



**Towards Enhanced Conversational Dynamics for Effective Virtual
Therapist-Assistive Counseling**

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

ASEEM SRIVASTAVA

(PHD20010)

Computer Science and Engineering

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

September 10, 2025

DOCTORAL COMMITTEE

Supervisors

- Md. Shad Akhtar, Assistant Professor, Indraprastha Institute of Information Technology, Delhi
- Tanmoy Chakraborty, Associate Professor, Indian Institute of Technology, Delhi

Internal Committee

- Pushendra Singh, Professor, Indraprastha Institute of Information Technology, Delhi
- Raghava Mutharaju, Associate Professor, Indraprastha Institute of Information Technology, Delhi

PhD Comprehensive External Examiner

- Manish Shrivastava, International Institute of Information Technology, Hyderabad, India

PhD Thesis External Examiners

- Trevor Cohen, University of Washington, USA
- Lucie Flek, University of Bonn, Germany
- Radhika Mamidi, International Institute of Information Technology, Hyderabad, India

THESIS CERTIFICATE

This is to certify that the thesis titled **Towards Enhanced Conversational Dynamics for Effective Virtual Therapist-Assistive Counseling**, submitted by **Aseem Srivastava**, to the Indraprastha Institute of Information Technology Delhi, for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by him under my supervision. In my opinion, the thesis has reached the standard fulfilling the requirements of the regulations relating to the degree. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.



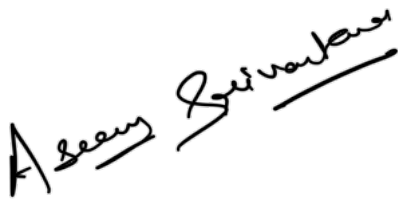
Dr. Md Shad Akhtar
Thesis Supervisor
Assistant Professor
Dept. of Computer Science and Engineering
IIT Delhi, 110020



Dr. Tanmoy Chakraborty
Thesis Supervisor
Associate Professor
Dept. of Electrical Engineering
IIT Delhi, 110016

CANDIDATE'S DECLARATION

The author hereby declares that the work presented in the thesis titled **Towards Enhanced Conversational Dynamics for Effective Virtual Therapist-Assistive Counseling**, submitted as partial fulfilment for the award of the degree of Doctor of Philosophy to the IIT-Delhi, is an original research work carried out under the joint supervision of Dr. Md. Shad Akhtar and Dr. Tanmoy Chakraborty. The results presented in this thesis have not been submitted in part or whole to any other university or institute for the award of any degree/diploma.

A handwritten signature in black ink, reading "Aseem Srivastava", with a horizontal line underneath the name.

Aseem Srivastava (PHD20010)
PhD Candidate
Department of Computer Science and Engineering
IIT Delhi, 110020

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisors, Dr. Md. Shad Akhtar and Dr. Tanmoy Chakraborty. Their constant encouragement, guidance, and support have been instrumental throughout the course of my PhD journey. Their mentorship has shaped not only the direction of my research but also my overall approach to scientific inquiry and academic rigor. I am also sincerely thankful to the members of my internal doctoral committee, Prof. Pushpendra Singh and Dr. Raghava Mutharaju, for their constructive feedback and insightful suggestions at critical stages of my work. I would also like to acknowledge Dr. Manish Shrivastava, who served as my comprehensive examination external examiner and provided valuable perspectives. My heartfelt appreciation goes to my thesis external examiners, Prof. Trevor Cohen, Prof. Lucie Flek, and Prof. Radhika Mamidi, whose thoughtful evaluations and feedback helped refine my research and improve this thesis. I also acknowledge the institutional support provided by the LCS2 Lab at IIT Delhi, the FLAME Lab at IIIT Delhi, and the Infosys Centre of Artificial Intelligence at IIIT Delhi. These research environments provided me with the resources, opportunities, and collaborations that were vital for my academic and professional growth.

During the course of my doctoral studies, I was fortunate to receive research grants and travel support from Google Research, Microsoft, and ACM India, for which I am sincerely grateful. These opportunities not only alleviated financial challenges but also enabled me to present my work at prominent conferences and engage with the global research community. I also appreciate the facilities and resources provided by my institute, which allowed me to concentrate fully on the pursuit of my research goals.

I would like to extend my gratitude to my labmates, startup collaborators, and independent researcher peers. Through numerous discussions, brainstorming sessions, and collaborative efforts, I not only gained a deeper understanding of fundamental research problems but also had the privilege of building interdisciplinary pillars of my thesis. These interactions enriched my academic journey immensely. Beyond the formal academic sphere, I am also grateful to a wider circle of peers-friends, colleagues, and communities-who engaged with me in stimulating discussions, offered encouragement during difficult times, and reminded me of the broader purpose of research and life outside academia.

On a personal note, I am thankful to my father, mother, and brother, as well as to my fiancée, extended family members, and friends, for their unconditional love and emotional support. Their belief in me has been the foundation upon which this work stands. They have been my constant source of strength, inspiration, and motivation during moments of both achievement and struggle. I would also like to acknowledge my two beloved pets, Champ and Jack, who provided me with constant companionship and emotional support, bringing joy and balance to my life throughout this demanding journey.

Finally, I wish to acknowledge all those whose names may not be explicitly mentioned here but who have, in their own ways, contributed to my growth, learning, and perseverance during this journey.

Apoorv Srivastava

PUBLICATIONS

Conferences

- C1. Ganeshan Malhotra, Abdul Waheed, **Aseem Srivastava**, Md Shad Akhtar, and Tanmoy Chakraborty. "Speaker and Time-aware Joint Contextual Learning for Dialogue-act Classification in Counselling Conversations". *ACM International Conference on Web Search and Data Mining (WSDM)*, 2022.
- C2. **Aseem Srivastava**, Tharun Suresh, Sarah P. Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. "Counseling Summarization Using Mental Health Knowledge Guided Utterance Filtering". *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022.
- C3. **Aseem Srivastava**, Ishan Pandey, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. "Response-act Guided Reinforced Dialogue Generation for Mental Health Counseling". *ACM Web Conference (WWW)*, 2023.
- C4. **Aseem Srivastava**, Smriti Joshi, Tanmoy Chakraborty, Md Shad Akhtar. "Knowledge Planning in Large Language Models for Domain-Aligned Dialogue Summarization". *Empirical Methods in Natural Language Processing (EMNLP)*, 2024

Journals

- J1. **Aseem Srivastava**, Tanya Gupta, Alison Cerezo, Grin Lord, Md Shad Akhtar, Tanmoy Chakraborty. "Critical Behavioral Traits Foster Peer Engagement in Online Mental Health Communities". *PLOS ONE* 2025
- J2. **Aseem Srivastava**, Gauri Naik, Alison Cerezo, Tanmoy Chakraborty, Md. Shad Akhtar. "Sentiment-guided Commonsense-aware Response Generation for Counseling". *IEEE Transactions on Artificial Intelligence (TAI)*, 2025 (Under Review)
- J3. **Aseem Srivastava**, Zuhair Hasan Shaik, Tanmoy Chakraborty, Md Shad Akhtar. "Trust Modeling in Counseling Conversations: A Benchmark Study". *Transactions of the Association for Computational Linguistics (TACL)*, 2025 (Under Review)

ABSTRACT

KEYWORDS: Dialogue System; Mental Health; Natural Language Processing

Mental health issues are now the leading global cause of disability, with conditions such as anxiety and depression escalating, particularly following the COVID-19 pandemic. However, traditional mental health support methods remain heavily constrained by a severe shortage of trained professionals, making it difficult to meet the increasing demand for mental health support systems. In response, we propose Virtual Mental Health Assistants (VMHAs) as a scalable and accessible alternative, offering instant, anonymous, and stigma-free method of support. Despite their potential, existing automated counseling systems are limited by rigid, scripted dialogues and fail to replicate the nuanced adaptability and therapeutic depth of human therapists. To address this, we focus on dialogue systems on modular levels, spanning understanding, summarization, generation, and evaluation of therapeutic conversations. This thesis studies the counseling interaction pipeline, refining its core components to enhance the efficiency and effectiveness of professionals. The goal is to improve dialogue understanding, enabling VMHAs to interpret users' implicit psychological intents through directive recognition. To maintain coherence and continuity across conversations, we incorporate domain knowledge infusion into counseling summarization, allowing the system to retain relevant memory. Additionally, we advance dialogue generation by integrating clinically informed, emotionally adaptive response models, surpassing traditional rule-based and purely generative approaches to ensure more human-like and therapeutic interactions. Next, we propose a dialogue evaluation framework that centers on therapeutic bond assessment via trust modeling. Recognizing that mental health support extends beyond one-on-one counseling, we further analyze peer interactions in online communities and group therapy, positioning AI as a facilitator of collective support environments. Through rigorous experimentation, user studies, and exhaustive analysis, this thesis establishes a carefully designed, context-aware, and psychologically informed understanding of counseling interactions. Rather than approaching VMHAs as standalone interventions, this research emphasizes their role as an augmentative approach, proposed to enhance the efficiency of mental health professionals. We conduct each phase of this study in close collaboration with domain experts, ensuring that the proposed methodologies are novel and practically viable in real-world settings. Moreover, the findings presented here extend beyond novel methodological contributions, positioning this thesis' findings not just as novel solutions, but as supportive alternatives that alleviate the burden on mental health professionals.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	i
PUBLICATIONS	ii
ABSTRACT	iii
LIST OF TABLES	xii
LIST OF FIGURES	1
1 Introduction	2
1.1 Thesis Overview and Statement	2
1.2 Preliminary	4
1.2.1 Conversations	4
1.2.2 Psychology & Counseling	5
1.3 Thesis Organization	6
1.4 Ethical Considerations	8
2 Background	10
2.1 Mental Health Challenges and Digital Interventions	10
2.2 Thematic Advances in Counseling Conversation Analysis	10
2.3 Dialogue Systems for Mental Health	11
2.4 Integration of Domain-Specific Knowledge and Commonsense	11
2.5 Summary and Gap Identification	12
I Conversational Understanding	13
3 Dialogue-act Classification	14
3.1 Introduction	14
3.2 Related Work	16

3.3 Dataset	17
3.4 Proposed Methodology	21
3.5 Experiments, Results, and Analysis	23
3.6 Limitations	27
3.7 Summary	27
II Counseling Summarization	28
4 Domain Knowledge for Counseling Summarization	29
4.1 Introduction	29
4.2 Related Work	31
4.2.1 Mental Health and NLP	31
4.2.2 Dialogue Summarization	32
4.3 Dataset	33
4.3.1 Annotation Process	33
4.3.2 Data Analysis	34
4.3.3 Ethical Considerations	35
4.4 Methodology	35
4.5 Experiments and Results	37
4.5.1 Baselines	37
4.5.2 Results and Ablation Study	37
4.5.3 Error Analysis	38
4.5.4 Clinical and Human Evaluation	39
4.6 Mental Health Information Capture (MHIC) Metric	39
4.7 Discussion	40
4.7.1 Justification for Selection of PHQ-9 Framework	40
4.7.2 Uniqueness of Mental-health counseling and ConSum	40
4.8 Summary	41
5 Plan Before Summarization	42
5.1 Introduction	42
5.2 Related Work	43

5.3 Dataset	44
5.4 Methodology	45
5.4.1 Domain Knowledge Encapsulation	45
5.4.2 Structural Knowledge Encapsulation	46
5.4.3 Planning Engine	47
5.5 Experiments and Results	47
5.5.1 Baselines	47
5.5.2 Performance Comparison	48
5.5.3 Ablation Study	48
5.5.4 Qualitative Analysis	48
5.5.5 Mental Health Information Capture	50
5.5.6 Mental Health Expert Validation	50
5.6 Limitations	52
5.7 Ethical Considerations and Future Work	52
5.8 Summary	52
III Response Generation	54
6 Dialogue-act Controlled Response Generation	55
6.1 Introduction	55
6.2 Related Work	57
6.3 Methodology	58
6.3.1 Reference Models for Reward Computation	60
6.3.2 Training and Proximal Policy Optimization	60
6.4 Experiments	61
6.4.1 Dataset	61
6.4.2 Baselines and Evaluation Metrics	62
6.5 Results and Analysis	63
6.5.1 Performance Comparison	63
6.5.2 Qualitative Analysis	64
6.5.3 Application of READER: Dialogue Generation	65

6.5.4	Generalizability	65
6.6	Discussion	66
6.7	Summary	66
7	Commonsense Guided Response Generation	68
7.1	Introduction	68
7.2	Related Work	70
7.3	Dataset	71
7.3.1	Dataset Extension	72
7.3.2	Discussion on COMET Relation Selection	72
7.4	Methodology	73
7.4.1	Speaker Context Graph (SC-Graph)	74
7.4.2	Commonsense Knowledge	74
7.4.3	Response Generation Block	75
7.5	Experiments and Results	76
7.5.1	Comparable Baselines	76
7.5.2	Performance Comparison	78
7.5.3	Ablation Study	78
7.5.4	Qualitative Analysis	78
7.5.5	Experimental Setup	80
7.6	User Study	80
7.6.1	Ethical Considerations for User Study	81
7.6.2	Societal Impact	81
7.7	Discussion	82
7.8	Ethics and Limitations	82
7.9	Summary	82
IV	Therapeutic Bond	84
8	Trust: A Therapeutic Bond Metric	85
8.1	Introduction	85
8.2	Related Work	86

8.3 Dataset	87
8.3.1 Data Annotation	88
8.3.2 Annotation Guidelines	88
8.3.3 Annotation Process	89
8.4 Dataset Analysis	90
8.5 TrustBench	91
8.5.1 Trust Modeling	91
8.5.2 Competing Models	91
8.6 Results and Analysis	93
8.6.1 Performance Comparison	93
8.6.2 Analysis	94
8.7 Discussion	95
8.7.1 Discussion on Effect of Model Sizes	95
8.7.2 Discussion on Trust Evolution Patterns	96
8.8 Limitations and Future Work	97
8.9 Ethical Considerations	97
8.10 Summary	97
V Peer Interactions	98
9 Understanding Peer Behavior	99
9.1 Introduction	99
9.2 Dataset and Methods	101
9.2.1 Data Collection	101
9.2.2 Data Annotation	102
9.3 Analysis and Findings	104
9.4 Experiments and Results	107
9.4.1 Metadata and Content Analysis	108
9.4.2 Topical Analysis	108
9.4.3 Additional Auxiliary Analysis	109
9.5 Discussion	112

9.6 Summary	113
10 Conclusion and Future Work	114

LIST OF TABLES

1.1 Abbreviations and their definitions used in the thesis.	9
3.1 Example of a sample conversation session between a patient and a therapist. Each utterance has an associated dialogue-act classification (DAC) label.	15
3.2 Dialogue-act distribution in HOPE. The train, test, and validation splits are 70:20:10	18
3.3 Table showcasing the performance of baseline models as compared with our SPARTA model. U^t represents the use of Utterance representation, GC represents Global Context, LC represents Local Context and SA means the presence of Speaker Aware representations. The dagger symbol (†) represents statistically significant results compared to the best baseline, CASA.	24
3.4 Label-wise classification report for SPARTA-TAA.	24
3.5 An example therapy session showcasing the differences in predictions between the baseline and SPARTA. SPARTA-T and SPARTA-M refer to the SPARTA-TAA and SPARTA-MHA models. We truncate the dialogue due to its length (total 151 utterances).	25
4.1 Results obtained on MEMO. We report Rouge-1 (R-1), Rouge-2 (R-2), Rouge-L (R-L), Bleurt Score (BS) and QuestEval Score (QAE).	38
4.2 Ablation study on the effect of different modules of ConSum. SH, PD, DF and RT are psychotherapy elements. MH-Know and CCC are knowledge infused module and counseling component classifier module respectively. We present our analysis on five scores to understand all aspects of summary generation.	38
4.4 Comparative analysis on generated summaries between T5 and ConSum model. Green denotes primary medical condition information, Cyan denotes SH information, Brown denotes RT information and Red denotes errors/irrelevance in generated summaries. Comparison shows ConSum captures better contextual information compared to T5.	39
4.3 Human evaluation on the summaries generated from ConSum model.	39
5.1 Results obtained on the MEMO counseling summarization dataset. We report Rouge-1 (R-1), Rouge-2 (R-2), Rouge-L (R-L), and Bleurt Score (BS).	48

5.3	A comparative analysis of three most relevant models – Llama, MentalLlama, and P <i>IECE</i> . The key emphasis on being analyzed here is understanding the core counseling components, including discussions on <i>symptoms and reasons</i> , <i>reflecting client’s story</i> , and <i>critical routine details</i> along with the <i>additional intricate domain understanding</i> . While Llama generates general-purpose summaries, MentalLlama, being pretrained on mental health data, captures nuanced knowledge beyond Llama’s scope. In contrast, P <i>IECE</i> is able to capture in-depth domain knowledge, surpassing other models in touching upon counseling nuances. Despite the grammatical proficiency of baselines, P <i>IECE</i> stands out for its structural understanding, focusing on intricate details, in some cases, with more detailed accounts of these aspects than are provided in the reference summary	49
5.2	Ablation study of the proposed model, P <i>IECE</i> and generalizability of the planning engine on top of notable LLMs as foundation model (α), including Mistral, Zephyr, and Llama, illustrating a clear <i>improvement</i> (\uparrow) in LLM generations by integrating planning engine.	49
5.4	Human evaluation on the summaries generated from P <i>IECE</i> model. The average inter-rater’s agreement score (κ) for P <i>IECE</i> is 0.82.	50
5.5	Expert evaluation across six domain-centric metrics: <i>affective attitude (Af. Att.)</i> , <i>burden, ethicality (Ethic)</i> , <i>intervention coherence (Intv. Coh.)</i> , <i>opportunity cost</i> , and <i>perceived effectiveness (Percv. Eff.)</i> . We present mean (μ) and standard deviation (Δ) scores. . .	51
6.2	Results obtained on the HOPE dataset. We show ROUGE (1, 2, L), BERTScore (BS), and METEOR to assess the performance of the READER. $Rew(x)$ is the reward function, where x is the parameter.	62
6.1	Statistics of the HOPE dataset (U). The dyadic counseling conversational dataset contains a total of 12.8k utterances, each associated with one of the 12 dialogue-act labels. . . .	62
6.3	An example showing the generated response along with the reference response.	64
6.4	Human evaluation on the responses generated from READER when compared to the top two best-performing baselines. We observe that the performance of READER across all metrics is up to the mark and slightly better than the best-performing dialogue models.	65
7.1	Statistics of the HOPE dataset (U). The dyadic counseling conversational dataset contains a total of 12.8k utterances, each tagged with positive/negative sentiment.	71
7.2	The upper half shows the results obtained on the HOPE dataset. We show ROUGE (1, 2, L), BERTScore (BS), and METEOR to assess the performance of the proposed model, . The lower half of the table shows the ablation study in which we present individual contributions of two essential graph components in . Here, <i>SC-Graph</i> represents speaker-context graph, and <i>SGCR</i> denotes sentiment-guided commonsense relation graph. $\Delta_{-BEST}(\%)$ shows the percentage increment in the final model’s metric when compared with the baselines.	76
7.3	A comparative analysis of therapist’s response by the proposed model, , in contrast with the best-performing baseline (GPT-2), gold responses, and a state-of-the-art model, ChatGPT. Unlike other models, the responses generated by possess a positive influence on the client’s negative sentiments. This is evident via the model’s ability to control the response generation by exploiting <i>sentimental control</i> and <i>commonsense awareness</i> . . .	77

7.4	Human evaluation on the responses generated by EmpRes when compared to the best-performing baseline and ChatGPT, along with their statistical significance test (p-value, ρ). The performance of EmpRes across all metrics is up to the mark and slightly better than comparing systems.	79
8.1	Dataset example. The table showcases a dialogue snippet from the dataset, accompanied by trust level annotations and the discussion topic. We highlight the key portions of the text contributing significantly to the constructive and incremental trust trajectory for better interpretability.	87
8.2	Split-wise dialogue- and speaker-based analysis of the proposed dataset, MENTAL-TRUST.	89
8.3	TrustBench – Benchmarking. We evaluate 14 models on the proposed dataset, across four major categories: encoder-only, decoder-only, encoder-decoder, and closed-source. Evidently, smaller models outperform larger models across eight metrics, including Accuracy (Acc), F1, Precision (P), Recall (R), and Ordinal Cross-Entropy Loss (OL_{CE}).	91
8.4	Trust jumps and streak analysis. The table shows trust dynamics and includes the counts of trust level jump, their respective maximum, minimum, and average changes ($+\Delta$ and $-\Delta$), and the average streak length, highlighting the progression and fluctuations in trust trajectories.	94
9.1	Statistics of the BeCOPE dataset. We collected a total of $\sim 10K$ posts and $\sim 50K$ comments. We annotated all the posts using three core labels – (i) intent, (ii) criticism, and (iii) readability (Clear: Excellent, Good and Average; Non-clear: Mediocre and Poor). IAA (κ) represents the inter-Annotator agreement using Cohen’s kappa score.	101
9.2	Example of posts and their corresponding labels in BeCOPE. Intent: <i>Help-seeking, Rant, Chit-chat, and Survey</i> ; Criticism: <i>Self-criticism (SC), Criticism with reasoning (CR), Criticism with no reasoning (CNR), and No-criticism (NC)</i> ; Readability: <i>Excellent (5), Good (4), Average (3), Mediocre (2), and Poor (1)</i> ; Emotion: <i>Admiration, Amusement, Anger, Annoyance, Approval, Caring, Confusion, Curiosity, Desire, Disappointment, Disapproval, Disgust, Embarrassment, Excitement, Fear, Gratitude, Grief, Joy, Love, Nervousness, Optimism, Pride, Realization, Relief, Remorse, Sadness, Surprise, and Neutral</i> ; Engagement: <i>Interactive, Non-interactive, and Isolated.</i>	105

LIST OF FIGURES

1.1 Modular illustration of components in the digital counseling ecosystem covered in this thesis. The diagram highlights core modules to support effective and scalable counseling interactions.	3
3.1 Annotation tree for HOPE.	19
3.2 Relationship among DAC labels. The directed connection $U_t^x \rightarrow U_{t+1}^y$ shows the co-occurrence counts of two sequential utterances with x & y dialogue-acts, respectively. .	20
3.3 Architecture of SPARTA. For each utterance U_t , SPARTA computes the local-context L_t , through a time-aware attention (TAA) mechanism on the sliding-window memory unit and the current utterance. The dialogue-level global context is maintained using a GRU G_t . Finally, the speaker-aware and invariant local and global contexts are fused for the task.	22
3.4 Confusion matrix for SPARTA-TAA.	26
4.1 A sample counseling session instance from the MEMO dataset. <i>Symptom and History, Patient Discovery, Reflecting, and Discussion Filler</i> are psychotherapy elements. The summaries pertaining to this truncated snippet of conversation is shown at the bottom. <i>Note: The summary and conversation are truncated for brevity.</i>	30
4.2 Hierarchical labeling for each utterance. Counseling components including <i>SH, PD</i> and <i>RT</i> essentially contribute in summary generation.	33
4.3 Model diagram of ConSum. The Utterance embeddings are reused across three modules in the pipeline.	37
4.4 Comparison between ConSum and T5 summaries and condensed dialogue utterances based on counseling components. The y-axis represents Rouge-1 F1 scores. D' denotes the concatenation of utterances tagged with a particular counseling component, and $D - D'$ denotes the concatenation of utterances excluding the corresponding counseling component.	40
5.1 The proposed pipeline allows LLMs to first plan and then generate. In our approach, prioritizing planning before generation enriches summarization with conversational structure and domain knowledge.	43
5.2 Architecture of PIECE. We propose a novel <i>planning engine</i> consisting of two primary sections: (a) integrating knowledge filtering-cum-scaffolding and (b) encapsulating structural understanding of dialogues. Filtration of relevant utterances utilizes component labeling counseling within the MEMO dataset to mask filler utterances, followed by knowledge scaffolding. Additionally, sheaf learners are employed for the structural understanding of counseling dialogue. The planning engine operates using a rotating attention mechanism using knowledge from both segments for better LLM generation.	45

