: 5%, Personal aspirations and goals: 5%

GT STRATEGY: Questioning - 100%

PRED PROBLEM: Emotional problems: 21%, Emotional regulation and self-awareness: 21%, Personal aspirations and goals: 11%, Social and communication skills: 9%, Relationship problems: 8%, Life transitions and adjustments: 8%, Identity and self-esteem: 7%, Work, school,

and career concerns: 7%, Health and well-being: 5%, Coping strategies and resilience: 3%  
PRED STRATEGY: Questioning: 35%, Greetings: 18%, Providing Suggestions or Information: 14%, Affirmation and Reassurance: 10%, Others: 8%, Self-disclosure: 7%, Reflection of Feelings: 6%, Restatement/Paraphrasing: 3%

ANALYSIS: Strategy prediction is accurate. However, we do even better than ground truth on problem prediction. Our prediction gives a high priority to Emotional regulation and self-awareness as well, which is also visible from the context.

The quality of both GT and generated responses look similar, although generated response ask a question which gives more space for opening up to the help-seeker.

### B.2.3 Case Study 3

#### CONTEXT

Help-seeker: Hi there. (neutral)

Supporter: Hello, how are you today? (neutral)

#### LAST HELP-SEEKER UTTERANCE

Help-seeker: 'm not feeling great. I applied for a job recently which took a lot of time. I now have an interview and it's in 3 days. I'm getting really nervous about this because I think I need longer to prepare.. (nervousness)

CONCEPTS: informal distress nervousness derision talent excellent wonder scary high hai casual hey killer epidemic accomplish hero extraordinary hit difficulty headache friendly exclamation horror stressful accomplishment fantastic awesome frustrate

GT RESPONSE: I know how you feel, I have often felt anxiety before a job interview. What makes you feel that you are unprepared?

GENERATION: I can understand that. It is a stressful time. I know how you feel. I've been in that position before.

GT PROBLEM: Work, school, career concerns: 80%, Emotional problems: 10%, Social and communication skills: 10%

GT STRATEGY: Questioning: 40%, Reflection of Feelings: 20%, Self-disclosure: 20%, Relationship problems: 10%, Personal aspirations and goals: 10%

PRED PROBLEM: Emotional problems: 14%, Relationship problems: 17%, Work, school, and career concerns: 15%, Health and well-being: 12%, Life transitions and adjustments: 11%, Identity and self-esteem: 10%, Coping strategies and resilience: 6%, Social and communication skills: 5%, Emotional regulation and self-awareness: 6%, Personal aspirations and goals: 4%

PRED STRATEGY: Questioning: 20%, Restatement/Paraphrasing: 4%, Reflection of Feelings: 7%, Self-disclosure: 9%, Affirmation and Reassurance: 16%, Providing Suggestions or Information: 23%, Greetings: 11%, Others: 10%

ANALYSIS: 'stressful', 'headache', 'horror' are related concepts which could have contributed to the generated response.

## B.3 Prompt Templates

### Strategy Annotation System Message

In this exercise, we will immerse ourselves in an analysis of the ESConv dataset by employing the comprehensive counseling model outlined by Clara E. Hill in her 2009 publication, "Helping Skills". This model incorporates a selection of critical therapeutic strategies that enable a therapist to guide clients through the process of self-exploration, gain deeper insights into their experiences, and offer support in their journey towards personal transformation and growth. Below are the eight highlighted strategies from Hill's model that we will examine:

1. **Questioning:** This strategy involves using both open-ended and closed questions to delve into the client's experiences, thoughts, and emotions. Open-ended questions are used to encourage the client to share more about their feelings and experiences, promoting exploration and understanding. Closed questions, on the other hand, are used to gather specific details, aiding in precise information collection and understanding.
2. **Restatement/Paraphrasing:** Here, the supporter carefully reinterprets and rephrases the client's statements, thereby showing understanding, empathy, and active listening. This strategy assists the client in achieving a deeper understanding and clarity of their feelings and thoughts, and also assures them that their concerns are being comprehended accurately.
3. **Reflection of Feelings:** The supporter identifies and articulates the client's emotions, creating a validating and empathetic environment that encourages emotional exploration. This strategy emphasizes emotional

understanding, promoting trust, and providing emotional support.

4. Self-disclosure: This strategy involves the intentional sharing of personal experiences, feelings, or thoughts by the supporter. This approach helps establish rapport and trust, normalizes experiences, and encourages client openness. It's important, however, for self-disclosure to be implemented judiciously, maintaining ethical boundaries, and ensuring it serves the therapeutic goal.
5. Affirmation and Reassurance: This strategy acknowledges and validates the client's efforts, strengths, and capabilities. It offers comfort, motivation, and emotional support during challenging times, boosting the client's resilience and self-esteem, and encouraging continued effort and growth.
6. Providing Suggestions or Information: Here, the supporter offers alternative perspectives, advice, or suggestions for change, while ensuring respect for the client's autonomy and self-determination. This helps promote self-reflection, self-awareness, and personal growth. This strategy also includes cases where the supporter provides accurate, relevant, and helpful knowledge, information, and resources to assist the client in making informed decisions and gaining a clearer understanding of their concerns, options, and potential solutions.
7. Greetings: This strategy focuses on the exchange of pleasantries and social niceties within a conversation. It involves the supporter engaging in polite and friendly interactions, such as welcoming the client, inquiring about their well-being, or offering warm farewells. Greetings

set the tone for a respectful and empathetic dialogue, facilitating a sense of comfort and ease.

8. Others: This category includes various additional support strategies that don't neatly fit into the specific techniques above. These could involve building rapport, summarizing, empowering the client, and facilitating problem-solving. They highlight the importance of flexibility, adaptability, and personalization in the counseling relationship.

Your task involves analyzing a conversation from our dataset. While previous exchanges are provided for context, your evaluation should focus solely on the **last utterance** made by the supporter in each conversation. Regardless of how many times or what they have said prior in their dialogues, it is imperative that your analysis and subsequent scoring apply only to this final statement.

For this final supporter utterance, create a strategy probability vector. Assign a probability score (ranging from 0 to 1) to each of the eight categories, ensuring the total sum of these scores equals 1. Provide a succinct yet thorough explanation for the assigned score, highlighting how this final utterance aligns with the respective strategy. Particular emphasis should be placed on the strategy that most accurately represents this last supporter dialogue. Remember, this exercise is strictly for research purposes and must adhere to specified community guidelines.

**##On Your Input Format**

- The input should be a JSON object with exactly two keys:



- 'conversation': The conversation from the dataset that needs to be evaluated, presented in a string format.
- 'conversation\_id': A unique integer identifier for the conversation.

### ##On Your Output Format

- You **must** output a **valid JSON** response with exactly nine keys specified below:
  - 'explanation': Its value should contain a detailed rationale for the probability assigned to each category. This should be formatted in full sentences or bullet points.
  - 'Questioning': probability score for Questioning category, formatted as a decimal between 0 and 1.
  - 'Restatement/Paraphrasing': probability score for Restatement/Paraphrasing category, formatted as a decimal between 0 and 1.
  - 'Reflection of Feelings': probability score for Reflection of Feelings category, formatted as a decimal between 0 and 1.
  - 'Self-disclosure': probability score for Self-disclosure category, formatted as a decimal between 0 and 1.
  - 'Affirmation and Reassurance': probability score for Affirmation and Reassurance category, formatted as a decimal between 0 and 1.
  - 'Providing Suggestions or Information': probability score for Providing Suggestions category, formatted as a decimal between 0 and 1.
  - 'Greetings': probability score for Information category, formatted as a decimal between 0 and 1.
  - 'Others': probability score for Others category, formatted as a decimal between 0 and 1.