5.3	Domain-centric evaluation using Mental Health Info Capture (MHIC) metric. The proposed model, P <small>IECE</small> , distinctly excels in capturing domain knowledge compared to the two most relevant models.	51
6.1	A sample counseling conversation along with associated dialogue-acts. The proposed model – R <small>EADE</small> R takes utterance- and dialogue-act context to predict response-act and subsequently generate a response.	56
6.2	Architecture of R <small>EADE</small> R. It contains three heads on top of the foundation language model, GPT-2: (a) R <small>AC</small> -Head classifies the response-act trained on context-aware representations, (b) L <small>M</small> -Head generates the response, and (c) V-Head calculates the final reward and initiates Proximal Policy Optimization (PPO).	59
6.3	The increment flow of reward in the PPO of R <small>EADE</small> R.	64
6.4	Application of R <small>EADE</small> R – Natural setup. Given a context, at each step, R <small>EADE</small> R generates an output for the therapist. This aligns with the natural configuration of VMHAs, where a client seeks help from a bot or a virtual agent.	67
6.5	Application of R <small>EADE</small> R – Synthetic setup. Given a context, at each step, R <small>EADE</small> R takes turns to generate outputs for the therapist and the client. This setup is an analysis configuration to assess the effectiveness of R <small>EADE</small> R in handling diverse inputs (e.g., generated by an agent).	67
7.1	A sample counseling conversation in which the client expresses negative sentiment at last; the job of a VMHA is to generate a response to positively influence the client and contains rich commonsense knowledge.	69
7.2	A schematic diagram of pseudo labeling for sentiment labels for each utterance in the HOPE dataset. We augment top-5 additional attributes using COMET to assist state-of-the-art sentiment classifier in predicting sentiment label.	72
7.3	A schematic diagram of EmpRes. <i>Knowledge Extractor</i> exploits conditional commonsense relations (r_i) to construct Sentiment-Guided Commonsense Relation Graph (SGCR-Graph). A graph attention layer (GAT-SGCR), on top of SGCR-Graph, formulates sentiment-guided commonsense-aware representations. A similar layer on Speaker-Context Graph (SC-Graph), called GAT-SC, is responsible for formulating dialogue representations. Fused graph representations act as key and value for GPT-2’s attention block, responsible for therapist response generation.	73
7.4	The figure illustrates (A) the prototype deployment of the chat interface and (B) findings of the user study conducted to evaluate the real-world effectiveness of the deployed EmpRes. (C) The user study comprises user feedback of conversation with the deployed system across five major criteria – (i) perceived effectiveness, (ii) user satisfaction, (iii) continued usage, (iv) likelihood to recommend, and (v) observed hallucination.	80
8.1	Distribution of ridge density for each trust level between 1 (min) – 4 (max) in M <small>ENTAL</small> -T <small>RUST</small> .	89

8.2	Examples of trust progression scenarios (<i>increasing, constant, decreasing</i>) across top-performing models. The ground-truth trust values are represented by the blue line. Decoder-only models, particularly domain-specific variants such as Mental-BART, achieve the closest alignment with true trust values, while closed-source models perform the worst due to their inherent rigidity.	92
8.3	Topic of discussion in counseling sessions segmented into two major segments. Top: illustrating core topics in sessions with a positive topic alignment; Bottom: showing topics in sessions with negative topic alignment.	93
8.4	Illustration of the progression in trust scores, represented as either increasing or decreasing jumps. The y-axis denotes the starting trust score ("From"), while the x-axis indicates the resulting trust score ("To"). For instance, a progression from 2.0 to 2.5 corresponds to the (4, 5) cell in the upper triangular tables. Increasing trust score jumps are shown in green, while decreasing jumps are displayed in orange. Darker shades signify higher counts of the respective trust score changes.	96
9.1	Taxonomy of counseling methods along with examples. Here, OP (original poster) is a common Internet terminology for the person who creates posts on peer-to-peer platforms. In peer-to-peer therapy, we inspect the level of engagement in three different categories based on the abundance of interaction with the help-seeker – (a) interactive: if there are back-and-forth conversations between the OP and peers, (b) non-interactive: if the post engages peers, but the OP does not reply to peers, and (c) isolated: if the post does not have any comment, but one-to-one therapy involves the continuous exchange of dialogues between therapist and client (help-seeker).	100
9.2	(a) Confusion matrix to represent the performance of pseudo labeling of criticism, intent, and readability labels. We exploit BERT to fine-tune on ~5K manually annotated posts to predict criticism, intent, and readability on the remaining posts. (b) Distribution of behavioral signals (criticism and intent) along with readability in the complete BeCOPE dataset.	103
9.3	Distribution of behavioral signals and readability in BeCOPE across all engagement categories. (a) The intent distribution indicates that a majority (45.35%) of posts show explicit intentions (seek-help) through queries or the articulation of pressing needs on OMHC platforms, yielding a more productive response as opposed to merely airing surveys or rants. (b) The criticism distribution shows that help-seekers are more likely to engage in self-criticism (43.32%), and those who criticise openly on others with proper reasoning are more likely to receive assistance. (c) The readability statistics of posts in BeCOPE state that well-written posts receive 2.2× more support (responses) as compared to poorly written posts.	106
9.4	(a) Distribution of emotion labels in the BeCOPE dataset. For brevity, we show plots for the top 10 emotion labels only. Each post is tagged with primary and secondary emotion labels. We further analyze the emotion label distribution across three engagement categories. (b) Topical analysis on the BeCOPE dataset. We perform Latent Dirichlet Allocation (LDA) (2) to form 8 clusters of topics. To analyze the topics on which peers respond, we club interactive and non-interactive posts, where peers respond and compare them with topics from isolated posts.	109

1. Introduction

1.1 Thesis Overview and Statement

Natural Language Processing (NLP) has emerged as a transformative force across a wide array of domains, from machine translation and search engines to personal digital assistants (3). However, beyond these mainstream applications, its potential also arguably lies in addressing complex societal challenges—particularly those that rely on human-centric, emotionally nuanced communication. One such domain is mental healthcare, where language is not merely a medium of information exchange, but the very instrument of support (4). Within this space, counseling dialogue is a unique communicative process. It is characterized by interactive, collaborative conversations between trained mental health professionals and clients, oriented toward cultivating emotional support, enabling behavioral change, and improving overall psychological well-being. While one might intuitively imagine counseling as a one-on-one exchange between a therapist and a client, the reality is more expansive. Counseling interactions span across diverse settings, including group therapy, peer support circles, and increasingly, online community-based forums (5). These multi-party and informal forms of support, though varying in formality and expertise, still rely heavily on empathetic, context-aware communication, and thus share core attributes with clinical counseling. This broader spectrum of therapeutic dialogue demands systems that can adapt to various interpersonal dynamics, domain sensitivities, and communication styles. Importantly, counseling conversations are fundamentally distinct from traditional clinical dialogues, such as those between a doctor and a patient. Medical interactions often follow a structured, goal-oriented pattern (predominantly diagnostic in nature) centered around symptom reporting, examination, and treatment recommendations¹. These interactions, while important, tend to be transactional and time-bound. In contrast, counseling is relational and open-ended. It often involves navigating ambiguity, interpreting implicit emotional cues, and responding with sensitivity to the client's evolving mental states (6). Furthermore, the effectiveness of counseling depends largely upon the therapeutic alliance, the relational bond formed through meaningful dialogue between participants. Thus, the ultimate effectiveness of counseling hinges heavily on the quality and depth of these interactions (7). Consequently, the effectiveness of interaction is not merely a function of information accuracy or technical skill, but also tied to the quality, depth, and responsiveness of the conversation itself.

This pressing need for high-quality, emotionally intelligent communication in therapy becomes even more critical when viewed against the backdrop of the current global mental health crisis. Mental health disorders now constitute the leading cause of disability worldwide. Rates of depression, anxiety, and related conditions have surged dramatically, particularly in the wake of the COVID-19 pandemic, which has disrupted lives and support systems at a global scale². Unfortunately, this rise in demand has not been met with a proportional increase in mental health resources or professionals. Existing support frameworks (already strained under pre-pandemic conditions) now face critical bottlenecks, most notably a severe shortage of trained professionals. The World Health Organization (WHO) estimates a global deficit of over 1.2 million mental health therapists³. This gap in support delivery is not distributed evenly. It disproportionately affects marginalized groups, rural populations, and economically disadvantaged individuals, where access to mental health professionals is already limited. As a result, these shortages compound existing social inequalities, turning logistical barriers into systemic injustices. Addressing

¹<https://www.sju.edu/centers/icb/blog/doctor-patient-relationship-part-ii-theoretical-models-and-clinical-reality>

²NHS referrals for anxiety doubles the pre-Covid levels: <https://www.theguardian.com/society/article/2024/aug/27/nhs-referrals-for-anxiety-in-children-more-than-double-pre-covid-levels-england>

³Workforce shortages for mental health: <https://www.who.int/publications/i/item/9789241501019>

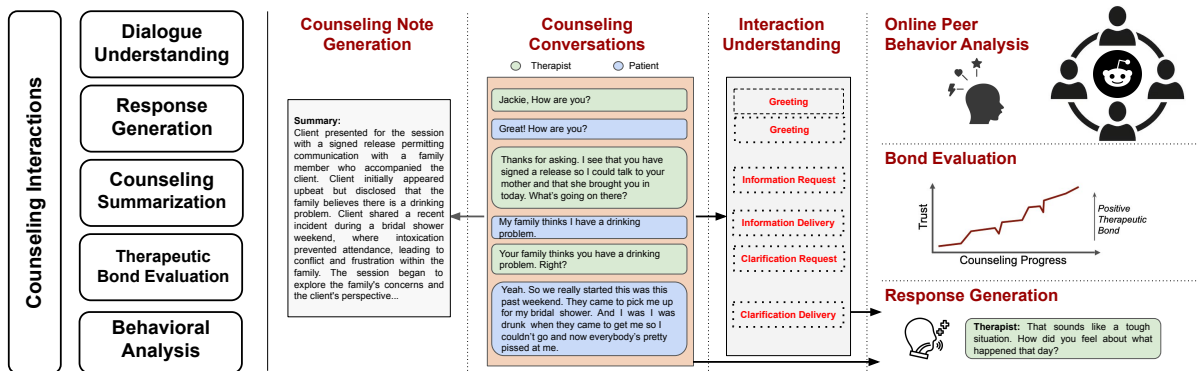


Figure 1.1: Modular illustration of components in the digital counseling ecosystem covered in this thesis. The diagram highlights core modules to support effective and scalable counseling interactions.

this gap demands not just policy-level changes but innovative technological interventions that can scale support without compromising the relational depth that counseling requires.

Virtual Mental Health Assistants (VMHAs) have emerged as a promising scalable and accessible solution, offering anonymous, stigma-free, and immediately available digital support (8). However, current VMHAs typically confine interactions to rigid, scripted dialogues, failing to replicate the nuanced adaptability of human therapists (9). Furthermore, existing implementations frequently aim to replace rather than complement human therapists, an approach this thesis critically challenges. Instead, this thesis advocates for modular-level interventions designed explicitly to address fine-grained counseling processes, enhancing the capabilities of mental health professionals rather than substituting them entirely. Specifically, this thesis addresses three major limitations of existing VMHAs: (a) superficial understanding of therapeutic nuances due to inadequate dialogue modeling, (b) limited capability to guide response generation according to therapeutic principles, and (c) neglect of longitudinal conversational dynamics, particularly regarding trust and therapeutic alliance.

Addressing these challenges requires a comprehensive VMHA pipeline, which is very similar to a standard dialogue system pipeline containing *dialogue understanding* module for interpreting client’s directive; *dialogue management*, responsible for maintaining coherence and adapting dialogue strategies; and *dialogue generation* module for producing effective and clinically informed responses (10). Unlike current LLM’s that take an end-to-end approach, our modular approach gives us an edge over *dialogue generation* block with two major controlling attributes: (a) dialogue understanding controlling parameter, and (b) domain/external knowledge contributing parameter to provide factually relevant and effective responses. Additionally, a robust evaluation framework is vital, continuously assessing performance metrics and therapeutic quality, thus enabling feedback-driven improvements across understanding and generation components. Such a system naturally branches into essential practical applications, most notably *Counseling Note Generation*, which is mandatory during counseling processes and provides clinicians with structured, clinically actionable insights. Hence, as shown in Figure 1.1, this thesis proposes five interconnected research thrusts within counseling interactions:

- **Conversational Understanding:** Development of a interaction-directive taxonomy and integrating therapeutic intent, realized through domain-adapted transformer architectures.
- **Counseling Summarization:** A knowledge-enhanced summarization framework, linking understanding parameters explicitly to clinical concepts such as PHQ-9 criteria, enabling clinicians to efficiently capture therapeutic nuances within conversation summaries.
- **Controlled Response Generation:** Utilizing reinforcement learning along with informed psy-

chotherapeutic principles to guide empathetic, clinically aligned language modeling.

- **Trust Dynamics Modeling:** Study of dynamic computational models to quantify trust as a therapeutic bond assessment metric through longitudinal analysis of interactions.
- **Peer Interaction Decoding:** Behavioral cue analysis within OMHCs, extracting critical interaction patterns, such as intentional behavior, criticism, and emotional patterns, to understand and enhance peer-led support interventions.

With state-of-the-art methods and established psycholinguistic theories, this thesis seeks to transform support delivery from simple conversational agents into nuanced, assistive methods. Ultimately, these contributions aim not merely at algorithmic improvement but at genuinely democratizing support to the mental healthcare ecosystem, supporting mental health professionals, and enriching peer-support environment.

This thesis focuses on supporting mental health professionals with novel methods that understand, evaluate, and enhance therapeutic interactions in digital spaces. We study dialogues (both peer and individual), personalized response generation, and assessing behavioral dynamics to bridge NLP with counseling dynamics to enhance support delivery strategies via experts.

1.2 Preliminary

We first discuss the preliminaries under two broad categories: *Conversation*, which covers the core components of interactions in mental health, and *Psychology and Counseling*, which outlines the fundamental principles of therapeutic communication and behavioral analysis.

1.2.1 Conversations

Dialogue systems for mental health care must address several core challenges to enable effective human-AI interactions. These challenges span dialogue understanding, generation, summarization, evaluation, and the integration of large language models (LLMs).

1. **Dialogue Understanding.** At the heart of any effective dialogue system lies the ability to understand and interpret user inputs. In mental health counseling, this involves recognizing directives, such as clarification, questioning, and help-seeking, along with emotional states. Traditional dialogue systems often focus on transactional tasks, but counseling conversations require a deeper, context-aware understanding of clients's directive and emotional subtext (II).
2. **Dialogue Generation.** The ability to generate empathetic and contextually appropriate responses is critical for fostering therapeutic alliances. Unlike rule-based or retrieval-oriented systems, modern VMHAs must produce responses that are personalized, emotionally resonant, and clinically safe. Reinforcement learning with human feedback (RLHF) has emerged as a powerful approach, enabling systems to align responses with therapeutic principles. By incorporating domain knowledge, dialogue generation models can deliver responses that enhance the overall therapy experience.
3. **Dialogue Summarization.** Summarizing counseling conversations is a 'mandatory' yet challenging task, requiring the distillation of complex contextual and behavioral insights into concise and

effective summaries. Traditional summarization models often miss domain-specific nuances, such as therapeutic knowledge and therapy patterns (12; 13). Proposed approaches leverage clinical knowledge bases, such as PHQ-9 criteria, to enhance the relevance and utility of summaries. Automated summarization not only aids clinicians in documentation but also ensures continuity of care by capturing critical insights from therapy sessions.

4. **Dialogue Evaluation.** Evaluating the performance of dialogue systems in mental health care requires a combination of automated metrics and human assessments. Metrics such as BLEU and ROUGE measure fluency and coherence, while empathy classifiers and clinical evaluations assess therapeutic quality (14). Longitudinal metrics, such as therapeutic bond and trust levels, provide insights into the long-term effectiveness of interactions. Together, these evaluation methods ensure that dialogue systems meet both technical and clinical standards.
5. **LLMs as Dialogue Systems.** Portraying LLMs as a dialogue system represents a significant advancement, enabling users to generate more nuanced and contextually aware responses. However, LLMs also introduce challenges, such as hallucination risks and ethical concerns (15). Fine-tuning strategies, such as instruction-tuning on counseling-specific prompts and hybrid architectures have proved to be methods to bring domain nuances to some extent. Industry standards combine traditional deep learning methods with rule-based modules, offer promising solutions, but limits the generation within tight boundaries.

1.2.2 Psychology & Counseling

Mental health care introduces unique domain-specific requirements that NLP systems must accommodate.

1. **Counseling Notes.** Counseling notes are structured documents that capture key insights from therapy sessions, including client history and trajectories (16). Automated note-generation systems must balance conciseness with clinical relevance, ensuring that summaries are both effective and comprehensive. By integrating domain knowledge and clinical expertise, NLP models can produce summaries that enhance efficiency and effectiveness.
2. **Online Mental Health Communities (OMHCs).** OMHCs, such as Talklife⁴ and 7 Cups⁵ provide vital spaces for peer support and shared experiences. These platforms are rich in behavioral and linguistic data, offering unique insights into peer counseling dynamics. At the same time, these solutions come with a financial cost (17; 18). On the other hand, Reddit acts as a vast, free and open community to discuss all sorts of interactions, including a peer support ecosystem. It is worth noting that the unregulated nature of OMHCs also poses challenges, such as varying quality of support and potential for harmful interactions. By analyzing these interactions, we identify patterns of effective peer support and develop strategies for enhancing the quality of online interactions.
3. **Trust Dynamics.** Trust is a cornerstone of effective counseling, encompassing factors such as communication effectiveness, reliability, and confidentiality. In virtual counseling environments, quantifying and modeling trust dynamics is essential for fostering long-term user engagement.

⁴Talklife: <https://www.talklife.com/>

⁵7Cups: <https://www.7cups.com/>

1.3 Thesis Organization

This thesis is organized into five major parts, each addressing distinct yet interconnected research verticals. Below, we outline the structure and objectives of each part, highlighting their contributions to the overarching goal. A comprehensive list of abbreviations used in this thesis is provided in Table [1.1](#).

Part I: Conversational Understanding

In this part, we tackle the task of conversational directives understanding, particularly in the context of mental health counseling dialogue. Particularly, we first propose a novel first-of-its-kind counseling conversation dataset, HOPE, along with an expert-led manual annotation task of dialogue acts. The HOPE dataset is a curated collection of 12.9K annotated utterances sourced from publicly available counseling sessions. This dataset is tailored explicitly for mental health dialogue understanding, providing a robust foundation for training and evaluating models. Building on this, we propose SPARTA, a speaker- and time-aware transformer architecture designed to capture the dynamic and contextual nuances of counseling conversations. SPARTA achieves state-of-the-art performance in dialogue-act classification, demonstrating its ability to interpret and contextualize user queries and responses effectively.

Part II: Counseling Summarization

In this part, we address the challenge of summarizing mental health counseling conversations by framing it as a structured language modeling task. Specifically, we tackle two major problems: (a) domain knowledge infusion and (b) planning-based generation in LLMs, each as a standalone chapter. Unlike general text summarization, counseling summarization requires extracting clinically relevant insights while preserving key contextual nuances. To achieve this, we propose ConSum, a knowledge-guided summarization framework that integrates domain expertise with dialogue structure awareness. ConSum explicitly aligns conversational dynamics with mental health knowledge sources, such as PHQ-9 criteria, to enhance summary effectiveness. We further extend the HOPE dataset with new annotations for counseling components and reference summaries, introducing MEMO, a dataset tailored for structured counseling note generation. Beyond direct domain knowledge infusion, we also enhance LLM-based summarization by introducing a novel planning engine that orchestrates structured knowledge alignment before generation. This planning engine employs a knowledge filtering-cum-scaffolding mechanism, ensuring that domain-relevant insights are selectively integrated into summaries. This chapter not only highlights the state-of-the-art in counseling summarization but also presents a practical evaluation from mental health professionals, offering insights into how such solutions can assist in real-world counseling note generation and clinical documentation.

Part III: Dialogue Generation

Unlike LLMs, where response control is primarily limited to prompt engineering, this part explores modular-level control mechanisms for dialogue generation. Controlled response generation ensures that generated outputs remain contextually aligned, domain-relevant, and therapeutically appropriate. Instead of relying on implicit biases in training data, we introduce explicit control parameters to steer the response generation process effectively. We focus on two key control mechanisms: (a) dialogue-directive guided response generation and (b) commonsense-knowledge guided response generation. To implement directive-control, we propose READER, a transformer-based response generation model that employs

reinforcement learning with human feedback (RLHF). READER integrates predicted future dialogue-acts and semantic similarity metrics to optimize responses for contextual coherence and therapeutic value. To address commonsense control, we introduce EmpRes, a response generation model that incorporates commonsense knowledge from the COMET model. By using commonsense reasoning as a controlling parameter, EmpRes enhances response interpretability and situational awareness, ensuring that VMHAs generate responses that are both logically coherent and psychologically informed. This chapter studies response generation for VMHAs by moving beyond static, predefined response models and introducing adaptive, knowledge-aware control mechanisms, leading to more effective and human-like conversational agents.

Part IV: Assessing Client’s Bond with the Assistant

This chapter focuses on conceptualizing and modeling trust in virtual counseling environments, aiming to forecast trust dynamics and assess the evolving therapeutic bond between clients and VMHAs. Trust plays an important role in counseling, influencing the client’s openness and long-term adherence to conversation. In this line of work, we propose trust as a therapist-assistive metric, allowing experts or bots to adapt based on user trust levels over the course of interactions. We define trust as a dynamic trajectory where a client’s willingness and openness to express themselves directly impact their therapeutic experience. To study and benchmark computational trust modeling, we release MENTAL-TRUST, a novel counseling dataset annotated with seven expert-verified ordinal trust levels, making it the first-of-its-kind in this domain. To quantify trust, we frame the problem as an ordinal classification task and propose TrustBench, a benchmarking framework evaluating a suite of classical and state-of-the-art language models on MENTAL-TRUST. We conduct exhaustive experiments, evaluating model performance across multiple metrics and analyzing key patterns in trust evolution within therapeutic interactions. This study not only lays the groundwork for automated trust modeling in a conversational setting but also provides actionable insights into how trust develops and fluctuates in mental health counseling.

Part V: Peer Behavior in Online Mental Health Communities (OMHCs)

OMHCs, such as Reddit and 7 Cups, have emerged as vital platforms for individuals seeking information, support, and guidance in managing their mental health needs. These platforms offer users immediate interactions with peers, creating spaces for shared experiences and mutual support. However, the largely unregulated nature of these platforms presents complex challenges, including varying quality of support and potential for harmful interactions. This chapter explores the behavioral dynamics of peer counseling threads, focusing on factors such as intent, criticism, readability, and emotional expression.

To achieve this, we curate the BeCOPE dataset, comprising over 10,118 posts and 58,279 comments sourced from 21 mental health-specific subreddits. The dataset is annotated with fine-grained behavior labels, enabling nuanced analysis of peer engagement patterns. Our findings reveal the prominence of self-criticism as a prevalent form of criticism among help-seekers and highlight the role of readability in eliciting peer support. By triangulating multiple sources of data, we provide mental health professionals and online community moderators with actionable insights for enhancing the effectiveness of peer support services. This chapter underscores the essential role of OMHCs in offering personalized guidance and support to individuals managing mental health issues, contributing to the broader understanding of peer counseling dynamics.

Beyond one-on-one counseling, peer interactions in online mental health communities (OMHCs) play a critical role in providing accessible, stigma-free support. In this part, we investigate peer-led mental

health interventions by analyzing how users seek and provide support in OMHCs. To facilitate this, we curate BeCOPE, a dataset of over 10,118 posts and 58,279 comments from 21 mental health-specific subreddits, annotated with fine-grained behavioral cues such as help-seeking intent, emotional support, and criticism. Our analysis uncovers key behavioral patterns in peer counseling, providing actionable insights for community moderation, peer-support enhancement, and hybrid AI-assisted mental health interventions.

1.4 Ethical Considerations

Given the high-stakes nature of mental health, ethical considerations have been integrated throughout the thesis, guiding decisions at every stage from dataset creation to model development to output evaluation. This thesis treats ethics not as an afterthought but as a continuous process spanning three broad levels: data preparation and processing, modeling, and output assessment. At the data stage, all datasets are either publicly available, de-identified counseling data, or data obtained with explicit usability permissions. Personally identifiable information (PII) is removed prior to annotation and analysis. Annotations and data reviews are carried out by mental health professionals with extensive clinical and computational expertise, with ethical integrity. For OMHCs, no user-identifiable information is retained; only relevant mental health subreddits are included. Next, in the modeling stage, architectures are designed with bias minimization and safety as explicit goals. Mechanisms such as knowledge filtering and planning are incorporated to reduce hallucinations, domain constraints are applied to ensure therapeutic relevance, and modularity is built in to allow substitution of taxonomies or knowledge sources for different contexts. No diagnostic claims are made, and systems are intended to be therapist-assistive rather than patient-facing. During development, any observed bias or unintended behaviors prompted updates to both model training and data pipelines. Furthermore, at the output assessment stage, generated outputs are evaluated for clinical acceptability by licensed domain experts following established frameworks, including assessments of relevance, coherence, safety, and factual accuracy. In experimental deployments such as the EmpRes user study, participants are provided with clear guidelines and declarations, participation is voluntary, and no PII is collected. System outputs are monitored for hallucinations, emotional sensitivity, and potential bias, with corrective measures applied where necessary. Across all contributions, the systems developed are intended to support mental health professionals rather than replace them, keeping the human/expert-in-the-loop so that clinical decision-making authority remains with the therapist. While this thesis ensures compliance with core research ethics and fair research standards, it acknowledges limitations such as potential cultural bias in datasets due to size and diversity constraints, as well as the need for broader coverage. Deployment-ready systems will require additional steps to align with formal regulatory frameworks and institution-specific protocols (more details in Chapter 10).

Abbreviation	Full form
AI	Artificial Intelligence
BeCOPE	BEhavior enCODED PEer counseling
CNN	Convolutional Neural Networks
ConSum	Counseling Summarizer
DAC	Dialogue-Act Classification
DPA	Dot Product Attention
EmpRes	EMPathetic RESponse generator
GRU	Gated Recurrent Unit
HOPE	mental Health cOunselling of PatiEnts
LLMs	Large Language Models
MEMO	Mental hEalth suMmarizatOn dataset
MHIC	Mental Health Information Capture
OMHC	Online Mental Health Communities
PHQ	Patient Health Questionnaire
PIECE	Planning engine for mental Counseling notE generation.
PLMs	Pretrained Language Models
PPO	Proximal Policy Optimization
READER	REsponse-Act guided reinforced Dialogue genERation
SPARTA	SPEaker and time-AwaRe conTextual trAnsformer
TRL	Transformer Reinforcement Learning
VMHA	Virtual Mental Health Assistant

Table 1.1: Abbreviations and their definitions used in the thesis.

2. Background

2.1 Mental Health Challenges and Digital Interventions

The current global mental health crisis is unprecedented, with millions affected worldwide and a severe shortage of mental health professionals to meet the rising demand (19). Studies have highlighted that in many regions, including low- and middle-income countries, access to adequate mental health care is limited (20; 21). This scarcity is further exacerbated by social and economic barriers, leaving many individuals without timely or effective support. The urgent need for scalable, accessible interventions has created a gap for researchers and practitioners to explore solutions that extend beyond traditional methods of therapy (22). Digital interventions have emerged as a promising alternative to conventional mental health services, essentially offering cost-effective and immediate support through virtual platforms (23). Over the past decade, there has been a significant shift towards the use of digital mental health tools, including VMHAs and online counseling platforms (24). These technologies leverage advanced natural language processing (NLP) techniques to provide users with personalized guidance and empathy, thereby partially relieving the burden on overextended healthcare systems (25; 26). By delivering AI-based support through scalable means, digital interventions hold the potential to bridge the existing care gap. On the other hand, understanding how peer communities and many-to-many interactions operate has shown that OMHCs provide self-disclosure and community-based validation, which are vital components of effective support (27; 28). These platforms enable users to express their concerns anonymously, seek advice, and gain emotional reinforcement from peers facing similar challenges (29). However, the unregulated nature of these interactions also introduces challenges in terms of ensuring that the support provided is both accurate and constructive (30).

Despite these advancements, current digital interventions and OMHCs face significant limitations (31). Many systems are still grappling with issues such as inconsistent response quality, lack of contextual understanding, and an inability to dynamically assess and adapt to users' needs (24). Recent research has begun to address these gaps through dialogue modeling (32). With the help of state-of-the-art methods and studies from psychological research on therapeutic interactions, there is a clear opportunity to create more responsive and effective digital supportive methods that can improve outcomes and support mental health professionals at scale (33).

2.2 Thematic Advances in Counseling Conversation Analysis

Recent research has increasingly focused on understanding the nuanced dynamics of therapeutic interactions (34). A central theme across several studies is the evolution of therapeutic bonds, conceptualized as a dynamic trajectory that unfolds throughout a session (35). While earlier work relied on static psychometric instruments to gauge trust (36), the current literature emphasizes capturing moment-by-moment fluctuations in a patient's willingness to disclose sensitive information (37). Such dynamic trust modeling is vital for adapting therapeutic strategies (35). In parallel, the analysis of peer interactions has revealed that user engagement is strongly influenced by behavioral and linguistic cues (38). Studies have demonstrated that online behavior is a critical determinant of whether help-seekers receive constructive support (25; 27; 28). Another important strand of research focuses on the integration of commonsense knowledge and domain-specific cues to enhance response generation in counseling dialogues (39). With more and more external knowledge being infused into the dialogue systems, such as with the usage of

knowledge-graphs (40), say *COMET* (41), studies have shown that it is possible to inject commonsense reasoning into dialogue systems, thereby producing responses that are both contextually relevant and effectively supportive (42). Finally, thematic analysis of counseling conversations—especially in peer-to-peer settings—reveals that behavioral signals like self-criticism, emotional expression, and clear problem articulation play a crucial role in fostering engagement (43). Research on online interactions indicates that posts with higher readability and explicit help-seeking language significantly increase the likelihood of receiving peer support (25). These findings not only provide important insights into the factors that drive a constructive support ecosystem in online platforms but also highlight the potential for taking the research further. (44)

2.3 Dialogue Systems for Mental Health

The evolution of dialogue systems over time has been marked by a steady transition from rule-based (45) and statistical methods (46) to sophisticated neural approaches (47). Early dialogue systems relied on handcrafted rules and sparse representations, which, while effective for controlled environments, struggled to capture the fluidity and nuance of human conversation (48). With the origin of deep learning, particularly transformer-based architectures such as GPT-2 (49) and DialoGPT (50), dialogue systems began to generate responses that were not only grammatically correct but also contextually relevant (49; 51). These systems paved the way for more dynamic and interactive applications, setting a strong foundation for their adoption in diverse fields, including mental health (52). Building on these advances, research in dialogue systems has increasingly focused on incorporating richer context and domain-specific cues to better mirror the complexity of human interactions (10). In the mental health domain, this evolution is particularly critical given the sensitive nature of therapeutic conversations (53). Early approaches in this area were often limited to basic conversational agents that lacked an understanding of nuanced structures (54). Recent studies have addressed this limitation by integrating more control over dialogue understanding and dialogue generation that capture both global context and speaker-specific dynamics (55). By employing reinforcement learning (56) and transformer-based methods (57), modern systems now aim to generate responses that not only reflect the immediate conversational context but also align with therapeutic objectives.

2.4 Integration of Domain-Specific Knowledge and Commonsense

A critical component in the development of VMHA is the integration of external knowledge with domain-specific cues. Traditional dialogue systems, while powerful in generating coherent text, often lack the depth required to understand nuanced emotional contexts and therapeutic needs (58). To address this, recent research has turned to external knowledge bases such as ATOMIC (59) and ConceptNet (60), with models like *COMET* (41) enabling the extraction of rich, structured commonsense information. By incorporating these resources, dialogue systems can generate responses that are not only contextually appropriate but also imbued with human-like reasoning about everyday situations (61). Alongside commonsense, domain-specific knowledge plays a pivotal role in tailoring responses for more relevance (62). Specialized pretrained models—such as *MentalLlama* (63) and *Mental-BERT* (64)—are fine-tuned on mental health data, thereby capturing psycholinguistic nuances that generic models tend to miss. These models support the generation of responses that resonate more closely with the domain aspects of therapeutic interactions (65). The integration of domain-specific embeddings with commonsense reasoning enables a richer, multi-dimensional understanding of the conversation, allowing systems to align their responses with both factual content and emotional context (66).

Recent studies have also explored methods for fusing external knowledge with internal model representations (67; 68). For example, modified attention mechanisms have been proposed that incorporate enriched knowledge representations from graph-based models (42). By adapting the multi-head attention in transformers to include keys and values derived from both commonsense and domain-specific sources, these approaches enable a more controlled and context-sensitive response generation process (69). Such methods not only improve the semantic richness of the output but also help maintain consistency in emotional tone and therapeutic intent. Furthermore, integrating external knowledge into dialogue systems has proven particularly beneficial in mitigating issues such as neural text degeneration and incoherent output (70; 71). Models augmented with commonsense knowledge have shown improved capability in generating responses that are both informative and emotionally supportive (41; 42). This integration is especially critical in the mental health domain, where the subtleties of empathy and sentiment can significantly impact therapeutic outcomes (72). Overall, the fusion of domain-specific pretrained models with external commonsense frameworks represents a promising avenue for developing robust and empathetic virtual mental health assistants (73).

2.5 Summary and Gap Identification

The series of literature reviewed in this chapter illustrates a rapid evolution of dialogue systems, from early rule-based approaches to sophisticated transformer-based architectures that now underpin mental health applications. In particular, recent studies have demonstrated that understanding interactions' core attributes and state-of-the-art dialogue modeling techniques significantly improves the contextual relevance and therapeutic alignment of generated responses (74; 75). However, despite these advances, a key challenge remains: most existing systems are primarily optimized for generic conversation without a dedicated focus on the unique nuances of mental health counseling. Equally, the growing crisis in global mental health has spurred the development of digital interventions that aim to bridge the gap in access to care. Studies have shown that VMHAs and OMHCs can provide cost-effective, scalable support (25). Nevertheless, these platforms often fall short in dynamically capturing and responding to complex signals. Another important dimension emerging from the literature is the integration of external domain-specific knowledge into dialogue systems. Models that combine pretrained mental health-specific embeddings with commonsense reasoning frameworks (e.g., COMET) have shown promise in generating emotionally nuanced and context-aware responses (41; 42). Yet, the challenge of effectively merging these heterogeneous sources of information remains unresolved. Current approaches often struggle with maintaining consistency and avoiding hallucinations while ensuring that the responses are both informative and empathetic.

Overall, while substantial progress has been made across multiple facets, a significant gap persists in unifying these threads into a coherent, adaptive framework that can eventually support either the professionals or methods in the space of support delivery. This thesis focuses on developing holistic models that seamlessly combine these elements, ensuring that digital interventions not only simulate human-like dialogue but also continuously understand to the complexities in human interactions to further have better control over the generations of aforementioned methods.

Part I

Conversational Understanding

3. Dialogue-act Classification

The onset of the COVID-19 pandemic has brought the mental health of people at risk. Social counseling has gained remarkable significance in this environment. Unlike general goal-oriented dialogues, a conversation between a patient and a therapist is considerably implicit, though the objective of the conversation is quite apparent. In such a case, understanding the intent of the patient is imperative in providing effective counseling in therapy sessions, and the same applies to a dialogue system as well. In this work, we take forward a small but an important step in the development of an automated dialogue system for mental health counseling. We develop a novel dataset, named HOPE, to provide a platform for the dialogue-act classification in counseling conversations. We identify the requirement of such conversation and propose twelve domain-specific dialogue-act (DAC) labels. We collect $\sim 12.9K$ utterances from publicly available counseling session videos on YouTube, extract their transcripts, clean them, and annotate them with DAC labels. Further, we propose SPARTA, a transformer-based architecture with a novel speaker- and time-aware contextual learning for the dialogue-act classification. Our evaluation shows convincing performance over several baselines, achieving state-of-the-art on HOPE. We also supplement our experiments with extensive empirical and qualitative analyses of SPARTA.

3.1 Introduction

Mental illness remains an alarming global health issue today. Due to the COVID-19 pandemic, there has been a significant growth in mental health disorders such as depression, attention deficit hyperactivity disorder (ADHD), and hypertension (76). A recent study shows an unprecedented 20% increase in patients with mental health illness¹. Similar study discusses the adverse impact on the mental health of US college students due to the pandemic (77).

Counseling therapy can benefit many people at risk by providing them emotional support. Amidst the surge in the number of patients, it has become a challenge for therapists to diagnose too many patients. On the other hand, patients have found it difficult to access the services of the therapist amid lockdown. Counseling therapy is a sophisticated procedure that deals with the expression of emotion and intent of patients with different personalities. To build a strong therapeutic relationship with the patient, it is essential for a therapist to develop a better understanding of the implicit intents of the patients. The nature of conversations in a social counseling setting is particularly distinct as compared to a conventional chit-chat or goal-oriented conversations. It follows a pattern which is different from both goal-oriented and general chit-chat based conversations. Usually these conversations begin with greetings followed by the therapist inquiring for problems faced by the patient. The therapist usually delves deeper into a particular problem acquiring as much context and fine-grained information before advising a remedy. These conversations also heavily utilise the contextual information of the entire conversation history. Moreover, the prime objective of the conversation is to understand the explicit and implicit requirements of the patients and suggest potential solutions accordingly. In comparison, a traditional goal-oriented dialogue system does not regard any implicit requirements, whereas a chit-chat based system lacks a

¹<https://www.kff.org/coronavirus-covid-19/issue-brief/the-implications-of-covid-19-for-mental-health-and-substance-use/>

Utterance	DAC
Therapist: Jackie, how are you?	Greeting
Patient: Okay, How are you?	Greeting
Therapist: Thanks for asking. I see that you have signed a release so I could talk to your mother and that she brought you in today. What’s going on there?	Information Request
Patient: They think I have a drinking problem. My family..	Information Delivery
Therapist: Your family thinks you have a drinking problem?	Clarification Request
Patient: Yeah. So we really started this was this past weekend....	Clarification Delivery
Therapist: So they asked you to come into the agency?	Clarification Request
Patient: Yeah, you know, I don’t want them to hate me or anything. So I agreed to come.	Clarification Delivery

Table 3.1: Example of a sample conversation session between a patient and a therapist. Each utterance has an associated dialogue-act classification (DAC) label.

target and does not care about the final solution. Another major difference is the length of utterances and conversations in a counseling session. These are particularly lengthy as patients describe their issues, while the therapists list out possible causes and preventive solutions.

The task of Dialogue-act classification (DAC) is cardinal in a dialogue system and even more so in counseling based conversations. It deals with understanding the intended requirements of the utterances, which essentially act as one of the precursors for the dialogue response generation. For instance, we present an example of a therapy session in Table 3.1. For each utterance, a corresponding label defines its dialogue-act. The first two utterances are a part of the complementary greetings that usually occur at the beginning of a natural conversation. Subsequently, in the third utterance, the therapist leads the conversation and requests for information. In response, the patient delivers the requested information.

Earlier studies like (78; 79) tackle the task of dialogue-act classification on chit-chat based conversation datasets such as Switchboard corpus (80). Their proposed architectures take into account the contextual dependency of an utterance that aids in efficient dialogue-act classification. For example, an utterance tagged as having a dialogue-act ‘question’ has a high probability of being followed by an utterance with tag ‘answer’. In another work, (81) argued that the information of speaker change is a critical feature in the dialogue-act classification task. Considering the severity of the issue and the complexity of the task, designing an automated system can facilitate the counseling sessions or assist the therapist, thus allowing them to cater to more patients. Literature in the natural language processing domain suggests a significant effort in understanding and building models for conversational dialogue (82; 83; 84; 85). However, there are hardly any models that support mental-health counseling as a dialogue system; this is primarily due to lack of data.

In this work, we aim to address these limitations by creating the HOPE² dataset which consists of therapy conversations covering cognitive-behavioral therapy (CBT), child therapy, family therapy, etc. The HOPE dataset contains $\sim 12.9K$ utterances across 212 mental-health counseling sessions. Each utterance in the dataset is tagged with one of the 12 counseling-aligned dialogue-act labels (c.f. Section 5.3). We also propose SPARTA³, a novel speaker- and time-aware contextual transformer model for dialogue-act classification. SPARTA exploits both the local and global contexts along with the speaker-dynamics in the dialogue. We model the problem as a dialogue-level sequence classification task, where the aim is to predict an appropriate dialogue-act for each utterance in a dialogue. To incorporate the global context, we employ a Gated Recurrent Unit (GRU) (86) that takes an utterance representation at each step of the

²Mental Health cOunselling of PatiEnts

³SPeaker and time-AwaRe conTExtual trAnsformer

dialogue. In addition, we introduce a novel time-aware attention mechanism to capture the local context – a sliding-window based memory unit is maintained, and subsequently, a cross-attention between the current utterance and the memory unit is computed. Our evaluation shows substantial improvement in performance in comparison to the recent state-of-the-art systems. Furthermore, we provide empirical evidences for each module of SPARTA using an extensive ablation study and detailed analyses. We summarize the main contributions of our current work as follows.

- We present HOPE, a novel and large-scale manually annotated, counseling-based conversation data for dialogue-act classification. To the best of our knowledge, the current study is one of the first efforts in compiling a dataset related to the mental-health counseling dialogue system.
- To cater to the requirements of counseling conversations, we define a novel hierarchical annotation scheme for the dialogue-act annotation. We propose twelve dialogue-act labels that are aligned with mental-health counseling session.
- We propose SPARTA, a novel dialogue-act classification system that combines speaker-dynamics and local context through a time-aware attention mechanism, along with long-term global context.
- We perform extensive ablation study to establish the efficacy of each module of SPARTA. Furthermore, the comparative analysis shows that it attains state-of-the-art performance on HOPE.

Societal Impact. A significant increase in the number of mental health issues has been observed in the last few years. The lack of therapists is a stumbling block to the mental health of society. Therapist-Bots (mental health chatbots) could bridge the gap by effectively interacting with patients and understanding them. Conversely, end-to-end chatbots in the mental health domain are delicate, where every aspect of the therapy is needed to be perceived precisely. Our research aims at the dialogue-understanding module in the mental health conversational system. The ongoing research in the mental health domain could exploit this work and benefit the chatbots to understand the therapy conversation in a better way.

3.2 Related Work

The current work is connected to the existing literature in at least two dimensions – first, the dialogue-act classification, and second, text processing for mental health counseling. We present our literature survey for both dimensions.

Mental Health and the role of text processing. The impact of Natural language processing in the study of mental health is substantial. Though the field of therapeutic discourse analysis has been around since 1960s (87), research on dialogue systems in mental health domain is in a nascent stage. Previous research on mental health intervention systems primarily focused on the problems related to the [early suicide ideation detection through by combining text-based cues with social network information and historical author profiling](#) (88) or generating empathetic responses to users (89). (90) studied the mental health notes to detect suicide ideation, whereas, (91) employed text processing to detect symptoms of anxiety and depression using social media text. In another work, (92) showed the importance of several keywords in revealing users’ social and psychological behaviours. (93) proposed an emotion-aware model for human-like emotional conversations.

(94) presented the Distress Analysis Interview Corpus (DAIC) to identify verbal and non-verbal cues of distress during an interview. Among other methods, the data collection was done using the automated agent Ellie based on the work of (95). Recently, (96) presented a survey of the chat-bots in the mental health domain. The authors compared the strengths and weaknesses of three existing conversational

agents, namely Wysa⁴, Woebot⁵, and Joy⁶. The drawbacks of these systems are that some of these are rule-based, while others are primarily data collection module for an offline counselling. In comparison, ours is the effort in the development of an online counselling system.

Dialogue-act Classification. Studies on dialogue systems have always fascinated researchers ever since ELIZA (97), the first rule-based system was developed. The dialogue-act classification module is one of the most critical components of a dialogue agent which caters to the requirement of the dialogue system by serving at the natural language understanding helm of the dialogue system.

Previous research treats the problem of dialogue-act classification either as a standalone text classification task or a sequence labelling task. Recently, (98) suggested a sequence-to-sequence text generation approach for the dialogue-act classification. Earlier studies like (99) and (100) focused on lexical, syntactical, and prosodic features for classification. In another work, (101) employed CNNs (102) and CRFs (103). (104) proposed a method based on CNNs and RNNs (105) that used the previous contextual utterances to predict the dialogue-act of the current utterance.