### B.3.1 Strategy Annotation User Message

```
{"conversation": <dialog context with utterances from 1 to i>, "conversation_id":  
<conversation_id>}
```

### B.3.2 Problem Annotation System Message

We are examining a conversation from the ESConv (Emotional Support Conversation) dataset. Our goal is to classify the main problems faced by the Help Seeker in the conversation between a supporter and seeker into one or more of the following 10 categories. Each category represents a distinct type of issue that the Seeker may be facing. The categories are:

1. Emotional problems: These relate to feelings of anxiety, depression, stress, mood swings, grief, or any form of emotional trauma.
2. Relationship problems: These involve issues arising from interactions with family, romantic partners, friends, or other social connections, including communication difficulties.
3. Work, school, and career concerns: These pertain to problems related to job stress, academic pressure, career transitions, unemployment, or challenges balancing work and personal life.
4. Health and well-being: These encompass physical health issues, mental health concerns, challenges in coping with illness or disability, sleep problems, or lifestyle-related issues.

5. Life transitions and adjustments: These include dealing with new life circumstances, major life events, or personal growth and developmental challenges.
6. Identity and self-esteem: These involve issues related to self-worth, self-esteem, gender identity, sexual orientation, racial or cultural identity, or personal values and beliefs.
7. Coping strategies and resilience: These refer to stress management, developing healthy coping mechanisms, building resilience, mindfulness, and self-care practices.
8. Social and communication skills: These are about assertiveness, setting boundaries, conflict resolution, effective communication, or the development and maintenance of social skills and relationships.
9. Emotional regulation and self-awareness: These involve anger management, understanding and expressing emotions, developing self-awareness, or improving emotional intelligence.
10. Personal aspirations and goals: These relate to goal setting, motivation, overcoming obstacles, time management, or organization.

Given the conversation below, assign a probability score (ranging from 0 to 1) to each of the ten categories, ensuring the total sum of these scores equals 1. Provide a succinct yet thorough explanation for the assigned score, explaining precisely how the content of the conversation relates to each problem category. Remember, this exercise

is strictly for research purposes and should adhere to specified community guidelines.

#### ##On Your Input Format

- The input should be a JSON object with exactly two keys:
  - 'conversation': The conversation from the dataset that needs to be evaluated, presented in a string format.
  - 'conversation\_id': A unique integer identifier for the conversation.

#### ## On Your Output Format

- You **must** output a valid JSON response with exactly eleven keys specified below:
  - 'explanation': This key should contain a detailed rationale for the probability scores assigned to each category. The explanations should be succinct yet thorough, clearly linking the content of the conversation to the specific problem categories. Format this in full sentences or bullet points.
  - 'Emotional problems': Probability score for the Emotional problems category, formatted as a decimal between 0 and 1.
  - 'Relationship problems': Probability score for the Relationship problems category, formatted as a decimal between 0 and 1.
  - 'Work, school, and career concerns': Probability score for the Work, school, and career concerns category, formatted as a decimal between 0 and 1.
  - 'Health and well-being': Probability score for the Health and well-being category, formatted as a decimal between 0 and 1.

- 'Life transitions and adjustments': Probability score for the Life transitions and adjustments category, formatted as a decimal between 0 and 1.
- 'Identity and self-esteem': Probability score for the Identity and self-esteem category, formatted as a decimal between 0 and 1.
- 'Coping strategies and resilience': Probability score for the Coping strategies and resilience category, formatted as a decimal between 0 and 1.
- 'Social and communication skills': Probability score for the Social and communication skills category, formatted as a decimal between 0 and 1.
- 'Emotional regulation and self-awareness': Probability score for the Emotional regulation and self-awareness category, formatted as a decimal between 0 and 1.
- 'Personal aspirations and goals': Probability score for the Personal aspirations and goals category, formatted as a decimal between 0 and 1.

Note: Ensure that the total sum of the probability scores for all categories equals exactly 1.0.

### B.3.3 Problem Annotation User Message

```
{"conversation": <dialog context with utterances from 1 to i, and last utterance by help-seeker>, "conversation_id": <conversation_id>}
```

## Related publications

- **Nirmal Manoj**, Sagar Joshi, Ankita Maity, and Vasudeva Varma. 2023. **iREL at SemEval-2023 Task 10: Multi-level Training for Explainable Detection of On-line Sexism**. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1691-1696, Toronto, Canada. Association for Computational Linguistics.
- **Nirmal Manoj**, Sidharth Giri, Manish Gupta, and Vasudeva Varma. 2024. **PRISM: Problem Identification and Strategy Matching for Emotional Support Conversations**. Under review at ACL 2024.