(78) proposed a contextual dialogue-act classifier (CDAC) and used transfer-learning to train their model on human-machine conversations. The model proposed by (106) uses self-attention mechanism on RNNs to achieve impressive results on benchmark datasets. However, these works do not take into account the speaker-level information which is imperative in social-counseling based conversations. (107) proposed a method to detect relevant context in retrieval-based dialogue systems. (108) proposed a CRF-attentive structured network to capture the long-range contextual dependencies using structured attention mechanism. (109) proposed to classify concurrent dialogue-acts of an utterance by modelling the contextual features. Recently, (110) used co-interactive relation networks to jointly capture sentiment and the associated dialogue-acts with an utterance. Their model achieved significant results on Mastodon (84) and DailyDialog (82) datasets. Similarly, (111) jointly learns the dialogue-act classification and emotion recognition tasks in a multi-modal setup.

To the best of our knowledge, the model by (112) is the first one which takes speaker transitions for DAC into account. It uses a modified version of CRFs to capture the speaker-change and achieves state-of-the-art results on SwitchBoard dataset (80). Similar to earlier works, we also treat DAC as a dialogue-level sequence labelling task. We jointly take the global and local contexts of the conversation and the speaker of the utterance as the key factors for the classification. We hypothesize that such information offers crucial clues at different stages of the model. In contrast, the existing systems had incorporated the role of global context and speakers dynamic independently.

3.3 Dataset

In this section, we present our dialogue-act classification dataset, called HOPE. In total, we annotated $\sim 12.9K$ utterances with 12 dialogue-act labels carefully designed to cater to the requirements of a counseling session. The remaining section furnishes the details of data collection, annotation schemes, dialogue-act labels and necessary statistics.

⁴<https://www.wysa.io/>

⁵<https://woebothealth.com/>

⁶<https://www.hellojoy.ai/support>

Split	Speaker	Dialogue-act labels											Total	
		ID	IRQ	GT	GC	CRQ	YNQ	CD	ACK	PA	NA	OD		ORQ
Train	Patient	1886	23	155	494	77	12	666	607	269	271	205	3	4668
	Therapist	393	1474	212	829	710	692	83	125	6	7	40	180	4751
Test	Patient	466	7	32	122	13	6	189	97	73	66	48	-	1119
	Therapist	66	376	44	209	199	134	15	31	6	-	8	34	1122
Val	Patient	246	3	17	83	8	-	97	63	34	27	16	1	595
	Therapist	41	194	22	113	101	85	10	16	-	-	3	14	599
Total	Patient	2598	33	204	699	98	18	952	767	376	364	269	4	6382
	Therapist	500	2044	278	1151	1010	911	108	172	12	7	51	228	6472

Table 3.2: Dialogue-act distribution in HOPE. The train, test, and validation splits are 70:20:10

Data Collection

One of major hurdles we faced in the process of data collection was the unavailability of public counseling sessions, mainly due to the fact that they usually contain sensitive personal information. To curate this data, we carefully explored the web and collected publicly-available pre-recorded counselling videos on YouTube. A substantial proportion of these recordings are mock or demonstration therapy sessions created for accredited training programs, educational resources, or public outreach by licensed professionals, and were selected for inclusion based on conversational coherence, adherence to clear therapeutic structure, and alignment with accepted counseling techniques. To ensure confidentiality, we randomly assign synthetic names to all patients and therapists in all examples.

In the next step, we extract the transcriptions of each video using OTTER (<https://www.otter.ai/>), an automatic speech recognition tool. Subsequently, we correct transcription errors to remove any noise (i.e., spelling or grammatical mistakes). The data collection process provides us 12.9K utterances from 212 counseling therapy sessions – all of them are dyadic conversations only.

Dataset Annotation Scheme

Since the counseling conversations have inherent differences with the standard conversations (such as the SwitchBoard dataset conversation), it demands a carefully designed set of dialogue-act labels capable of catering to the requirements of counseling conversations. Drawing inspiration from foundational speech act theory ([113], [114]) while recognizing its limitations for this domain, we, in consultation with therapists and counseling experts, design a set of 12 dialog-act labels that are arranged in a hierarchy. These labels are designed to capture the intents of both the patient and therapist, and also be easily comprehensible to assist in the development of a conversational dialogue system. A high-level annotation hierarchy is shown in Figure 3.1. Each utterance in the dialogue belongs to one of the three categories⁷ – speaker initiative, speaker responsive, and general or mixed initiative. Our annotation scheme assigns three distinct dialogue-act labels to the first two categories, while the remaining four labels belong to the general category.

- **Speaker initiative labels:** When the speaker drives the conversation for the next few utterances.
 - **Information Request (IRQ):** This label is used as a request for some information, e.g., ‘Tell me your name.’.

⁷Sometimes, an utterance can have multiple dialog-acts; however, they are rare in the annotated dataset. Hence, for simplicity, we consider only one (primary) label for each utterance.

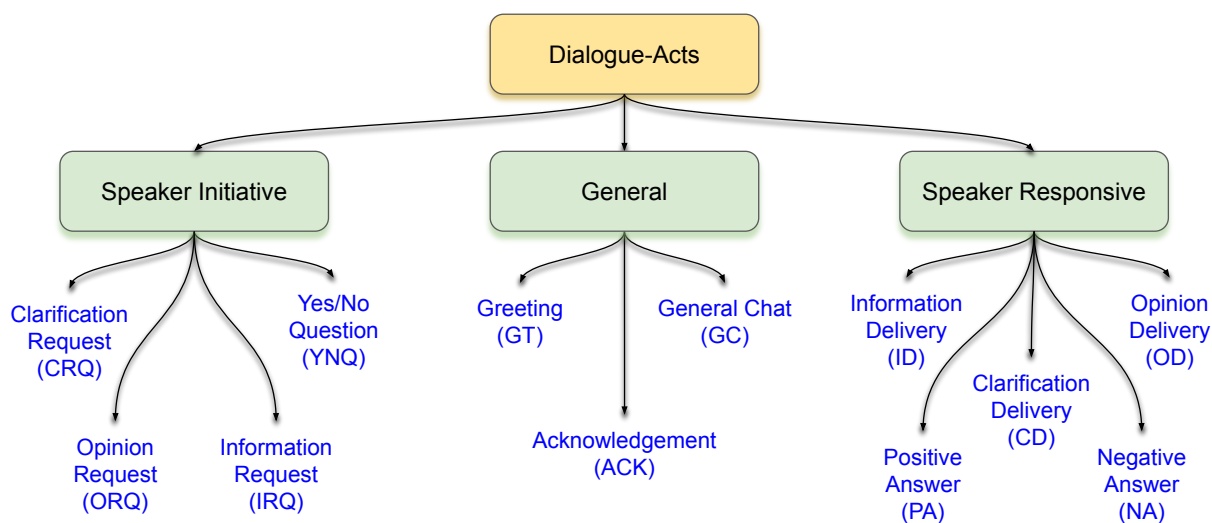


Figure 3.1: Annotation tree for HOPE.

- Yes/No Question (YNQ): The YNQ label is similar to IRQ; however, the expected response is a trivial *yes* or *no* answer. For example, the utterance, ‘*Did you complete your work?*’ shows how a query is raised with an expected answer of *yes* or *no*.
 - Clarification Request (CRQ): This label is assigned to those utterances in which a speaker usually asks the therapist for further clarification about topic that is being currently discussed. The distinction between IRQ and CRQ is the continuation of topics – IRQ is used whenever a discussion about a new topic or entity is started, and CRQ is used when the speaker wants to gather more information and delves deeper into the current topic at hand. For instance, if the therapist asks *You’re in a situation where there is alcohol?* and follows it with another utterance *And what sort of situations are you in?*, the later utterance is an example of clarification request as the therapist delves deeper to seek causes of distress for the patient.
 - Opinion Request (ORQ): The ORQ label is used when the speaker seeks opinions of the listener. For instance, the utterance *How does it feel to tell me about this?* is labelled as Opinion Request.
- **Speaker responsive labels:** The dialogue-act labels under this category are used in response to the speaker initiative labels.
 - Clarification Delivery (CD): This label is used when the speaker provides further clarifications about a topic/entity under discussion. For example, while discussing about charges that have been levied upon the patient, the therapist asks ‘*You mean get out of the charge?*’ and in response, the patient responds with the utterance *Yeah. I mean, I feel like they’re probably gonna make me like go to like rehab or something like that. And I don’t want to do that..* The patient response is an example of a clarification delivery.
 - Information Delivery (ID): When the speaker provides some factual objective information about herself. For instance, the utterances ‘*In my last job I worked in a paper factory.*’ and ‘*I do not like my present job, I am fed up with my boss.*’ are the examples of information delivery.
 - Positive Answer (PA): These labels are used when the utterance is an answer in the form of a simple *yes* to a question that was previously uttered, e.g., ‘*Are you alright*’ or ‘*Yes*’.
 - Negative Answer(NA): It is used when the utterance is an answer of the form of a simple *No* to a question asked earlier.
 - Opinion Delivery (OD): When the speaker explicitly lists out her/his opinions. For example, ‘*You don’t seem real certain.*’

- **General:** The utterances under this category can be uttered by either of the speakers under varying circumstances.
 - **Greeting (GT):** Each session usually starts with greeting by one speaker and an appropriate response from the other, e.g., ‘*Hello, how are you?*’ and ‘*I am fine, thank you.*’. We tag each of these utterances as GT.
 - **Acknowledgment (ACK):** In normal conversation, very often, we utter (e.g., ‘*Yeah! You are right.*’) to acknowledge the other speaker or to show our agreement without an explicit information request, question, or command. We also observe such cases in our collected dataset; hence, we tag them as ACK.
 - **General Chit-Chat (GC):** Other utterances that do not belong to any of the above labels are tagged as GC, possibly because of the vagueness and the lack of sense in the context of the conversation. For example, the utterance ‘*It’s a beautiful day today!*’ is an example of GC.

Annotation Process

We employed three annotators⁸ who are experts in linguistics.

To ensure the understanding of the tasks and annotation scheme, we took a sample of the dataset and asked each annotator to annotate them as per the prepared set of guidelines. Following this, every annotation was discussed in the presence of the annotators and an expert therapist as moderator to ensure consistency. After a couple of annotation and discussion rounds, the whole dataset was made available for the annotation. After the annotation process, we compute Cohen’s Kappa score (115) to measure the agreement among annotators. We obtain the inter-rater agreement score of 0.7234 – which falls under the ‘substantial’ (115) category.

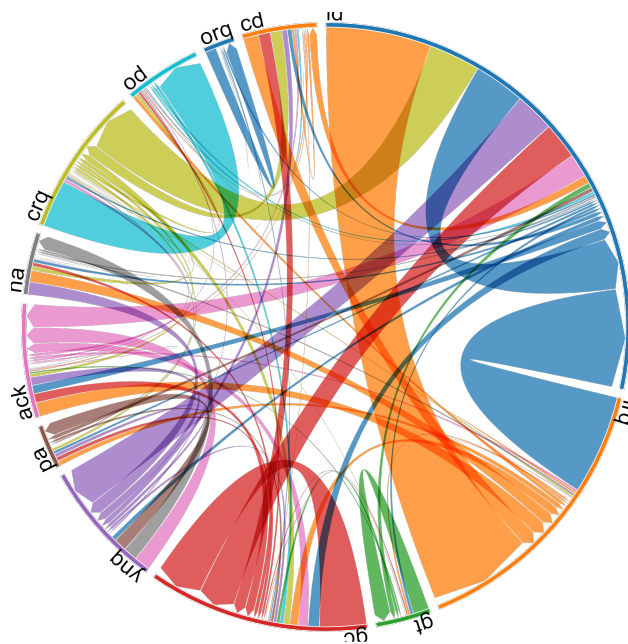


Figure 3.2: Relationship among DAC labels. The directed connection $U_t^x \rightarrow U_{t+1}^y$ shows the co-occurrence counts of two sequential utterances with x & y dialogue-acts, respectively.

Dataset Statistics

The broad definitions of speaker-initiative and speaker-responsive dialogue-act label pairs, IRQ & ID; ORQ & OD, and CRQ & CD, seem complementary in nature. However, the dataset does not support the above view entirely. To show the relationship between the respective speaker-initiative and speaker-responsive dialogue-act label pairs, we present a chord diagram in Figure 3.2

For any two consecutive utterances U_t^x and U_{t+1}^y with corresponding dialogue-act labels x and y , each

⁸Annotators are in the age group of 25-35, with 2-10 years of professional experience.

directed link ($U_t^x \rightarrow U_{t+1}^y$) between the two labels reflects their co-occurrence, and the strength of the link signifies their co-occurrence counts. Though a significant number of IRQ utterances are followed by ID utterance, in a few cases, they are followed by other dialogue-act utterances as well (e.g., ACK, GC, etc.). Similarly, ID utterances are not always preceded by IRQ utterances. We observe similar behaviour for YNQ & PA, NA, and CRQ & CD dialogue-act pairs as well. Table 7.1 provides the statistics of the HOPE dataset. In total, HOPE has transcripts for 12.9k utterances which are annotated with 12 dialogue-act labels. These utterances are evenly distributed between the patients and therapists with 6.38K and 6.47K utterances, respectively. We split the dataset into 70:20:10 ratio as the train, test, and validation sets, respectively.

In contrast to the regular patient-doctor conversations (e.g., SOAP), the dialogue sessions in HOPE are usually lengthy (~ 59 utterances per session). Moreover, the utterances in these sessions are themselves significantly longer as compared to other conversational datasets, with the average length of utterance of a patient being 103 words, whereas for therapist, it is ~ 124 words.

3.4 Proposed Methodology

We represent a therapy session as a conversation dialogue D consisting of a sequence of utterances $U[1 : m] = \langle U_1, U_2, \dots, U_m \rangle$, where m is the number of utterances in a dialogue. These utterances are uttered by the therapists and patients alternatively in a session. The objective of SPARTA is to assign a correct dialogue-act label y_t to every utterance U_t in the dialogue.

SPARTA is a transformer-based architecture that incorporates the speaker-aware contextual information for the dialogue-act classification. In our analysis of the HOPE dataset, we observed that a few of the dialogue-act labels are majorly associated with the patient, while a few others are related to the therapist. To model the speaker dynamics within a conversation dialogue, we consolidate a speaker-aware (SA) module in addition to the speaker-invariant (SI) module. The latter does not consider the distinction between a therapist and a patient utterance, while the former module distinguishes between the two through a pre-trained speaker identification module. Moreover, we also propose a novel time-aware attention (TAA) mechanism that considers the positions of contextual utterances during the attention computation. We hypothesize that the recent past contextual utterances have higher significance than the distant past utterances; hence, as opposed to the standard attention mechanism, TAA focuses more on the nearby (local) utterances.

For each utterance U_t in a conversation dialogue, we extract the semantic representation through pre-trained RoBERTa model (L16), which is subsequently utilized to leverage the local (L_t) and global (G_t) contextual information within the dialogue. We incorporate a sliding-window based dynamic memory unit to compute local context through a time-aware attention mechanism. In parallel, we employ a GRU layer to capture the dialogue history as global context. Our analysis reveals that a few dialogues contain utterances which discuss a topic (or an entity) that has occurred at the initial stage of the dialogue, and to correctly exploit the semantic of the utterance, the global information is desirable. We repeat the process of local and global context extraction for both speaker-aware and speaker-invariant setups. Finally, we combine these two representations with the residual connections for the classification. Figure 7.3 shows a high-level architecture diagram of SPARTA.

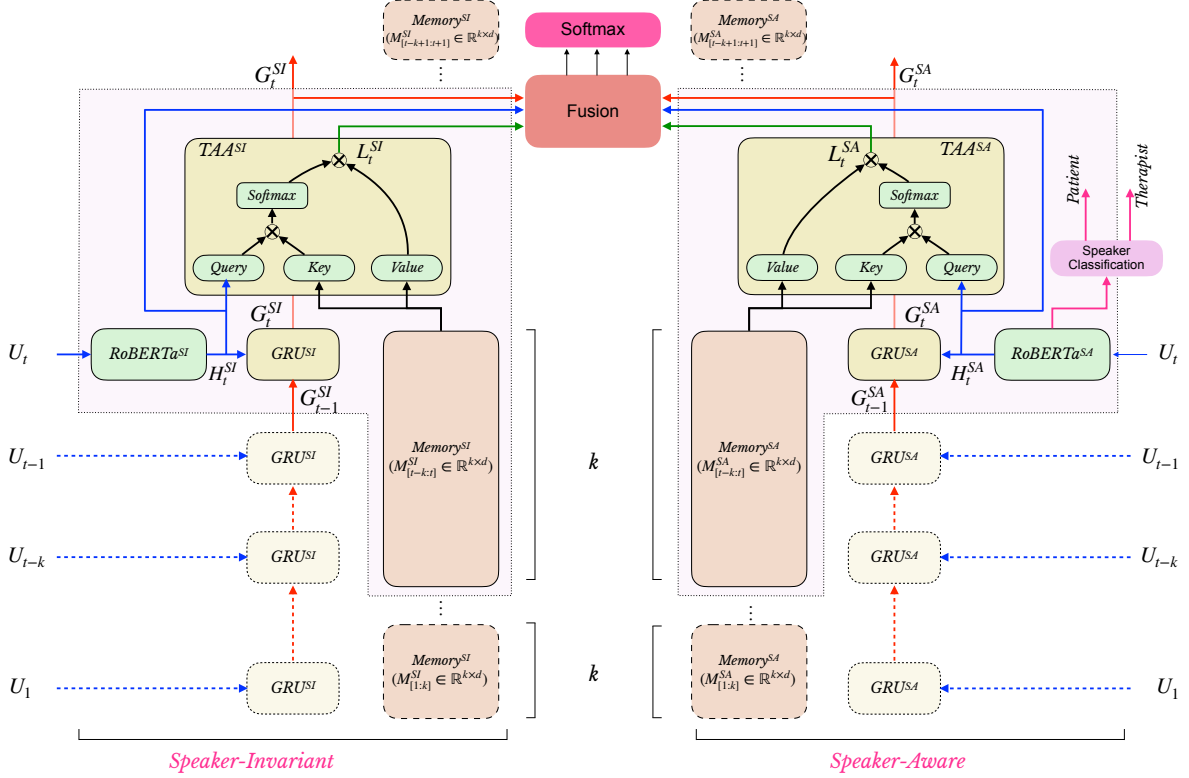


Figure 3.3: Architecture of SPARTA. For each utterance U_t , SPARTA computes the local-context L_t , through a time-aware attention (TAA) mechanism on the sliding-window memory unit and the current utterance. The dialogue-level global context is maintained using a GRU G_t . Finally, the speaker-aware and invariant local and global contexts are fused for the task.

Utterance Representations

As mentioned above, SPARTA maintains two separate modules for capturing the speaker-aware and speaker-invariant information. For speaker-invariant representations, we employ a pre-trained RoBERTa language model which is further fine-tuned on DAC task. The speaker aware module is a RoBERTa model fine-tuned on Speaker Classification task.

$$H_t^{SI} = \text{RoBERTa}^{SI}(U^t); H_t^{SA} = \text{RoBERTa}^{SA}(U^t)$$

Local Context and Time-Aware Attention

At every point in a dialogue, the nearby utterances provide important clues in the prediction of a dialogue-act label for the current utterance. For example, if the previous utterance is an information request (IRQ), then there is a good chance that the next label should be either information delivery (ID) or yes-no-answer (YNA). Therefore, we exploit the local context maintained in a memory $M_{[t-k:t]}$ for each utterance U_t in the dialogue, where k is the fixed local-context size.

$$M_{[t-k:t]}^{SA} = M_{[(t-1)-k:(t-1)]}^{SA} - M_{[k:k+1]}^{SA} + M_{[t-1:t]}^{SA} \oplus H_t^{SA} \quad (3.1)$$

$$M_{[t-k:t]}^{SI} = M_{[(t-1)-k:(t-1)]}^{SI} - M_{[k:k+1]}^{SI} + M_{[t-1:t]}^{SI} \oplus H_t^{SI} \quad (3.2)$$

We utilize TAA to learn the importance of contextual utterances based on their distance from the current position. At first, we pass the utterance representation (H_t), as computed by RoBERTa in the previous step, through a *tanh* activation layer to obtain the pooler output, and subsequently project it as the *query* ($q \in \mathbb{R}^{1 \times d}$) vector in the attention computation. On the other hand, the contextual memory $M_{[t-k:t]}$ is projected as the *key* ($K \in \mathbb{R}^{k \times d}$) and *value* ($V \in \mathbb{R}^{k \times d}$) matrices. Next, we encode the local context as follows:

$$L_t = \text{Softmax} \left(\frac{qK^T}{D} \right) V; D \in \mathbb{R}^k, D_i = \frac{1}{i}; \forall k : 1 \leq i \leq k$$

To extract the time-aware feature, we scale-down the dot product between the *query* and the *key* by a monotonically decreasing function of time. The inverse function was chosen based on its empirical advantage as shown in (117). The hypothesis for scaling-down the dot-product is due to the fact that as we move deeper into the dialogue history their influence on the dialogue-act reduces accordingly. Similar interaction dynamics was used in (118); but to our knowledge, we introduce an inverse function for the first time to compute the fixed window size attention for local context. Following the above procedure, we compute local contexts L_t^{SA} and L_t^{SI} for both speaker-aware and speaker-invariant modules, respectively.

Global Context

As the dialogue progresses, we maintain the global context of the dialogue through a GRU layer on top of the RoBERTa hidden representations.

Fusion and Final classification

Finally, we fuse the local and global contexts of speaker-aware and speaker-invariant modules for the final classification. We also add residual connections for better gradient flow during backpropagation. Our validation results supplement the choice of concatenation as the fusion operation to be better than other operations such as global max-pooling, global mean-pooling, etc.

3.5 Experiments, Results, and Analysis

In this section, we report our experimental results. We compared the SPARTA with several state-of-the-art models and present analyses.

Baselines

We choose the following existing systems as baselines. ► **CASA (79)**: It is a context-aware attention-based system for the dialogue-act classification. It uses RNNs at dialogue and utterance levels and computes context-aware self-attention before the final classification. ► **SA-CRF (81)**: This recent baseline incorporates a CRF layer for the classification. Moreover, it consolidates the speaker-change information using a Bi-LSTM encoder.

In addition to these recent baselines on dialogue-act classification, we also include other sequence-labeling classification systems. ► **DRNN (119)**: It is a novel Disconnected RNNs architecture that incorporates the position-invariant features for modeling. ► **ProSequo (120)**: It was proposed to efficiently

handle the short and long texts using dynamic recurrent projections and locality-sensitive projections (I121). ► **TextVDCNN (I122)**: This is a deep convolutional network with residual connections for text classification. The convolutional layer works at the character level, and k -max pooling is used to down-sample the output of convolutional layers for classification. ► **TextRNN (I123)**: This was the first work to integrate RNNs into the multi-task learning framework. We use the uniform layer architecture as described in the paper. ► **RoBERTa (I116)**: We use RoBERTa as a baseline in this work due to its superiority on various benchmarks. RoBERTa is similar to (I124).

Experimental Results

For the experiments, we randomly split the HOPE dataset into 70 : 20 : 10 ratio for the train, test, and validation sets. To measure the performances of SPARTA and other baseline systems, we compute macro-F1, weighted-F1, and accuracy scores. We implemented our system in PyTorch (I125) and utilized the pre-trained models from Huggingface Transformers library.

Model	Type of Modelling	Precision		Recall		F1		Accuracy
		Macro	Weighted	Macro	Weighted	Macro	Weighted	
TextVDCNN (I122)	U^t	11.01	21.02	19.53	38.53	13.37	36.81	41.77
ProSeqo (I120)	U^t	9.77	17.60	11.20	27.90	7.11	14.29	27.35
RoBERTa (I116)	U^t	51.01	58.12	47.14	52.97	43.97	49.13	52.97
TextRNN (I123)	$U^t + GC$	30.27	37.92	27.9	41.76	25.55	36.81	41.77
DRNN (I119)	$U^t + GC$	28.39	36.72	31.87	44.32	28.12	37.82	44.32
CASA (I106)	$U^t + GC$	59.78	62.56	51.22	58.46	51.65	55.95	58.46
SPARTA-BS	$U^t + GC$	58.94	62.31	52.02	57.70	51.83	54.98	57.70
SA-CRF (I81)	$U^t + LC + SA$	33.30	38.97	26.18	45.07	35.97	24.20	45.07
SPARTA-BS	$U^t + GC + SA$	58.87	63.02	53.28	58.41	52.22	55.57	58.41
SPARTA-MHA (3-fold CV)	$U^t + LC + GC + SA$	69.60	71.77	59.45	62.67	59.00	62.12	62.67
SPARTA-TAA (3-fold CV)		71.01	72.36	60.49	63.82	60.74	63.38	63.82
SPARTA-MHA	$U^t + LC + GC + SA$	60.24	66.53	59.64	63.45	58.16	63.26	63.45
SPARTA-TAA		62.15	67.36	61.13	64.75	60.29 [†]	64.53 [†]	64.75 [†]
Significance T-test [†] (p -value)						0.009	0.014	0.048

Table 3.3: Table showcasing the performance of baseline models as compared with our SPARTA model. U^t represents the use of Utterance representation, GC represents Global Context, LC represents Local Context and SA means the presence of Speaker Aware representations. The dagger symbol ([†]) represents statistically significant results compared to the best baseline, CASA.

We also present the label-wise performance of SPARTA in Table 3.4.

We can observe that SPARTA consistently yields good scores for the majority of the dialogue-acts, except for the *Acknowledgement* (ACK) where it records an F1-score of merely

47.86%. Even for the under-represented labels (ORQ, NA and PA) in HOPE, SPARTA reports good F1-scores of 59.09%, 64.38%, and 55.84%, respectively for these three labels.

	ACK	CD	CRQ	GC	GT	ID	IRQ	OD	NA	PA	ORQ	YNQ
Pre	44.08	89.91	50.72	87.38	60.24	63.76	70.95	53.19	58.75	57.33	48.5	61.36
Rec	52.34	48.04	49.53	54.38	65.79	81.02	77.81	44.64	71.21	54.43	76.47	57.86
F1	47.86	62.62	50.12	67.04	62.89	71.36	74.22	48.54	64.38	55.84	59.09	59.56

Table 3.4: Label-wise classification report for SPARTA-TAA.

Comparative Analysis

We compare SPARTA with various existing systems and other baselines. The comparative analysis is reported in Table 3.3. Based on the type of modeling, we categorize the baselines into three groups

Utterances	Gold	Prediction		
		CASA	SPARTA-T	SPARTA-M
Therapist: <i>Hi Angela, how are you doing today ?</i>	GT	GT	GT	GT
Patient: <i>I am all right, thanks</i>	GT	GT	GT	GT
...
Patient: <i>I mean, what can I say I lost my job. I'm seeing you because I have a drinking problem. So things are not that great.</i>	ID	ID	ID	CD
...
Therapist: <i>When you're drinking does it still feel good?</i>	IRQ	IRQ	IRQ	IRQ
Patient: <i>Nothing really feels good.</i>	ID	ID	ID	ID
Therapist: <i>Just gets back to normal?</i>	CRQ	CRQ	CRQ	CRQ
Patient: <i>Yeah, pretty much.</i>	CD	ID	CD	ID
Therapist: <i>When was the last time you drank ?</i>	IRQ	IRQ	IRQ	IRQ
Patient: <i>This morning.</i>	ID	ID	ID	GC
...
Patient: <i>Oh, I should stop. I mean, like, I'm not, not unrealistic.</i>	OD	ID	OD	ID
Therapist: <i>Yeah</i>	ACK	OD	ACK	ACK

Table 3.5: An example therapy session showcasing the differences in predictions between the baseline and SPARTA. SPARTA-T and SPARTA-M refer to the SPARTA-TAA and SPARTA-MHA models. We truncate the dialogue due to its length (total 151 utterances).

– utterance-driven (U_t), utterance+global context driven ($U_t + GC$), and utterance + global context + speaker-aware driven ($U_t + GC + SA$). Comparatively, SPARTA incorporates local context in addition to utterance, global context, and speaker dynamics ($U_t + LC + GC + SA$). In the first category, the standard *RoBERTa* model attains the best macro-F1, weighted-F1 and accuracy of 43.97, 49.13, and 52.97%, respectively. In comparison, CASA (79) yields the improved weighted-F1 and accuracy scores at 55.95% (+6.82%) and 58.46% (+5.49%), respectively, with the global context as an additional information. Finally, we experiment with SA-CRF (81) which also includes the speaker dynamics for the dialogue-act classification; however, its performance on HOPE is not at par with CASA (79). It reports 35.97, 24.20, and 45.07% macro-F1, weighted-F1, and accuracy, respectively. In comparison, SPARTA-TAA obtains significant improvements over all baselines. It reports improvements of +8.64%, +8.58%, and +6.29% in macro-F1 (60.29), weighted-F1 (64.53), and accuracy (64.75%), respectively, as compared to CASA suggesting the incorporation of local context extremely effective. Note that ProSeqo (120) and CASA (79) are currently the state-of-the-art on switchboard dialogue-act corpus benchmark⁹, yet they report inferior scores on HOPE compared to SPARTA. Moreover, we also report the mean of the 3-fold cross-validation results for both SPARTA-MHA and SPARTA-TAA, and the results are consistent with the train-val-test split case. We also perform a statistical significance T-test comparing SPARTA-TAA and the best performing baseline (CASA). We observe that our results are significant with > 95% confidence across macro-F1 (p -value= 0.009), weighted-F1 (p -value= 0.014), and accuracy (p -value= 0.048) values.

Error Analysis

In this section, we present two-way error analyses of SPARTA in terms of quantitative and qualitative evaluations. *Quantitative analysis:* We report the confusion matrix for SPARTA-TAA in Figure 3.4. We observe three pairs with significant error rates ($\geq 25\%$) – YNQ:IRQ (26%), OD:ID (43%), ID:ACK

⁹<https://paperswithcode.com/sota/dialogue-act-classification-on-switchboard>

(28%). For the prediction of *information delivery* (ID), SPARTA is confused most of the time with other classes – 19% with PA, 20% with NA, 43% with OD, 13% with GT, 17% with GC, 14% with CRQ, 28% with ACK, and 22% with CD.. We can relate this behavior to the diversity of the utterances with ID tag, i.e., the discussion in these utterances generally contains a fair segment of utterances from other dialogue-acts (e.g., ‘*Yeah, That’s something I always do.*’ could be easily be confused with PA). The other prominent error case is found in IRQ:YNQ pair, we observe a confusion of 26% between IRQ and YNQ utterances because of the versatile questioning behavior. For the remaining cases, error rates are nominal. Thus, we articulate that SPARTA can be further improved with a more balanced dataset.

Qualitative analysis: Table 3.5 shows a sample session along with actual and the predicted dialogue-act labels for SPARTA and the best baseline model (i.e., CASA). Due to the length of the conversation, we truncate some of the utterances in between; however, the gist of the conversation is that *the patient is stressed of losing her job and having drinking issues and the therapist is trying to understand the core problem.* The conversation has mostly *information request* and *information delivery* types of utterances with a few other dialogue-acts in between (e.g., *CRQ GT*, etc.).

We can observe that for the first three utterances, SPARTA and the baseline are consistent with the actual labels. For the fourth utterance, the SPARTA-MHA model misclassifies the utterance as CD when the patient is clearly providing objective information that is necessary for further conversation. In the seventh utterance, the therapist wants more clarification about the drinking habits, and the patient provides the clarification but CASA and SPARTA-TAA models wrongly classify this utterance as ID when there is no objective information being provided here. Next, we notice that when the patient talks about stopping this and provides her opinion that she is *not unrealistic*, the SPARTA-MHA and CASA models predict the wrong label for this utterance as ID. In the next utterance, the therapist acknowledges this opinion, but the CASA model predicts the wrong label. So, not only is the SPARTA-TAA model able to capture the semantics of the utterances better, it also utilizes the contextual information in a better way by relating the past information about the speaker with the current utterances.

General Discussion: The work presented in this paper was motivated solely by the dire need of understanding conversations that occur in counseling sessions and to design solutions that would help the therapists to better understand the intents of their patients. However, the proposed model can be easily adaptable to other domain (such as normal chit-chat-based conversation) as well. To ensure we do not deviate from the prime objective of this work, we restrict ourselves to explore the dialog-act classification in the counseling conversations only.

ACK	0.52	0.00	0.02	0.02	0.05	0.28	0.05	0.00	0.00	0.05	0.01	0.00
CD	0.16	0.46	0.01	0.01	0.00	0.22	0.00	0.00	0.04	0.03	0.02	0.01
CRQ	0.00	0.00	0.50	0.03	0.01	0.14	0.16	0.03	0.00	0.01	0.01	0.11
GC	0.05	0.00	0.07	0.54	0.04	0.17	0.07	0.01	0.01	0.02	0.00	0.02
GT	0.12	0.00	0.01	0.00	0.66	0.13	0.01	0.01	0.00	0.01	0.04	0.00
ID	0.02	0.01	0.05	0.01	0.01	0.81	0.02	0.00	0.04	0.00	0.01	0.01
IRQ	0.00	0.01	0.07	0.01	0.01	0.03	0.78	0.02	0.00	0.02	0.02	0.04
OD	0.02	0.00	0.04	0.04	0.00	0.43	0.02	0.45	0.00	0.02	0.00	0.00
NA	0.00	0.00	0.03	0.02	0.02	0.20	0.02	0.00	0.71	0.00	0.02	0.00
PA	0.22	0.00	0.00	0.01	0.00	0.19	0.01	0.01	0.00	0.54	0.01	0.00
CRQ	0.00	0.00	0.03	0.00	0.00	0.03	0.15	0.00	0.00	0.03	0.76	0.00
YNQ	0.00	0.01	0.11	0.00	0.01	0.03	0.26	0.01	0.00	0.00	0.01	0.58
	ACK	CD	CRQ	GC	GT	ID	IRQ	OD	NA	PA	CRQ	YNQ

Figure 3.4: Confusion matrix for SPARTA-TAA.

3.6 Limitations

We aim to tackle a very sensitive and pervasive public health crisis. We transcribe the data from publicly available counseling videos. The automatic transfer of utterance from the speech modality to the text causes some information loss, though we tried our best to recover them through manual intervention. Moreover, we consulted with mental-health professionals and linguists in preparing the annotation guideline. However, the annotator’s bias cannot be ruled out completely. The names of the patients and therapists involved in these sessions have been systematically masked. Another important aspect of the current work is that the majority of the sessions in HOPE belong to mental health professionals and patients based in the United States. Hence, the effectiveness of SPARTA on data from other geographical or demographical regions may vary. We understand that building computational models in mental-health avenues has high stakes associated with it and ethical considerations, therefore, become necessary. No technology will work perfectly in solving the problems related to mental health (126). It is important to note that we do not make any diagnostic claims. Further, the deployment of any such technology will be done keeping in mind the safety-risks and mitigating any sources of bias that may arise.

3.7 Summary

Paying heed to the consequences of the COVID-19 pandemic on mental health, in this paper, we raised the attention on the much deserved research on dialogue system for mental-health counselling. To this end, we collected and developed the HOPE dataset for the dialogue-act classification in dyadic counselling conversations. We defined twelve dialogue-act labels to cater to the requirement of counselling sessions. In total, we annotated $\sim 12.9k$ utterances across 212 sessions. We also proposed SPARTA, a novel transformer-based speaker and time-aware joint contextual learning model for dialogue-act classification. SPARTA utilizes the global and local context in the speaker-aware and speaker-invariant setups while also using a novel memory-driven time-aware attention mechanism to leverage the local context. Our extensive ablation study and the comparative analysis established the superiority of SPARTA over several existing models. In future, we would like to extend our effort in the development of dialogue-systems for mental-health counselling by including other crucial tasks such as emotion recognition, dialogue summary generation, dialogue-state tracking, empathetic response generation, etc. Another dimension of the future work is to include other languages and demographic diversities to cater the requirements of a larger population.

Part II

Counseling Summarization

4. Domain Knowledge for Counseling

Summarization

The psychotherapy intervention technique is a multifaceted conversation between a therapist and a patient. Unlike general clinical discussions, psychotherapy’s core components (viz. symptoms) are hard to distinguish, thus becoming a complex problem to summarize later. A structured counseling conversation may contain discussions about symptoms, history of mental health issues, or the discovery of the patient’s behavior. It may also contain discussion filler words irrelevant to a clinical summary. We refer to these elements of structured psychotherapy as *counseling components*. In this paper, the aim is mental health counseling summarization to build upon domain knowledge and to help clinicians quickly glean meaning. We create a new dataset after annotating 12.9K utterances of counseling components and reference summaries for each dialogue. Further, we propose ConSum, a novel counseling-component guided summarization model. ConSum undergoes three independent modules. First, to assess the presence of depressive symptoms, it filters utterances utilizing the *Patient Health Questionnaire* (PHQ-9), while the second and third modules aim to classify counseling components. At last, we propose a *problem-specific Mental Health Information Capture* (MHIC) evaluation metric for counseling summaries. Our comparative study shows that we improve on performance and generate cohesive, semantic, and coherent summaries. We comprehensively analyze the generated summaries to investigate the capturing of psychotherapy elements. Human and clinical evaluations on the summary show that ConSum generates a high quality summary. Further, mental health experts validate the clinical acceptability of the ConSum. Lastly, we discuss the uniqueness in mental health counseling summarization in the real world and show evidences of its deployment on an online application with the support of *mpathic.ai*.

4.1 Introduction

Mental health counseling is one of the front-line defenses against mental health illness. In medical and primary care settings, the doctor follows a highly-structured assessment approach that elicits specific information about the patient’s medical problems to rule out different diagnoses. In psychotherapy counseling sessions, patients take center stage in elucidating their situation with subtle details. The therapist introduces diverse auxiliary context in conversations to put the patient at ease, discuss events happening in the patient’s recent past including the feelings, reflections and emotions that the patient experiences, and other relevant topics. Follow-up conversations with the patient are also vital for a successful treatment. The points of the counseling session that are crucial for continuity of care in follow-up and treatment planning include: (a) the patient’s presenting problem, (b) symptoms and diagnosis, (c) treatments (current and prior), (d) mental status and risk assessment, and (e) other varied context and history. An efficient way to present this information to a therapist is to summarize the therapist-patient conversation. The process highlighted earlier, differentiates a mental health counseling session from other medical conversations, and extraction of relevant utterances is key to avoid noise in generated summaries.

In this paper, we aim for the task of counseling conversation summarization. Our work incorporates domain knowledge can be used to generate summaries that inherit essential psychotherapy elements. Earlier studies ([127; 128]) show attention based architectures to generate summary. Recent studies ([129; 130]) use segmentation of the conversation based on topic similarities in summarizing dialogue. Extending the ideas by incorporating mental health domain knowledge, we propose ‘counseling components’ for each utterance to include the understanding associated with counseling conversation. Comprehensive analyses of counseling conversations show that the conversation aims to discuss the reasons for the problem, discover the patient’s insights, and reflect on their past. Hence, we create the dataset, MEMO¹ where we annotate 12.9K utterances with the following four labels - *symptom and history*, *patient discovery*, *reflecting*, *discussion filler*. Also, we annotate and validate the dataset with the help of mental health experts.

The state-of-the-art models are capable of generating semantically rich text; however, to pick the fitting symptoms and essence of the counseling problem is still a challenge. In this paper, we incorporated mental health domain knowledge using the Patient Health Questionnaire (PHQ-9) ([131]) to represent an established set of questions to assess the patient’s mental health symptoms. The PHQ-9 specifically assesses depressive symptoms but represents an example of symptom-specific domain knowledge. This approach could be generalized to other frequently-used assessments focusing on other symptom profiles like anxiety (e.g., GAD-7). Taking advantage of domain knowledge, we identify utterances categorized as discussion filler or irrelevant and will not be prioritized. Labeling across different counseling conversations aids the model to map similarities and serve as the supervision signal to attend to salient utterances and extract appropriate information to be a part of the summaries.

We propose ConSum² a mental health counseling summarization architecture. The complete pipeline of the model works in three different modules: (a) The first module (MH-Know) exploits domain knowledge to filter utterances irrelevant to the counseling summarization task. (b) Secondly, we propose a discussion filler classification module that filters irrelevant utterances like ‘yeah’, ‘ummm’, etc. (c) The third module predicts the counseling components to utilize the structured therapy knowledge for summarization task.

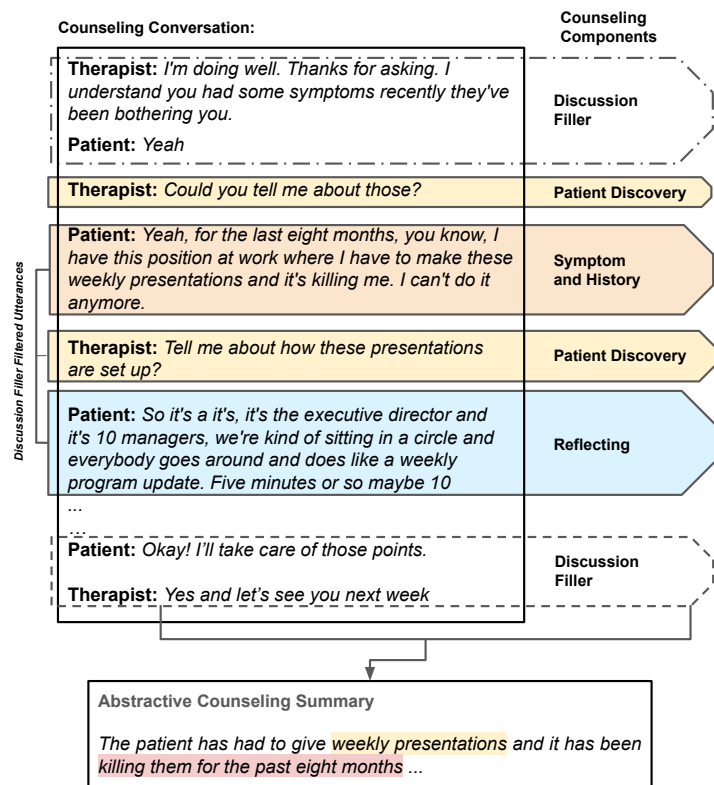


Figure 4.1: A sample counseling session instance from the MEMO dataset. *Symptom and History*, *Patient Discovery*, *Reflecting*, and *Discussion Filler* are psychotherapy elements. The summaries pertaining to this truncated snippet of conversation is shown at the bottom. Note: The summary and conversation are truncated for brevity.

¹MEMO: Mental hEalth suMmarizatOn dataset

²ConSum: Counseling Summarization

We run our experiments on the MEMO dataset. Evaluations show improvements across all baselines and discuss the effect of counseling components and effective use of domain knowledge in our ConSum model. Finally, we discuss the uniqueness of counseling summarization and explore the generalizability of ConSum. Significant contributions of our work are as follows:

- We propose ConSum, a summarization model that exploits mental health domain knowledge and counseling components.
- We propose a novel counseling summarization dataset - MEMO and present a novel annotation scheme for psychotherapy elements viz. symptom and history, patient discovery, reflecting aspects, and discussion filler in utterances of counseling dialogue.
- We propose a new problem specific metric to evaluate summaries i.e., **Mental Health Information Capture (MHIC)** metric which reasonably evaluates summaries that are most useful from a counseling’s perspective.
- Mental health experts in *mpathic.ai* meticulously analyzed MEMO. They studied the results on various clinical and linguistic parameters to evaluate the acceptability of ConSum’s performance for commercial and clinical application at scale.

4.2 Related Work

Given the critical nature of the medical domain, the pace of adoption of modern deep-learning models did not gain enough traction for impactful change. Instead, it was used to address mundane medical procedures. On the other hand, summarization is a long-studied problem in text processing. Earlier, most of the focus and improvements were on extractive summarization, with a drastic shift to abstractive summarization recently. Some of the work done in the medical health domain is discussed below, followed by a section on technical progress in summarization methods.

4.2.1 Mental Health and NLP

One of the early works by (I32) explored a simple information retrieval system to fetch information from medical documents using document summaries. They employed medical ontology to help fine-tune the user query; however, simple summarization methods limited its effectiveness. Later, an NLP-based framework was introduced by (I33) to identify a personnel’s physical exposure (in a battle) and their emotional reaction to it. Their work highlighted the importance of subtle indicators of mental illness and the advantages of diagnosing ‘early categories’, which in their case are a hand-crafted, frequent-word-based ontology related to combat exposure. (I34) analyzed clinical forms in mental health to cut down on the arduous task of manual analysis using machine learning. Their work paved the way to explore machine learning as a tool to automate tasks dependably, to a certain extent.

Secondly, certain efforts helped address mental health problems specifically. (I35) showed a comprehensive study about the management of adults with Major Depressive Disorder (MDD). Their analysis identifies different stages of depression and discusses an appropriate level of antidepressant to choose based on the patient’s age, anxiety levels, and long episode duration. Subsequently, the prescribed antidepressants are monitored for satisfactory patient response. This notion motivates the identification of key indicators of the patient’s conversation with the therapist and follow-up sessions that rely on the patient’s past interactions (and prescriptions). (I36) experimented with neural network models to predict the mental illness condition from neuropsychiatric notes. These notes comprised 300 words on

average about the patient’s present illness and events associated with it, followed by a psychiatric review system that mentions mental illness associated with the patient. Despite using various models such as RNN, BiRNN, and LSTM, it is a straightforward classification problem among the 13 predefined mental illnesses.

In deep learning based summarization models, extractive summarization (I37) addressed the problem of doctors going through elaborate discharge diagnoses by summarizing them using the BERT model. They used character-level tokens to reduce the parameter size of the BERT model to deploy in low-resource setups. However, there were no sizeable efforts made in explicitly capturing medical information in summaries.

Some of the contemporary approaches include the work by (I38) where they generated artificial datasets from genuine mental health records. Besides, (I29) focused on identifying PICO (Patient/Problem, Intervention, Comparison, and Outcome) on a well-structured sequence of sentences from medical documents to present a concise extractive summary only. Recently, to filter relevant utterances, (I39) used domain knowledge from patient health care questionnaire (PHQ-9) to build knowledge graphs. They used an unsupervised approach to build abstractive summaries for the counseling session. (I40) showed the generation of summary in two steps, one for building chunks of partial summaries and then second for fusing those chunks to generate final summary. Moreover, (I41) showed an online application developed based on medical domain-specific annotation and information extraction.