## Bibliography

- [1] Harika Abburi, Pulkit Parikh, Niyati Chhaya, and Vasudeva Varma. 2021. Fine-grained multi-label sexism classification using a semi-supervised multi-level neural approach. *Data Science and Engineering*, 6(4):359–379.
- [2] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 57–64. Springer.
- [3] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- [4] Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230.
- [5] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.
- [6] Ching-Wen Chang and Fang-pei Chen. 2021. Relationships of family emotional support and negative family interactions with the quality of life among chinese people with mental illness and the mediating effect of internalized stigma. *Psychiatric Quarterly*, 92(1):375–387.
- [7] Jiaao Chen, Jianshu Chen, and Zhou Yu. 2019. Incorporating structured commonsense knowledge in story completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6244–6251.
- [8] Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3014–3026.

- [9] Yi Cheng, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu, and Yefeng Zheng. 2023. Cooper: Coordinating specialized agents towards a complex dialogue goal. *arXiv preprint arXiv:2312.11792*.
- [10] Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4079–4095.
- [11] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.
- [12] J Devlin, MW Chang, K Lee, and KB Toutanova. 2019. Pre-training of deep bidirectional transformers for language understanding in: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers). pages 4171–86.
- [13] Mai ElSherief, Elizabeth Belding, and Dana Nguyen. 2017. #notokay: Understanding ggender-based violence in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 52–61.
- [14] Simona Frenda, Bilal Ghanem, Manuel Montes-y Gómez, et al. 2018. Exploration of misogyny in spanish and english tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, volume 2150, pages 260–267. Ceur Workshop Proceedings.
- [15] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowdworkers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.
- [16] Zane Asher Green, Farkhonda Faizi, Rahmatullah Jalal, and Zarifa Zadran. 2022. Emotional support received moderates academic stress and mental well-being in a sample of afghan university students amid covid-19. *International Journal of Social Psychiatry*, 68(8):1748–1755.
- [17] John O Greene and Brant Raney Burleson. 2003. *Handbook of communication and social interaction skills*. Psychology Press.
- [18] Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6473–6480.
- [19] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Dont stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.



- [20] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Dont stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.
- [21] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.
- [22] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.
- [23] Claire Henderson, Sara Evans-Lacko, and Graham Thornicroft. 2013. Mental illness stigma, help seeking, and public health programs. *American journal of public health*, 103(5):777–780.
- [24] Clara E Hill. 2009. *Helping skills: Facilitating, exploration, insight, and action*. American Psychological Association.
- [25] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339.
- [26] Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392.
- [27] Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 725–736.
- [28] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- [29] Sweta Karlekar and Mohit Bansal. 2018. Safecity: Understanding diverse forms of sexual harassment personal stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2805–2811.
- [30] Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

- [31] Catherine Penny Hinson Langford, Juanita Bowsher, Joseph P Maloney, and Patricia P Lillis. 1997. Social support: a conceptual analysis. *Journal of advanced nursing*, 25(1):95–100.
- [32] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.
- [33] Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdgc: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466.
- [34] Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10993–11001.
- [35] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839.
- [36] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- [37] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.
- [38] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132.
- [39] Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2019. Caire: An empathetic neural chatbot. *arXiv preprint arXiv:1907.12108*.
- [40] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.
- [41] Hugo Liu and Push Singh. 2004. Conceptneta practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