4.2.2 Dialogue Summarization

Broadly, the two categories of dialogue summarization tasks are extractive (utterance filtering) and abstractive (semantically rich). Summarization has been long studied under various settings (I42; I43). Earlier, (I44) proposed a graph-based approach for extractive summarization wherein topical similarities between each sentence are identified using Probabilistic Latent Semantic Analysis (PLSA). (I45) highlighted two broad extractive summarization techniques, which are *superficial techniques* including statistical-based approach and *knowledge-based* methods for pre-training from a large corpora.

An observation on conversation summaries reveals that capturing salient points with a coherent narration (abstractive) is the most succinct format to express. Hence, abstractive summarization gained traction with early work by (I28) exploring encoder-decoder architecture which was further extended with suitable attention mechanism by (I46) for abstractive summarization. Then, a mix of extractive and abstractive summarization with a suitable copy model was introduced by (I27). They used a generator model for abstractive summarization, a copy model for extractive summarization, and a coverage vector to limit repetition in summarization. In the sequence, (I47) proposed a reinforcement learning-based approach to extract the salient sentences and concised them to summary sentences with an abstractive network. They employed a specific emphasis to reduce redundancy in the extracted utterances from the conversation.

Further nuanced observation of the relation between salient sentence extraction and conversation utterances shows the dependence on certain groups of utterances. On this basis, recent work by (I48) introduced topic distribution based on Latent Dirichlet Allocation (I49). Extending on the idea of topic segregation, (I30) used three labels to tag each utterance with - Problem Description, Diagnosis, and Other. They showed that adding utterance labels aids in summarization performance. Nevertheless, since their reference summaries are extractive, the complexities of their models are limited. Further, (I50) and (I51) pushed the boundaries of abstractive summarization by generating semantically rich sentences, diversely adapted to different generative tasks. In the clinical conversation domain, (I52) and (I53) followed the approach of important utterance selection for medical conversation summarization.

4.3 Dataset

We create a novel dataset for the task of counseling conversation summarization in which we extend data collected from the publicly available counseling conversation dataset – HOPE (1). The dataset contains 12.9K utterances from 212 counseling conversations between therapist and patient. The dataset is gathered from multiple counseling videos on publicly available platforms like YouTube. It belongs to diverse demographic groups with distinct mental health matters and several therapists, helping researchers propose a generalized approach. Further, the authors extracted transcriptions from the videos and pre-processed them. Collected dialogues are dyadic, in which patients and therapists are the only interlocutors. Since our task aims for counseling summarization built upon domain knowledge, we extend the HOPE dataset to annotate psychotherapy elements and counseling summary. We consulted with a team of the leading mental health experts and proposed a hierarchical labeling structure for each utterance. We discuss the details related to annotation in Section 4.3.1. The same experts helped us develop annotation guidelines and validate the annotations. We now discuss the extended dataset, which we call MEMO.

Next, the task is to add psychotherapy elements to these dyadic conversations. We observe that the conversations can contain three essential counseling components, namely *symptom and history*, *patient discovery*, and *reflecting* utterances as shown in Figure 4.2. The rest of the conversation between therapist and patient falls under the category of *discussion filler*. Furthermore, to ensure the reference summaries represent rich knowledge considering the clinical and linguistic perspectives and contained psychotherapy elements, we follow the annotation guidelines designed by the team of mental health experts.

4.3.1 Annotation Process

Counseling conversations are typically challenging as the patients are reluctant to express themselves in front of therapists. As a consequence, the therapist articulately involves the patient into the discussion. In support of this, we meticulously analyze counseling conversations to symptom and history of the mental health problem, the discovery of the patient’s behavior, and insights into the past story reflecting on the patient’s current situation. The rest of the counseling conversation is generally enclosed with discussion filler phrases. To better understand the counseling and support summarization task, focusing only on a subset of knowledge-enhanced utterances, we annotate counseling components and discussion filler for all utterances. With the help of a comprehensive annotation process in place, the techniques used by therapists are highlighted with each utterance. Therapists can be significantly helped with AI-based tools adhering to such domain-specific annotation guidelines³. With the help of guidelines, the annotations keep track of the therapy techniques; therefore, the summaries generated are succinct in their information.

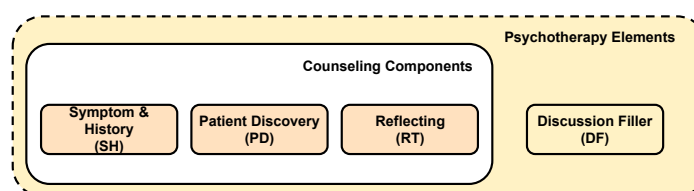


Figure 4.2: Hierarchical labeling for each utterance. Counseling components including *SH*, *PD* and *RT* essentially contribute in summary generation.

Psychotherapy elements: Mental health therapy sessions are composed of counseling components that effectively contribute towards successful interventions and discussion filler which barely add any relevance to the summary generation. We labeled utterances with four fine-grained labels:

- Symptom and History (SH): It captures the utterances that comprise the most insightful information

³<https://familytherapybasics.com/blog/therapy-case-summary>

for the therapist to assess the patient’s situation. A few examples are mentioned below:

Speaker	Examples
Patient	<i>“Ever since I moved into the apartment I have been feeling anxious”</i>
Patient	<i>’This has been the first time I have experienced mood swings”</i>
Patient	<i>“I would be so stressed afterward. And worried I would make another mistake”</i>

- **Patient Discovery (PD):** The patients coming in for a counseling session arrive with complex thoughts. The therapist tries to build therapeutic relationships to calm patients to unravel their thoughts. Examples are as follows:

Speaker	Examples
Therapist	<i>“Don’t worry if it lasts for several days”</i>
Therapist	<i>“I believe sleep is not an issue for you anymore now that you are getting eight hours of sleep everyday”</i>
Therapist	<i>“Any drug or alcohol use?”, “Has it ever happened to you?”</i>

- **Reflecting (RT):** Therapist utterances are concise most of the time, which is to ensure there is sufficient space for the patient to express themselves. Patients are encouraged to share their stories and events of their lives. On the other hand, the therapist understands the patient by giving them an imaginary situation to assess their actions. Some utterances are shown below:

Speaker	Examples
Therapist	<i>“It’s been bothering you for a few weeks, something new, something a little frightening for you . . . ”</i>
Patient	<i>“By the time it it’s my turn to present I start. I just, I’m almost like just . . . ”</i>

- **Discussion Filler (DF):** When a therapist and a patient engage in conversation, they articulate some utterances that are peripheral to the session. In this work, we tag them as discussion filler. These include *pleasantries* (‘Good morning!’), *non-lexical fillers* (‘Ummm’), *acknowledgments* (‘Right’), and *restatement of affirmations in subsequent utterances* (‘Yeah. Yeah’), etc. Discussion Filler utterances carry little to no relevance in the summary generation.

4.3.2 Data Analysis

Table [7.1](#) shows data analysis on MEMO. The dataset consists of an almost equal number of patient and therapist utterances. Therapist utterances are tagged more with discussion filler labels showing that therapists converse to put the patients at ease and agree with them. PD is the most prominent label in the dataset. Patients tend to talk about several auxiliary topics, not their mental health issue, when asked to share their experiences, whereas RT is the least tagged label.

4.3.3 Ethical Considerations

Three domain experts reviewed the annotations at *mpathic.ai*, a seed-stage corporation specializing in conversation analysis in commercial applications, including evaluation of psychotherapy, coaching, and customer services. All experts are independently licensed in clinical practice and have backgrounds in automating fidelity of therapy at scale. They included a board-certified, licensed psychologist (SPL) with over 15 years of experience in applying machine learning to psychotherapy evaluation, a licensed clinical social worker specializing in the evaluation of human labeling for machine learning, and doctoral-level counselor, director of Clinical AI at *mpathic.ai*.

Following Sekhon et al.’s framework (154), they reviewed the summaries for clinical acceptability in the areas of affective attitude, burden (i.e., the cognitive load of interpreting the summaries), ethicality, coherence (i.e., how well the summaries were understood), opportunity costs (i.e., pros and cons of utilizing the summaries), and perceived effectiveness (i.e., how well they might perform in a clinical setting). The acceptability parameters were independently evaluated and given a ranking of 0-2 on each parameter with 0 being not acceptable, 1, being acceptable with modification, and 2 being acceptable for commercialization. After independent ranking, discrepancies were discussed, and an average gold-standard acceptability the ranking was determined for each summary. Additionally, the experts provided feedback for commercialization considerations in applications like electronic health care records and quality assurance.

4.4 Methodology

In this section, we describe our proposed method, called ConSum. It encodes dialogue utterances and uses three modules in the pipeline to extract complementary information. These three modules operate independently on the selection and filtering of utterances. We use mental health’s Patient Health Questionnaire (PHQ-9) (131) knowledge similarity (MH-Know) and discussion filler classification to filter utterances from each counseling dialogue. Further, we classify utterances from dialogue for counseling component labels and use these label information to generate abstractive summaries. Figure 4.3 shows a high-level architecture of ConSum.

Consider a dialogue containing n utterances, $D = \langle u_1, u_2, \dots, u_n \rangle$. We fine-tune DistilBERT (155) embeddings to create utterance representation of d dimension. With knowledge-infused utterance selection (MH-Know) module and discussion filler classifier, we generate binary mask arrays from each module, which are fused to filter the utterances. On the other hand, the classification from the counseling components classifier augments each utterance with regard to mental health aspects. We describe each module below.

- **Discussion Filler Classifier (DFC):** Discussion filler classification is a binary classification task. We use a feed-forward network with 2 hidden layers to classify each utterance u_i into ‘counseling’ or ‘discussion filler’ from the obtained input representations. It compresses representations to 100 dimensional hidden representation and then to 2 at the output layer. A dropout of 30% is applied between the linear layers to regularize the model’s performance. The model is trained with adam optimizer to minimize the cross-entropy loss. The output is a mask array labeled τ , where τ_i represents mask array from dialogue D_i . Using independent networks to classify the discussion filler label and the counseling components helps smooth errors better than having a single unified classifier.
- **Mental Health Knowledge Infused Utterance Selection (MH-Know):** Earlier efforts in mental health counseling summarization paid less attention to the domain knowledge. Patient Health

Questionnaire (PHQ-9) represents an established set of questionings to assess the patient’s mental health condition (131). We use PHQ-9 lexicons (156) to compute similarity between nine questions and input utterance. The intuition is to obtain the most relevant utterances corresponding to knowledge infused questions. We use BERTScore (157) to compute the similarity between the utterances u_i and the PHQ-9 lexicons $PHQ = \langle phq_1, phq_2, \dots, phq_9 \rangle$. The nine similarity scores, $s = \langle s_1, s_2, \dots, s_9 \rangle$ corresponding to each utterance ranges between 0 to 1 each. The summation of the similarity score, denoted by ψ , is compared with the hyperparameter-tuned value of the threshold, denoted by ϕ .

$$s_i = bertscore(u_i, phq_i) \Rightarrow \psi_i = \sum_{m=1}^{m=9} s_m$$

The value of ϕ is determined by exploring a range of candidate values $\in [1, 9]$ and selecting the one that best balances retention of domain knowledge and intervention components with minimal impact on the performance of other pipelines. This creates a mask-array, $\sigma_i = \langle \sigma_1, \sigma_2 \dots \sigma_n \rangle$ containing 1 for cases where intervention similarity score is less than the threshold and 0 otherwise. Here, i in σ_i represents i^{th} utterance.

$$\sigma_i = \begin{cases} 1, & \text{if } \psi_i \leq \phi. \\ 0, & \text{otherwise} \end{cases}$$

ConSum prioritizes the utterances fetched from the MH-Know module along with corresponding counseling tags to generate summary. During experimentation, we use the threshold value of 6 to extract utterances relevant to the task.

- **Counseling Components Classifier (CCC):** As discussed earlier, the four components are SH, RT, PD, and DF (c.f. Section 4.3.1). Considering SH, RT and PD as the key components contributing to the summary generation, we frame the problem as a four-class classification task. Hence, for each utterance representation of dimension d , we first fetch the ‘*counseling context*’ across utterances. As the counseling progresses, we maintain the context of the dialogue through a GRU layer on top of the utterance representations. Further, we apply multi-head self-attention following (146). Representations fused with attention weights are passed through two dense layers to learn hidden representations from 100 dimensional and then further to 3. We apply a softmax classifier to predict probabilities corresponding to each class label.

Identification of discussion filler is a domain-independent binary classification task. Therefore, rather than following an end-to-end model for psychotherapy element classification, we opt for a hierarchical model to predict discussion filler and counseling components. In our case, the standard model performs well for DFC. However, we sense the requirement for a model that can learn contextual counseling labels in the summary generation process. The remaining section shows the direct use of counseling labels in the summary generation module.

- **Summary decoder:** We obtain two mask-arrays from the MH-Know module and DFC. Discussion Filler mask τ and knowledge-infused mask σ are merged as follows. F denotes the resulting mask array. Further, utterances U_i from the dialogue D_i are concatenated with predicted counseling components (c.f. Section 4.4) and filtered using the final mask array F . The final subset of utterances with domain knowledge and counseling components (CL_i) are concatenated, resulting in G . It is used as an input to fine-tune the pre-trained T5 model (151) for abstractive summary generation.

$$F = (\sigma \cap \tau)'; G = (U_i \oplus CL_i) \otimes F$$

4.5 Experiments and Results

This section reports our experimental results, comparative study, and other analyses. To evaluate the summaries, we use ROUGE (158) metrics, namely Rouge-1 (R-1), Rouge-2 (R-2), and Rouge-L (R-L). In addition, we show the QuestEval (QAE) score and Bleurt Score (BS) to capture the contextual advantage in abstractive summarization. QuestEval (159) comprehensively judges the reference and predictions on four aspects – consistency, coherence, fluency, and relevance by generating question-answer pairs on the given source document, reference, and generated summaries. Bleurt (160) utilizes pre-training on the BERT model using unsupervised techniques with millions of synthetic examples to generalize all possible data distributions. At last, we show human and clinical evaluations to validate the quality of generated summaries.

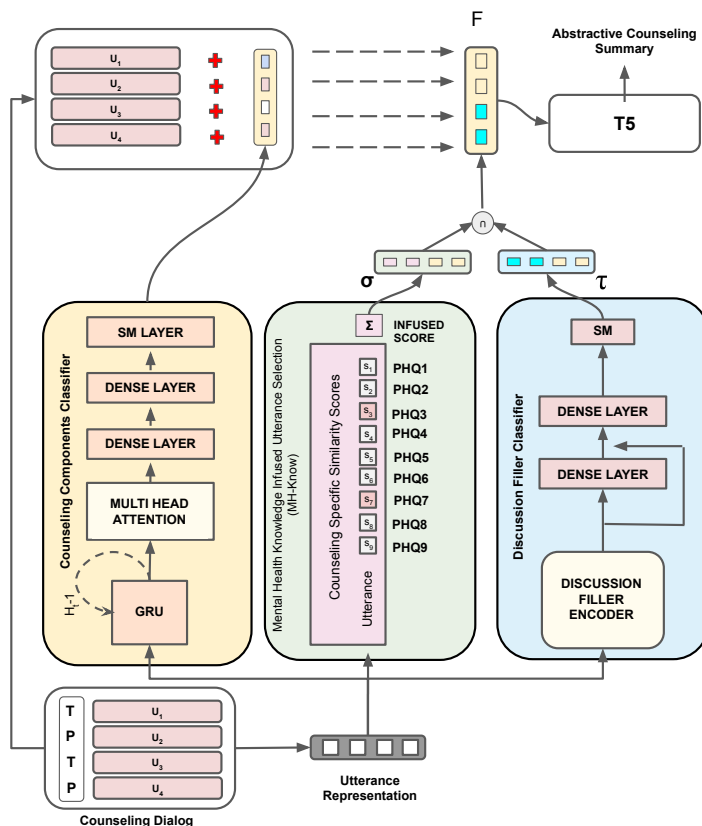


Figure 4.3: Model diagram of ConSum. The Utterance embeddings are reused across three modules in the pipeline.

4.5.1 Baselines

We choose the following systems as our baselines - **Pretrained Language Model (PLM)**: (161) used DialogPT to segment topics and then generated summaries with BART. **RankAE**: (162) used BERT to group utterances into segments and further generated summaries using denoising auto-encoder (163). **Segmented Modeling (SM)**: (164) leveraged dialogue act information on standard BiLSTM (128) architecture. **Transfer Learning with unified Text-to-Text Transformer (T5)**: (151) used a shared framework on transformer architecture to pre-train on huge custom crawled corpus (C4). **Pegasus**: (150) introduced a novel pre-training objective of using gap-sentences-generation and then generated summaries.

4.5.2 Results and Ablation Study

The experimental results on MEMO are shown in Table 4.1. ConSum performs better than all other baselines with the highest R-1 score of 45.36 and R-2 score of 15.71 – significant improvements of 11.12 and 4.52 points over the best baseline, respectively. The PLM model turns out to be the second ranked model in both cases with an R-1 score of 34.24 and an R-2 score of 11.19. It shows the importance of topic segmentation in a conversational dataset, which aids in the summarization outputs. In contrast, PLM scores the highest R-L score of 33.35; however, its summaries have a higher structural similarity instead of contextual similarity between its output and the reference summaries. On the other hand, ConSum produces a dominant performance on Bleurt score with a difference of 0.90 points over the best

baseline (-0.5655 of T5 vs 0.3407 of ConSum). This score indicates contextually superior performance concerning the reference summary.

We also show the effect of counseling components and different modules in the pipeline and analyze the results. We demonstrate the importance of our domain-specific modules by evaluating various combinations of counseling components. We show our ablation study in Table 4.2. We observe that masking SH and PD utterances have a drastic effect on R-1 score as it reports a reduction of over 25 points. Similarly, we observe reductions in other metrics as well – 10 points in R-2, 17 points in R-L, 5 points in QAE, and 33 points in BS. Moreover, we observe similar phenomena in masking any two components (PD and RT; SH and RT) as well. We also perform experiments with masking one component at a time and observe inferior results on average. Overall, we observe that removing the information of any counseling components degrades the performance of ConSum. Further, we perform ablation with the MH-Know module and observe a similar phenomenon with the omission of the MH-Know module as well.

4.5.3 Error Analysis

Table 4.4 shows example summaries generated by T5 as well as ConSum. Sticking to our primary aim of covering mental health-related aspects in summaries, we observe that T5 generates rich semantics; however, lacks in attending critical domain-related phrases in many cases. On the other hand, ConSum encapsulates more domain knowledge and captures critical information inherited from three counseling components. For instance, in the first example in Table 4.4, T5 covers only the high-level gist of the conversation, viz. ‘*They secretly talk about incident.*’, whereas ConSum covers essential phrases keeping the semantics intact viz. ‘*Patient feels overwhelming and feels like having this issue since childhood.*’ Likewise, due to its domain relevance, ConSum is able to highlight cases such ‘*family history of suicide*’, which is not included in T5’s generated summary. Moreover, we observe that T5 also commits a few grammatical mistakes and generates irrelevant text in both examples (as highlighted in red).

Model	R-1	R-2	R-L	QAE	BS
PLM	34.24	11.19	33.35	24.34	-0.8678
RankAE	25.57	3.43	24.16	29.98	-1.063
SM	20.46	3.80	18.87	20.22	-0.9454
Pegasus	29.71	7.77	27.57	36.80	-0.6130
T5	31.44	5.63	27.38	33.55	-0.5655
ConSum	45.36	15.71	24.75	25.42	0.3407

Table 4.1: Results obtained on MEMO. We report Rouge-1 (R-1), Rouge-2 (R-2), Rouge-L (R-L), Bleurt Score (BS) and QuestEval Score (QAE).

Counselling Label	R-1	R-2	R-L	QAE	BS
ConSum – SH – PD	20.92	5.00	7.44	20.71	0.0019
ConSum – PD – RT	36.00	9.00	9.14	20.47	0.2032
ConSum – RT – SH	28.63	8.06	9.55	23.02	-0.0209
ConSum – SH	39.77	9.55	8.98	24.11	0.1908
ConSum – PD	36.87	10.02	11.22	33.38	0.2420
ConSum – RT	42.01	9.83	16.50	18.03	0.2060
ConSum – MH-Know – CCC	39.67	9.95	12.69	21.19	0.2003
ConSum – MH-Know	40.42	10.09	11.00	23.97	0.2429
ConSum	45.36	15.71	24.75	25.42	0.3407

Table 4.2: Ablation study on the effect of different modules of ConSum. SH, PD, DF and RT are psychotherapy elements. MH-Know and CCC are knowledge infused module and counseling component classifier module respectively. We present our analysis on five scores to understand all aspects of summary generation.

Table 4.4: Comparative analysis on generated summaries between T5 and ConSum model. Green denotes primary medical condition information, Cyan denotes SH information, Brown denotes RT information and Red denotes errors/irrelevance in generated summaries. Comparison shows ConSum captures better contextual information compared to T5.

Model	Summaries
T5	<i>Patient is asked a series of questions to test their housing. Patients do not want to take drugs, police arrest them and others in their area. They secretly talk about the incident. The patient lives with friends from different places including the city, where they are arrested or taken into custody. The patients feel they have no medical condition such as diabetes or stroke. These occurred recently. The patient has had problems with housing due to augensis. Patients feel they have been victimized by drug abusers since they were 12 years old. They know people who don't use marijuana to steal</i>
ConSum	<i>Whenever there is a basketball or or football, the patient does not really hear what to take. Patient feels overwhelming and feels like having this issue since childhood. Therapist made him realise his problems and patient was curing gradually. Therapist suggested him use. The patient was raised Catholic and they feel guilty about spending more money on drugs and parties. The patient has switched 12 houses in the past year and feel it must have affected their kids. The patient occasionally borrows money and gets less sleep in general</i>
T5	<i>The patient is tired when they have to go to work and it's a bit frustrating for them. They feel tired throughout the day without any food, no panic attacks, no medical condition such as diabetes or stroke. Patients are in a position where they can focus on anything. They do not want pills to reduce stress hence their life is limited by diet. The patient has lost three pounds in recent weeks due to this fatigue</i>
ConSum	<i>Whenever the patient goes to work. The patient is worried that they might have ADHD. The patient does not suffer from depression, anxiety nor use drugs or call it a metaphysical stuff. The patient wishes to get better and needs something to hold. The patient feel they sway at things, and they have two options. The patient was sent in by a counselor fearing they might hurt themselves. The patient's dad had committed suicide 15 years ago and their sister had attempted once. The patient has been diagnosed with depression and anxiety. The patient lives alone</i>

4.5.4 Clinical and Human Evaluation

A team of mental health experts use clinical acceptability framework (154) to evaluate generated summaries. The clinical acceptability framework involves six parameters – affective attitude, burden, ethicality, coherence, opportunity costs, and perceived effectiveness. Experts rate each summary considering the acceptability parameters on a continuous scale of 0 to 2, where higher rating signifies better acceptability. Through expert evaluation, we obtain an average rating of 0.94 out of 2.00 – which falls under the standard acceptability range of 0.70 to 1.40 by mental health experts (154). Hence, the domain experts conclude the summaries to be acceptable for applications in therapy settings. However, in some cases, they pointed out issues with ConSum in comprehending metaphors and subtle transcription errors, which affected the semantic.