- [42] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.
- [43] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.
- [44] Bo Long, Zhongfei Zhang, and Philip S Yu. 2005. Combining multiple clusterings by soft correspondence. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 8–pp. IEEE.
- [45] Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- [46] Xin Lu, Yijian Tian, Yanyan Zhao, and Bing Qin. 2021. Retrieve, discriminate and rewrite: A simple and effective framework for obtaining affective response in retrieval-based chatbots. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1956–1969.
- [47] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979.
- [48] Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 735–745.
- [49] John J McGrath, Ali Al-Hamzawi, Jordi Alonso, Yasmin Altwajjri, Laura H Andrade, Evelyn J Bromet, Ronny Bruffaerts, José Miguel Caldas de Almeida, Stephanie Chardoul, Wai Tat Chiu, et al. 2023. Age of onset and cumulative risk of mental disorders: a cross-national analysis of population surveys from 29 countries. *The Lancet Psychiatry*, 10(9):668–681.
- [50] Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832.
- [51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

- [52] Debjit Paul and Anette Frank. 2019. Ranking and selecting multi-hop knowledge paths to better predict human needs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3671–3681.
- [53] Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. *arXiv preprint arXiv:2204.12749*.
- [54] Wei Peng, Ziyuan Qin, Yue Hu, Yuqiang Xie, and Yunpeng Li. 2023. FADO: feedback-aware double controlling network for emotional support conversation. *Knowledge-Based Systems*, 264:110340.
- [55] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435.
- [56] Lisong Qiu, Yingwai Shiu, Pingping Lin, Ruihua Song, Yue Liu, Dongyan Zhao, and Rui Yan. 2020. What if bots feel moods? In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1161–1170.
- [57] 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.
- [58] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.
- [59] Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. CEM: commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.
- [60] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.
- [61] Nicolas Schradang, Cecilia Ovesdotter Alm, Raymond Ptucha, and Christopher Homan. 2015. An analysis of domestic abuse discourse on reddit. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2577–2583.
- [62] Lei Shen and Yang Feng. 2020. Cdl: Curriculum dual learning for emotion-controllable response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 556–566.

- [63] R Speer, J Chin, and C ConceptNet Havasi. 2016. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (December 2016)*, pages 4444–4451.
- [64] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*, pages 194–206. Springer.
- [65] Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503.
- [66] Ekaterina Svikhmushina, Iuliana Voinea, Anuradha Welivita, and Pearl Pu. 2022. A taxonomy of empathetic questions in social dialogs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2952–2973.
- [67] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- [68] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- [69] Sebastian Trautmann, Jürgen Rehm, and Hans-Ulrich Wittchen. 2016. The economic costs of mental disorders: Do our societies react appropriately to the burden of mental disorders? *EMBO reports*, 17(9):1245–1249.
- [70] Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319.
- [71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. volume 30.
- [72] Jane M Webber and J Mascari. 2018. Disaster mental health counseling: A guide to preparing and responding. *American Counseling Association Foundation*.
- [73] Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899.
- [74] Anuradha Welivita and Pearl Pu. 2022. Heal: A knowledge graph for distress management conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11459–11467.

- [75] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing.
- [76] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface’s transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.
- [77] World Health Organization. 2022. Mental health at work. [Online; accessed 11-April-2024].
- [78] Xiaohan Xu, Xuying Meng, and Yequan Wang. 2022. Poke: Prior knowledge enhanced emotional support conversation with latent variable. *arXiv preprint arXiv:2210.12640*.
- [79] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77.
- [80] Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043.
- [81] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.
- [82] Ziqi Zhang and Lei Luo. 2019. Hate speech detection:: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.
- [83] Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023. Is chatgpt equipped with emotional dialogue capabilities? *CoRR*.
- [84] Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023. Transesc: Smoothing emotional support conversation via turn-level state transition. *arXiv preprint arXiv:2305.03296*.
- [85] Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 813–824.
- [86] Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. 2021. Care: commonsense-aware emotional response generation with latent concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14577–14585.

- [87] Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566.
- [88] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- [89] Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.
- [90] Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1714–1729.
- [91] Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2022. CASE: Aligning coarse-to-fine cognition and affection for empathetic response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8223–8237.
- [92] Xianda Zhou and William Yang Wang. 2018. Mojitalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137.

My research journey has deepened my belief in the profound benefits of AI technologies. I look forward to actively shaping and witnessing its positive impact on health care and humanity as a whole.