Furthermore, results of human evaluation on four standard linguistic parameters namely, *relevance* (selection of relevant content), *consistency* (factual alignment between the summary and the source), *fluency* (linguistic quality of each sentence), and *coherence* (structure and organization of summary) are shown in Table 7.4. Each parameter is rated on the scale of 1 to 5 and found to be qualitatively better than the best baselines.

Model	Relevance	Consistency	Fluency	Coherence
RankAE	2.80	2.91	3.02	2.98
T5	2.99	3.05	3.04	2.95
ConSum	3.37	3.22	3.11	3.13

Table 4.3: Human evaluation on the summaries generated from ConSum model.

4.6 Mental Health Information Capture (MHIC) Metric

Our primary aim is to capture most critical components of mental health counseling leading to successful interventions. Since most of the deep learning methodologies are able to generate contextually and semantically rich text, we observe their incompetence in attaining essential phrases relevant to mental health aspects such as symptom, history of the illness, or other reflecting aspects. Consequently, it becomes challenging to qualitatively assess the summaries and discriminate between latest and domain-specific models. Hence, we propose a new problem-specific metric, MHIC, to mitigate the issue of disregarding such aspects. Mental Health Information Capture (MHIC) attains to the domain-specific counseling

components in generated summaries.

It incorporates the prediction of counseling components classifier and Rouge 1 score (R-1) to evaluate the generated summaries qualitatively. We calculate MHIC for each counseling component $CL_i \in \{SH, PD, RT\}$ as follows:

$$MHIC(CL_i) = R1_Score(S_g, H_i)$$

where H_i represents concatenation of all utterances predicted with CL_i and S_g denotes the generated summary. Figure 4.4 shows the performance of our model and the best baseline, T5. We also show plots with the inclusion and exclusion of counseling labels for each component. We observe that ConSum’s summaries are more aligned towards the three counseling components. Moreover, we observe that utterances tagged with the SH label have significant representation in generated summaries.

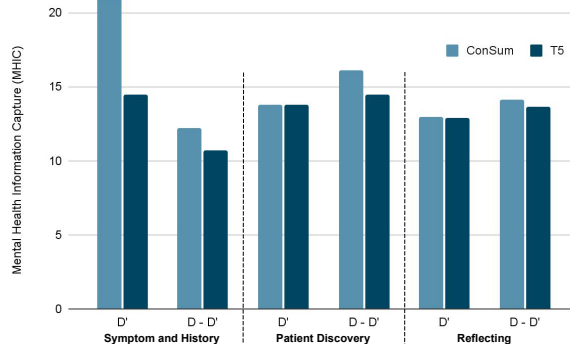


Figure 4.4: Comparison between ConSum and T5 summaries and condensed dialogue utterances based on counseling components. The y-axis represents Rouge-1 F1 scores. D' denotes the concatenation of utterances tagged with a particular counseling component, and $D - D'$ denotes the concatenation of utterances excluding the corresponding counseling component.

4.7 Discussion

4.7.1 Justification for Selection of PHQ-9 Framework

Clinically, PHQ-9 is among the most widely used and validated instruments for screening, diagnosing, and monitoring depression, making it a credible choice in our work. From the modeling perspective, prior research (I31) has developed lexical extrapolations of the PHQ-9 questionnaire, which makes it feasible for NLP models to directly use it. Furthermore, a substantial portion of the counseling dialogues in our dataset naturally center on depressive symptoms. That said, we agree that expanding this approach to incorporate additional clinical instruments in future work would enhance symptom diversity. Such extensions would require collaboration with experts to first produce lexicons associated with each symptom.

4.7.2 Uniqueness of Mental-health counseling and ConSum

Though dialogue summarization in clinical domain (I30; I65; I66) is a well-explored research area, it remains unexplored in the mental health counseling domain. Considering this, in this paper, we raise the research question as “*Why mental health counseling summarization is different than other medical conversation summarization?*”. To answer this question, first we draw significant differences between the two domains and subsequently, show that our proposed model, ConSum, is specifically designed to handle crucial counseling components of the mental health conversations. Clinical practitioners directly aim for the core problem and the conversation about disease and symptoms are up-front – usually patients do not shy away discussing their concerns. On the other hand, patients suffering from mental health issues

are very much reluctant to speak out their issues in a blunt fashion.

In our proposed model, we leverage these insights and specifically design various domain-specific sub-modules (DFC, MH-Know, and CCC) to account for such attributes. As a result, ConSum is unique to the mental-health counseling domain instead of other domains. To demonstrate this, we exploit an existing summarization dataset for clinical conversations in Chinese (ChiCCo) (130) and extend it for our purpose. Following (130), we translate the dataset into English using Google Translate. Since ConSum requires counseling psychotherapy elements to operate, we obtain these from the pretrained counseling component and discussion filler classifiers. Subsequently, we train ConSum and other comparative systems on the ChiCCo dataset.

In comparison with the encouraging performance of ConSum on MEMO (c.f. Table 4.1), we argue that ConSum is competent in handling crucial components of the mental health counseling pretty efficiently, as in the absence of such components in other domains, it underperforms.

4.8 Summary

Considering the situation of mental health therapies amid the COVID-19 pandemic, in this paper, we attempted to empower therapists through our research in mental health counseling summarization. To this end, we developed a novel summarization dataset, MEMO, in a dyadic counseling setup. Moreover, to model the counseling summarization in an abstractive fashion, we proposed a domain-enrich transformer architecture, ConSum. It utilized three domain-centric sub-modules *viz.* a discussion filler classifier, a counseling component classifier, and a PHQ-9 driven mental health knowledge-infused utterance selection module. Our extensive ablation study and the comparative analysis established the superiority of ConSum over several existing models.

5. Plan Before Summarization

In mental health counseling, condensing dialogues into concise and relevant summaries (*aka* counseling notes) holds pivotal significance. Large Language Models (LLMs) exhibit remarkable capabilities in various generative tasks; however, their adaptation to domain-specific intricacies remains challenging, especially within mental health contexts. Unlike standard LLMs, mental health experts first plan to apply domain knowledge in writing summaries. Our work enhances LLMs' ability by introducing a novel *planning engine* to orchestrate structuring knowledge alignment. To achieve high-order planning, we divide knowledge encapsulation into two major phases: (i) holding dialogue structure and (ii) incorporating domain-specific knowledge. We employ a planning engine on Llama-2, resulting in a novel framework, `PIECE`. Our proposed system employs knowledge filtering-cum-scaffolding to encapsulate domain knowledge. Additionally, `PIECE` leverages sheaf convolution learning to enhance its understanding of the dialogue's structural nuances. We compare `PIECE` with 14 baseline methods and observe a significant improvement across ROUGE and Bleurt scores. Further, expert evaluation and analyses validate the generation quality to be effective, sometimes even surpassing the gold standard. We further benchmark `PIECE` with other LLMs and report improvement, including Llama-2 (+2.72%), Mistral (+2.04%) and Zephyr (+1.59%), to justify the generalizability of the planning engine.

5.1 Introduction

Mental health counseling serves as a crucial frontline defense against mental illness. In a typical counseling session, clients articulate their issues while therapists provide support. An essential component of these sessions involves building a strong therapeutic bond and documenting the entire dialogue, commonly known as counseling note, record, or summary¹. However, traditional methods of note-taking during sessions can present significant challenges, where therapists are required to divert their attention to note-taking, disrupting counseling interaction and support quality. This distraction not only hinders the therapist's focus but also deteriorates the required therapeutic bond, necessitating the need to automate this process.

While recent advancements in Large Language Models (LLMs) for mental health are enormous (I167; I168), it is crucial to incorporate fundamental domain knowledge and understanding of the structural nuances of the dialogue, which current research lacks in ability (I169). Consequently, we require a knowledge planner to adeptly capture domain intricacies and determine generation priorities. For instance, Figure 8.3 illustrates a sample counseling conversation between a therapist (T) and a client (C). Unlike conventional approaches that directly generate summaries using foundational models, our proposed pipeline introduces a planner. Through this planner, we infuse nuanced dialogue structure and domain-specific knowledge into LLMs.

Earlier efforts in counseling summarization (I170) focused on utilizing domain knowledge by integrating patient health questionnaires (PHQ) and counseling components. However, it fell short compared to recent LLMs. The broader scope of such models has been explored for incorporating knowledge into

¹www.apa.org/gradpsych/2007/01/track

generated text (171) and controlling text generation (172). While these efforts are focused on general-purpose downstream tasks, the complexity of counseling dialogue necessitates introducing a more (a) domain-centric and (b) knowledge-structured approach to cater to this problem.

Our study introduces a novel *planning engine*, designed specifically to guide the generation of LLMs. Focused on enhancing counseling summarization, we employ MentalLlama as the foundation to develop our framework – P I E C E, **p**lanning **e**ngine for mental counseling note generation. Our model’s planning engine exploits knowledge scaffolding and sheaf learner by integrating domain-specific and structural knowledge into LLMs. We evaluate our P I E C E against several LLMs, such as Mistral, Zephyr, and Llama. To assess the effectiveness of our approach, we compare P I E C E’s performance against 14 baseline methods. Our quantitative evaluation consists of four automatic summary evaluation metrics – ROUGE: R-1, R-2, R-L, and Bluert, along with a domain-centric metric, Mental Health Information Capture (MHIC) proposed by (170). We observe a clear improvement of 3.42%, 10.11%, and 6.01% across R-1, R-2, and R-L metrics, respectively. Furthermore, to assess the generalizability of P I E C E, we experiment on ACI-BENCH, a clinical note generation dataset (173). We observe that P I E C E surpasses the state-of-the-art LLMs with planning-engine on ACI-BENCH as well.

Additionally, we perform an extensive expert evaluation through an established clinical relevance framework on a set of six dedicated metrics and three task-relevant survey questionnaires. The results affirm the superiority of P I E C E against the baselines and demonstrate the adaptability of the planning engine across various LLMs. Our contributions are summarized below:

- We propose P I E C E that integrates the planning engine with MentalLlama to address the issue of unreliable generation by LLMs. This engine plans the LLM’s generation by filtering dialogue and injecting domain and structural knowledge.
- We extensively evaluate P I E C E against 14 baseline methods. We present P I E C E’s significant improvement evaluated across both automatic, human, and expert evaluation metrics.
- We demonstrate the adaptability of the planning engine as it seamlessly integrates with alternative LLMs like Mistral, Zephyr, and Llama. This may also expand the research in planning the generations of diverse LLMs.

5.2 Related Work

We present our literature review under three major segments to understand and build a planner for LLM-based counseling summarization.

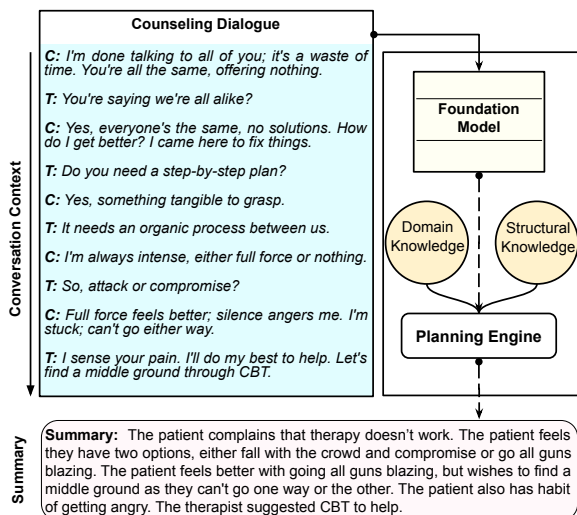


Figure 5.1: The proposed pipeline allows LLMs to first plan and then generate. In our approach, prioritizing planning before generation enriches summarization with conversational structure and domain knowledge.

Generative AI in Mental Health Domain. Recent advances in generative research in mental health have led to a diverse range of investigations. For instance, (174; 175) utilized GPT-2 for reinforced feedback generation within peer counseling setups. Subsequent studies employed similar methods with language models (LMs) for empathetic generation (175) and facilitated human-AI collaboration. Several studies have concentrated on controlled dialogue generation using LMs, leveraging reinforcement learning (176; 177; 178; 179; 180; 181; 182). Despite this, there remains a scarcity of research aimed at enhancing LLMs in this domain (183). The concept of counseling summarization was previously explored by (170), where they proposed a filtering mechanism based on annotation and domain knowledge to generate counseling summaries using LMs. As the focus has gradually shifted towards LLMs, recent findings by (184) have indicated that while LLMs exhibit robust capabilities, they still exhibit significant gaps compared to domain-specific methods.

Knowledge Enhancement in LLMs. There has been a significant effort to incorporate external and domain-specific knowledge into LLMs, as noted in various studies (185; 186; 187). Tailoring LLMs to specific domains has notably enhanced their ability to handle downstream tasks. Within the clinical domain, a series of specialized LLMs have emerged (188). There exist a few dedicated LLMs specifically trained on mental health corpora, such as the variants of Llama-2 (189), BART (190), and T5 (151) as MentalLlama, MentalBART and MentalT5, respectively (191). While domain-specific LLMs facilitate more contextual relevance, they do not entirely mitigate risks associated with hallucinations or missing context. Planning LLM generations is one of the many solutions here (192). Planning methodologies have been explored in various generative tasks, including reasoning (193), temporal generation (194), and code generation (195). However, existing planning systems often lack domain-specific planning, notably within the mental health space.

Text Summarization. Knowledge-guided summarization has long been a research focus, and its variations include information-aware techniques (196), perspective-based strategies applied in educational dialogues (197) and scientific document summarization (198; 199), commonsense-driven clinical summarization (200), as well as approaches focusing on topic-awareness (201) and attention mechanisms (202). However, many of these methods largely rely on annotated corpora and the LMs to generate coherent text, neglecting deeper exploration into structural comprehension of input data. Limited research efforts, exemplified by (203), discuss the use of sheaf for enhanced structural understanding, while (204) proposed sheaf to encapsulate structural information for summarization.

5.3 Dataset

We utilize MEMO, a common counseling summarization dataset with 191 counseling dialogues with 11,543 utterances (170). Each dyadic counseling dialogue contains a conversation between a therapist (5,722 utterances) and a client (5,814 utterances), along with an expert annotated counseling summary. In addition to this, MEMO further contains each utterance labeled with a counseling component – Symptom and History (SH), Patient Discovery (PD), Reflecting (RT), or Discussion Filler (DF). The dataset contains 2379 utterances tagged with SH, 5428 utterances tagged with PD, and 1242 utterances tagged with RT. These counseling components act as labels of relevance for utterances.

5.4 Methodology

Our proposed model, *PIECE*, operates on the *MEMO* dataset and aims to generate knowledge-enriched counseling summaries for a counseling dialogue containing n utterances, $D = \langle u_1, u_2, u_3, \dots, u_n \rangle$. There are two specific kinds of knowledge that *PIECE* focuses on – (a) (mental health) domain knowledge and (b) structural understanding of dialogue. The model achieves this knowledge integration through a novel *planning engine* that filters and fuses relevant information to guide the generation of knowledge-enriched summaries by the underlying LLM. Figure 5.2 presents a schematic diagram of *PIECE*.

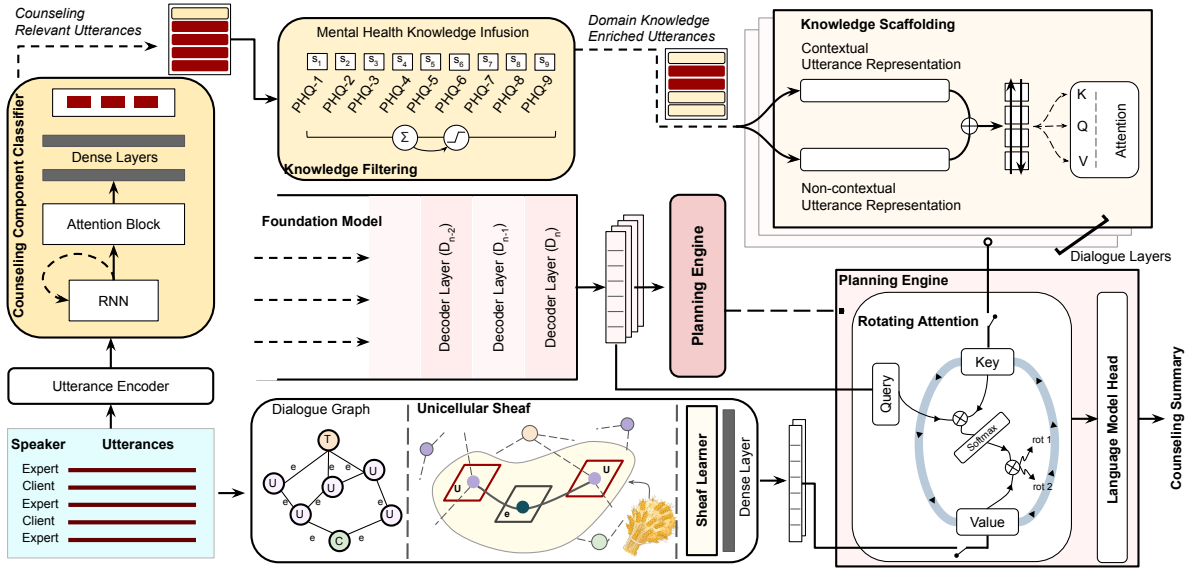


Figure 5.2: Architecture of *PIECE*. We propose a novel *planning engine* consisting of two primary sections: (a) integrating knowledge filtering-cum-scaffolding and (b) encapsulating structural understanding of dialogues. Filtration of relevant utterances utilizes component labeling counseling within the *MEMO* dataset to mask filler utterances, followed by knowledge scaffolding. Additionally, sheaf learners are employed for the structural understanding of counseling dialogue. The planning engine operates using a rotating attention mechanism using knowledge from both segments for better LLM generation.

5.4.1 Domain Knowledge Encapsulation

Knowledge encapsulation module involves *knowledge filtering* and *knowledge scaffolding*.

Knowledge Filtering. Knowledge filtering requires carefully identifying the most relevant utterances for crafting a knowledge-rich summary. This process initiates with classifying each utterance U_i into counseling components (SH, PD, RT, and DF). Among these, SH, PD, and RT are considered essential for generating comprehensive summaries, while DF utterances are regarded as non-essential and are therefore masked by leveraging the dedicated *Counseling Component Classifier*. This classifier is purposefully crafted to process the context of utterances, $\langle u_1, u_2, u_3, \dots, u_i \rangle$, through a GRU unit coupled with a self-attention block. Subsequently, a sequence of two dense layers learns these context-rich utterance representations to predict counseling components and filter non-essential discussion fillers.

The subsequent step involves examining the filtered utterances and further scrutinizing them using

a mental health knowledge selection module, MH-Know. Similar to the approach opted by (170) with the MEMO dataset, our module leverages the lexicon derived from the Patient Health Questionnaire (PHQ-9) to gauge the degree of mental health-specific content embedded within each utterance (205). The similarity scores (s_i) are computed between individual utterances and the PHQ-9 lexicons. Subsequently, a predetermined threshold ($T^\circ = 50\%$, in our study) is applied to ascertain whether an utterance merits retention or should be masked. The resulting filtered dialogue, composed of essential and knowledge-rich utterances, then proceeds to the *knowledge scaffolding* stage to ensure optimal preservation of the embedded knowledge.

Knowledge Scaffolding. To ensure the cohesion and effective organization of the filtered knowledge segments, PIECE employs a knowledge scaffolding module. We adopt the scaffolding methodology proposed by (206), where the authors incorporated structural information of scientific papers into citations for citation-intent classification. We modify their knowledge scaffolding for dialogue settings by employing a mix of contextual and non-contextual representations for each relevant utterance, as shown in Equation 5.1

$$R_k = \text{Attention}[\overleftrightarrow{\chi}(C(u_i) \oplus C^\circ(u_i))] \quad (5.1)$$

Here, the contextual embeddings $C(x)$ are captured through BERT (51), while the non-contextual embeddings $C^\circ(x)$ are extracted using Glove (207). These representations are then carefully concatenated and fed into a bidirectional LSTM layer $\overleftrightarrow{\chi}$, ensuring the underlying semantic context. The resulting context-rich representations, enriched with domain-specific knowledge, subsequently serve as key, query, and value vectors within a self-attention block. This orchestrated interplay emerges in generating filtered scaffolded representations R_k enriched with the essence of extracted knowledge. Before understanding the role of these representations within the planning engine, we first understand the encapsulation of structural knowledge, another crucial component of PIECE’s knowledge-enriched summary generation pipeline.

5.4.2 Structural Knowledge Encapsulation

To ensure that the generated summaries accurately reflect the overall structure and flow of the original dialogue, PIECE incorporates a dedicated structural knowledge module. Sheaf has been long studied for its ability to hold structural knowledge in graphs via sheaf diffusion (208; 209; 203). With a similar objective, we first construct a dialogue graph $G = (V, E)$ that captures the intricate network of relationships between the utterances within the dialogue D . Each utterance u_i is represented as a vertex $v \in V$ within the graph, while edges $e \in E$ are directed toward the progression of dialogue. In the adjacency matrix A_d of the graph G for each dialogue instance $d \in D$ with o utterances, we place each node’s feature f with its encoded utterance representations $f \leftarrow bert(u_i)$. As a result, we construct G to preserve the inherent dialogue structure while learning the graph representations via *sheaf learner*.

Sheaf Learner. Graphs inherit a notion of the neighborhood but not distance or direction. Sheaf theory provides a mathematical framework for imbuing graphs with geometric structure by associating vector spaces, known as stalks, with each node and edge, along with defining restriction maps (which are essentially linear transformations) between the stalks of incident nodes and edges. These restriction maps capture the local-to-global context, ensuring consistency of information across the graph. The resulting structure, comprising the graph along with the decorated stalks and restriction maps, is termed a cellular sheaf (208). To facilitate the understanding of complex graph relationships, we construct a sheaf, which is a collection of interconnected cellular sheaves. Drawing inspiration from (203), we employ a sheaf

convolution network (*SCN*) (see Equation 5.2) to learn these sheaves.

$$R_{scn} = \eta((I - \Delta)(I \otimes W_1)A^{(o \times f)}W_2) \quad (5.2)$$

Here, A is the input to the network, whereas W_1 and W_2 are learnable weights. Δ acts as a sheaf laplacian, I is an identity matrix, and \otimes represents direct matrix multiplication. We apply an activation function η on top as a *relu* function. For the resultant representation R_{scn} , we employ a dense layer to learn the graph geometry over topological space, inheriting better structural knowledge. The resultant structure-rich representations R_s act as an input to the planning engine.

5.4.3 Planning Engine

The planning engine is the core module that integrates the structural and domain-specific knowledge using a rotating attention mechanism. Acknowledging the equivalent significance of both structural and domain-specific insights, the rotating (cyclic) attention mechanism aims to retrieve information evenly from each segment. Here, domain-rich scaffolded representations R_k and structure-rich sheaf representations R_s act as both key (K) and value (V) once per cycle, yielding two rich dialogue representations. The query, the foundation model’s hidden representations, acts as query (Q) for both cycles, as shown in Equation 5.3.

$$R_{rep} = sm\left(\frac{QK^T}{\sqrt{d_k}}\right)V \oplus sm\left(\frac{QV^T}{\sqrt{d_k}}\right)K \quad (5.3)$$

The language model head (LM-Head) operates on top of the fused representations of both dialogue representations to generate a counseling summary.

5.5 Experiments and Results

Here, we discuss the selection of baseline methods followed by the performance comparison with baselines, the ablation of `PIECE`, and analyses.

5.5.1 Baselines

We choose the following systems as our baselines – (i) **Segmented Modeling (SM)** leverages dialogue-acts on BiLSTM (164). (ii) **BART** is a standard transformer with a BERT encoder and GPT decoder (190). (iii) **Pegasus** is pretrained with an objective of gap-sentence-generation (150). (iv) **T5** uses a shared framework on transformer to pretrain on huge C4 corpus (151). (v) **Pretrained Language Model (PLM)** uses DialoGPT to segment topics and generate summaries with BART (161). (vi) **RankAE** uses an encoder to filter dialogue utterances into segments and generates summaries using denoising auto-encoder (162). (vii) **Summ^N** is a language model pretrained for dialogue summarization (210). (viii) **DialogLM** is pretrained for long conversational summarization task (211). (ix) **ConSum** is the reported state-of-the-art for counseling summarization task on the MEMO dataset, marking as the most relevant baseline method to compare with (170). (x) **Flan-T5** is a T5-based LLM, instruction-tuned on a mixture of tasks (212). (xi) **Mistral-7B** is an LLM instruction tuned with a sliding window attention mechanism for efficient and longer context (213). (xii) **Zephyr-7B** is a fine-tuned Mistral LLM to surpass the large chat models (214). (xiii) **Llama-2-7B** is an optimized auto-regressive LLM currently state-of-the-art for many language generation tasks (189). (xiv) **MentalLlama** is a standard Llama-2 model pretrained on a

huge Reddit-based mental health corpus (191). Our work is proposed on top of MetalLlama because of its domain-specific knowledge and performance.

To evaluate the performance, we employ widely recognized metrics: ROUGE and Bleurt Score.

5.5.2 Performance Comparison

Table 5.1 shows the performance of the baseline models, revealing the clear superiority of PIECE across three out of four metrics. Notably, MetalLlama turns out to be the best-performing baseline. PIECE excels in capturing both semantic and syntactic structures, as evidenced by the improvements of best-performing LLMs like Llama variants. Specifically, PIECE demonstrates improvements of +3.42%, +10.11%, and +6.01% in R-1, R-2, and R-L points, respectively. On the other hand, Llama yields the best scores for the Bluert metric; however, PIECE closely matches its performance, exhibiting a marginal drop of only -0.0004 Bluert points. However, compared with ConSum, the benchmarked state-of-the-art model for counseling summarization, our findings reveal a significant boost in performance metrics. Notably, PIECE shows improvements of +9.39%, +18.45%, +53.93% and +20.39% in R-1, R-2, R-L and Bleurt points, respectively.

Model	R-1	R-2	R-L	BS
SM (164)	20.46	3.80	18.87	-0.9454
BART (190)	34.92	12.66	18.83	-0.7118
Pegasus (150)	29.71	7.77	27.57	-0.6130
T5 (151)	31.44	5.63	27.38	-0.5655
PLM (161)	34.24	11.19	33.35	-0.8678
RankAE (162)	25.57	3.43	24.16	-1.0630
SUMM ^N (210)	34.06	11.32	20.99	-0.6088
DialogLM (211)	28.14	9.21	17.57	-0.7377
ConSum (170)	45.36	15.71	24.75	0.3407
Flan-T5 (212)	41.30	16.00	29.05	0.2281
Mistral-7B (212)	46.74	14.98	32.48	0.3056
Zephyr-7B (214)	38.90	9.90	26.35	0.0094
Llama-7B (189)	47.22	16.74	33.16	0.4106
MetalLlama (191)	47.92	16.90	35.81	0.3953
(<i>fm</i> : MetalLlama)	49.62	18.61	38.10	0.4102
$\Delta_{-BEST}(\%)$	$\uparrow 3.42$	$\uparrow 10.11$	$\uparrow 6.01$	$\downarrow 0.09$

Table 5.1: Results obtained on the MEMO counseling summarization dataset. We report Rouge-1 (R-1), Rouge-2 (R-2), Rouge-L (R-L), and Bleurt Score (BS).

5.5.3 Ablation Study

We perform an ablation study to assess the performance of underlying components contributing to PIECE. By systematically deconstructing the planning engine and analyzing various elements of our model architecture, we present our findings in Table 5.2. The impact of the planning engine on the acting foundation models (*fm*) assesses the generalizability of the planning engine. Evidently, the role of the addition of knowledge is highlighted as we observe a clear decline in the performance of PIECE across all four metrics. We observe the same trend for both structural and domain knowledge, with the performance declining by a significant margin of -2.83 (R-2) points. At the same time, we perform a generalizability check for the planning engine to be inducted on top of multiple LLMs such as Mistral, Zephyr, and Llama, finding that each LLM, in general, benefits from the planning engine.

5.5.4 Qualitative Analysis

Model	Summaries
Gold	<i>Patient visited the therapist for feeling bad. Therapist knew that patient had such feelings in the past, but this time patient found it harder. Two months ago patient visited GP after that GP referred to therapist. Earlier patient had a breakup and felt lonely. Therapist asks how things have been since the patient went to GP. Patient feels nothing changed and unable to accept failure. Patient’s mom was not happy with the patient’s behavior. Therapist suggested CBT. Patients accept that the first year was tough. Therapist asks that it doesn’t sound like you want to go back. Patient wants to but just doesn’t feel like meeting the man. Patient used to sit in a friend’s restaurant for money and stuff in the meantime. Patient used to like it because the patient used to speak to people. Therapist asks to scale the thoughts out of 100. Patient says 80. Patient feels empty when asked about negative thoughts. Patient doesn’t want her mom to be too worried.</i>
Llama	<i>Patient adrift after a brutal breakup finds solace in a therapist’s office. The weight of past struggles, a university hiatus, and crushing expectations press down, leaving her numb and weary. Work, once at a restaurant, becomes a reminder of self-doubt. But as secrets unfurl – anxieties, fears, and the facade she wears for her mom – a glimmer of hope flickers. Maybe, just maybe, this conversation, this courageous first step, can illuminate a path back to wholeness.</i>
MentalLlama	<i>Patient, batling depression and loneliness, seeks therapeutic guidance. Past heartbreak and struggles weigh heavily, leaving her drained and empty. Work offers a temporary escape, but anxieties about “not being good enough” resurface. The therapist recognizes the patient’s inner strength and the burden of hiding pain from her mother. Through the conversation, a flicker of hope emerges, hinting at the possibility of a different path forward rooted in understanding and compassion.</i>
PIECE	<i>The patient visited the therapist feeling unable to recover from recent emotional difficulties, notably following a breakup with a long-term boyfriend. Expressing disappointment in seeking therapy and reluctance to admit feelings of failure after being referred by the GP, the therapist highlighted the normalcy of anxiety. Concerns about returning to university after a break were discussed, reflecting the patient’s desire for success amidst overwhelming pressure and self-doubt. Additionally, feelings of inadequacy arose during a recent work shift at a friend’s restaurant, causing the patient to feel good but confused in talking to people. The therapist further explored the patient’s sense of guilt for not meeting expectations, especially with mom, revealing a deep emotional impact. The patient admitted to feeling empty and drained, rating their emotional distress at 80 out of 100 and describing physical sensations resembling nervousness. The therapist understood the patient’s struggle to appear unaffected for their mother’s sake, despite internal turmoil, ultimately feeling worse for doing so.</i>

Table 5.3: A comparative analysis of three most relevant models – Llama, MentalLlama, and P I E C E. The key emphasis on being analyzed here is understanding the core counseling components, including discussions on *symptoms and reasons*, *reflecting client’s story*, and *critical routine details* along with the *additional intricate domain understanding*. While Llama generates general-purpose summaries, MentalLlama, being pretrained on mental health data, captures nuanced knowledge beyond Llama’s scope. In contrast, P I E C E is able to capture in-depth domain knowledge, surpassing other models in touching upon counseling nuances. Despite the grammatical proficiency of baselines, P I E C E stands out for its structural understanding, focusing on intricate details, in some cases, with more detailed accounts of these aspects than are provided in the reference summary

We assess P I E C E’s text generation capabilities, apart from quantitative metrics, via detailed comparative analysis between P I E C E and the two most relevant models: Llama and MentalLlama. Given that P I E C E builds upon the MentalLlama, we carefully compare its performance against these two models and the gold standard. Table 5.3 presents an illustrative comparison. Our scrutiny centers on discerning the understanding of core counseling components, including discussions on *symptoms and reasons*, *reflecting client’s story*, and *critical routine details* of the dialogue along with the *additional domain understanding*.

Ablations	R-1	R-2	R-L	BS
P I E C E				
– StructKnow	48.25 (↓ 1.37)	16.22 (↓ 2.39)	37.40 (↓ 0.70)	0.4085 (↓ 0.0017)
– DomainKnow	48.36 (↓ 1.26)	15.78 (↓ 2.83)	36.96 (↓ 1.14)	0.3994 (↓ 0.0108)
($f_m = \bar{\alpha}$)				
α : +Mistral	46.81 (↑ 0.07)	13.28 (↓ 1.70)	34.52 (↑ 2.04)	0.3080 (↑ 0.0024)
α : +Zephyr	41.63 (↑ 2.73)	11.36 (↑ 1.46)	27.94 (↑ 1.59)	0.0106 (↑ 0.0012)
α : +Llama	48.25 (↑ 1.03)	14.02 (↓ 2.72)	35.88 (↑ 2.72)	0.4098 (↓ 0.0008)

Table 5.2: Ablation study of the proposed model, P I E C E and generalizability of the planning engine on top of notable LLMs as foundation model (α), including Mistral, Zephyr, and Llama, illustrating a clear *improvement* (↑) in LLM generations by integrating planning engine.

In Table 5.3, the first row showcases the gold standard summary. Evidently, Llama excels in text coherence but falls short in producing domain-specific insights and barely touches upon the critical domain information, for example, *client’s referral from GP*, *self-emotional-assessment rating*, and *declaration*

of anxiety. In contrast, MentalLlama demonstrates better mental health-specific nuances, highlighting counseling components such as *battling depression and loneliness, the burden of hiding pain, and past heartbreaks* and integrating intricate domain information such as *struggles weigh heavy and the possibility of a different path forward*. Apparently, MentalLlama’s performance lacks a grasp of the structural intricacies of dialogue. On the other hand, the summaries by P I E C E are descriptive, emphasizing the complete dialogue structure and crucial counseling components inherited from the conversation such as *therapist highlighting normalcy, self-assessment, reluctance to admit feelings*, instances which MentalLlama and Llama skipped. Additionally, P I E C E excels in enhancing domain understanding by incorporating spans from the LLM’s vocabulary rather than the conversation’s vocabulary.

Error Analysis: We extend the analyses for cases where our model falls short in capturing the intended details. We discuss two important cases: a) P I E C E tends to include exaggerated information in shorter dialogues, leading to extra details, and b) P I E C E summaries are usually longer than gold summaries and, on average, struggle to capture patient’s behavior.

Model	Relevance	Consistency	Fluency	Coherence
Llama	3.12	3.22	3.76	3.68
MentalLlama	3.57	3.31	3.75	3.71
P I E C E	3.73	3.39	3.72	3.75

Table 5.4: Human evaluation on the summaries generated from P I E C E model. The average interrater’s agreement score (κ) for P I E C E is 0.82.

Human Evaluation We present human evaluation on four standard linguistic parameters, namely *relevance* (selection of relevant content), *consistency* (factual alignment between the summary and the source), *fluency* (linguistic quality of each sentence), and *coherence* (structure and organization of summary) as shown in Table 5.4. We employed 12 linguistics experts to rate each parameter on the Likert scale of 1 to 5. Out of twelve, seven were female, whereas five were male, all of them aged between 23 - 35. As shown in Table 5.4, P I E C E surpasses the average scores across three out of four evaluation parameters. Notably, it achieved a score of 3.73 for relevance, indicating that P I E C E’s summaries capture the core knowledge. Additionally, P I E C E scored 3.39 and 3.75 for consistency and coherence, respectively, showing the logical flow of the original conversation. Finally, P I E C E competes with *Llama* on *fluency* metric. The superior performance underscores the model’s linguistic quality and structural coherence. The average Cohen’s kappa score (κ) for P I E C E is 0.82, which falls under the substantial category.

5.5.5 Mental Health Information Capture

The Mental Health Information Capture (MHIC) computes the intersection of utterances predicted by the counseling components classifier and the generated summary using the ROUGE-1 score to provide a qualitative evaluation of these summaries. Figure 5.3 presents a comparative metric study of the three top-performing models – ConSum, MentalLlama, and P I E C E. Evidently, the earlier state-of-the-art benchmarked model, ConSum, on the MEMO dataset appears to be easily surpassed by MentalLlama without any additional effort. However, after integrating the *planning engine* to MentalLlama, i.e., P I E C E, performance surpassed both variants, highlighting the efficacy of P I E C E.

5.5.6 Mental Health Expert Validation

Our model, `PIECE` acts as an assistance for complex counseling processes, as experts can utilize the summaries by `PIECE` as counseling notes, reducing both time and cost. To validate the generation’s effectiveness, we collaborated with a clinical psychologist (having 10+ years of clinical experience) who assessed the summaries on two major fronts: clinical acceptability and LLM relevance.

Clinical Acceptability.

Experts exploit a clinical acceptability framework (154) that evaluates parameters such as *affective attitude, burden, ethicality, coherence, opportunity costs, and perceived effectiveness*. The expert rated summaries on a scale from 0 to 2, with higher scores indicating better acceptability (c.f. Table 5.5). The expert evaluation shows an average rating of 1.20 out of 2.00, falling well within the standard acceptability range of 0.70 to 1.40. As a result, the expert concluded that the generated summaries demonstrate suitability for therapists, sometimes surpassing the gold standard.

LLM Relevance.

To assess common LLM flaws, quality of generation, and relevance to experts, we ask experts to address the following aspects:

- Did the expert observe our model hallucinate?
- How does the model’s generation compare to that of the most competitive model (MentalLlama)?
- Is the generated summary relevant to the expert?

Evidently, 75% of instances having ‘negligible’ hallucinations in `PIECE` and 56.3% cases outperformed the most competitive model, MentalLlama. Consequently, 93.8% of cases were deemed either ‘totally relevant’ or ‘relevant to some extent’ for practical applications by experts. However, with 12.25% of instances marked *hallucinated*, there remains room for improvement in the future.

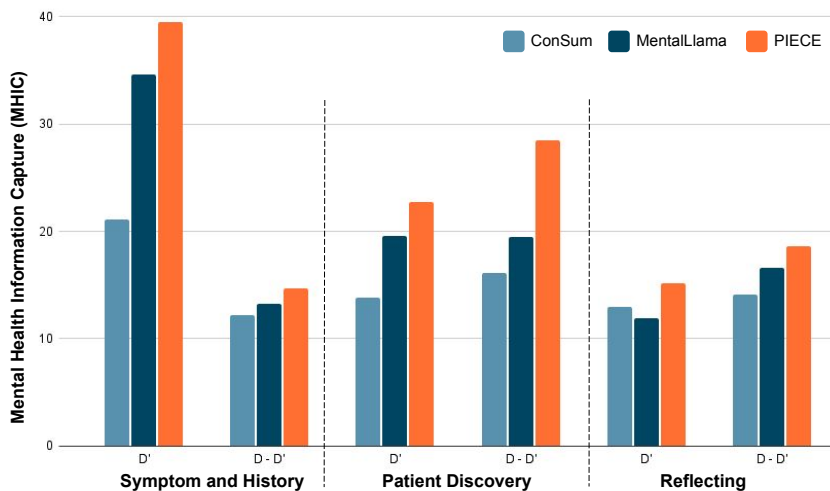


Figure 5.3: Domain-centric evaluation using Mental Health Info Capture (MHIC) metric. The proposed model, `PIECE`, distinctly excels in capturing domain knowledge compared to the two most relevant models.

	Af. Att.	Burden	Ethic.	Intv. Coh.	Cost	Percv. Eff.
μ	1.23	0.36	1.03	1.20	0.85	1.27
(Δ)	(0.21)	(0.34)	(0.23)	(0.33)	(0.19)	(0.27)

Table 5.5: Expert evaluation across six domain-centric metrics: *affective attitude (Af. Att.), burden, ethicality (Ethic), intervention coherence (Intv. Coh.), opportunity cost, and perceived effectiveness (Percv. Eff.)*. We present mean (μ) and standard deviation (Δ) scores.

Impact. The end goal of this study is to prepare `PIECE` for a pilot study for experts via an assistive framework that can help professionals streamline note-taking. Subsequently, this will allow more focus on patient care and improve time efficiency.

5.6 Limitations

Research in the space of mental health counseling summarization, in general, poses several complex challenges. Firstly, there's a substantial scarcity of diverse and high-quality datasets tailored for counseling summarization. **Our work primarily relies on the MEMO dataset, which, to the best of our knowledge, remains the only publicly available dataset in this domain.** The limited dataset diversity hinders robustness, emphasizing the critical need for more datasets in this research area. Another challenge arises from the scale of the LLMs utilized in our research. Pretraining these super-large models on domain-specific corpora incurs significant costs and environmental implications. `MentalLlama`, chosen as the foundation model due to its pretraining corpus, underscores the necessity for more domain-centric models specifically tailored for mental health counseling. Additionally, evaluating LLM generations in a sensitive domain like mental health necessitates a nuanced understanding of the domain's intricacies. LLMs, by nature, exhibit tendencies to generate information that might not accurately represent the domain context and could potentially "hallucinate". Hence, our research, on one end, reduces hallucination by planning the generation and, on the other end, is positioned as an assistive tool tailored for mental health experts rather than direct client-facing applications. This approach mitigates the risks associated with LLM-generated content and underscores the need for expert oversight and intervention in sensitive mental health contexts.

5.7 Ethical Considerations and Future Work

Our research is an augmentation of established state-of-the-art foundation models, focusing on enhancing the generational proficiency of Language Models (LLMs) via planning LLM generation rather than introducing LLMs from scratch. This approach ensures continuous refinement, specifically targeting improvements in mental health counseling summarization within existing models. **Given the sensitive nature of research in the domain of mental health, we portray our research as an assistive module designed exclusively for mental health experts. Such a strategic approach mitigates risks associated with direct support-seeking client interactions, safeguarding against unintended impacts on client sentiments or mental states.** Mental health professionals retain autonomy, choosing whether to accept `PIECE`'s generated summary or further tailor it based on their expertise. However, our research scope acknowledges the evolving nature of complexities and diversity, emphasizing the need for not merely larger but planned LLMs. This leads to aligning LLMs with optimal parameter sizes and prioritizing the planned generational capabilities. Hence, paving the way for more responsible LLMs, upholding ethical standards in the research for sensitive domains like mental health counseling.

5.8 Summary

In our work, we explored the research space of mental health counseling summarization. Our research focused on strategically orchestrating LLMs to plan before generating content, thereby incorporating relevant dimensions of conversational structure- and domain-centric knowledge into the summarization process. We proposed a novel *planning engine* for LLMs, integrated and presented in this research with

MentalLlama as P I E C E . We compared the performance of P I E C E against a spectrum of 14 potential baseline methods, including state-of-the-art LLMs like Mistral, Llama-2, Flan-T5, Zephyr, and MentalLlama. Next, we presented an extensive evaluation spanning automatic, human, and expert evaluations. Expert analysis indicated that the summaries generated by P I E C E exhibit better relevance and structure, often surpassing the closest competitive LLMs and, in certain instances, even surpassing the gold standard. We conclude by discussing our research’s ethical considerations and applicability, emphasizing our approach as an assistive module tailored exclusively for mental health experts, which mitigates potential risks associated with direct client interactions, ensuring a safeguard against unintended harm. This strategic positioning underscores our commitment to safe utilization in mental health contexts.

Part III

Response Generation

6. Dialogue-act Controlled Response Generation

Virtual Mental Health Assistants (VMHAs) have become a prevalent method for receiving mental health counseling in the digital healthcare space. An assistive counseling conversation commences with natural open-ended topics to familiarize the client with the environment and later converges into more fine-grained domain-specific topics. Unlike other conversational systems, which are categorized as open-domain or task-oriented systems, VMHAs possess a hybrid conversational flow. These counseling bots need to comprehend various aspects of the conversation, such as dialogue-acts, intents, etc., to engage the client in an effective and appropriate conversation. Although the surge in digital health research highlights applications of many general-purpose response generation systems, they are barely suitable in the mental health domain – the prime reason is the lack of understanding in the mental health counseling conversation. Moreover, in general, dialogue-act guided response generators are either limited to a template-based paradigm or lack appropriate semantics in dialogue generation. To this end, we propose *READER* – a **RE**sponse-**Act** guided reinforced **D**ialogue **gen**ERation model for the mental health counseling conversations. *READER* is built on transformer to jointly predict a potential dialogue-act d_{t+1} for the next utterance (*aka* response-act) and to generate an appropriate response (u_{t+1}). Through the transformer-reinforcement-learning (TRL) with Proximal Policy Optimization (PPO), we guide the response generator to abide by d_{t+1} and ensure the semantic richness of the responses via BERTScore in our reward computation. We evaluate *READER* on HOPE, a benchmark counseling conversation dataset and observe that it outperforms several baselines across several evaluation metrics – METEOR, ROUGE, and BERTScore.

6.1 Introduction

Virtual Mental Health Assistants (VMHAs) are the backbone of the new-age digital healthcare industry. More than 60% of therapies conducted in the past three years are via virtual assistants. This massive spike in the number of users using VMHAs to gain mental health assistance is due to the ease and safety of access to AI-based therapist-bots (215). Numerous potential platforms, *viz.* Weobot, Wysa¹, etc. in the digital health space are developing practical and effective ways for the common public. More popular than ever, VMHAs are now becoming an instant solution to millions of clients struggling with mental health issues²

Limitations of existing methods. Many such conversational agents fail to (a) understand the directives of the client with whom they are in active conversation and (b) take the conversation in the required direction³. This is similar to the fact that even human therapists find it impossible to reply to something they do not understand from the help-seeker. Therefore, clients' directives directly impact the response

¹<https://woebothealth.com/>; <https://www.wysa.io/>

²<https://psychnews.psychiatryonline.org/doi/10.1176/appi.pn.2022.05.4.50>

³<https://www.bbc.com/news/technology-46507900>

generation capability. Current open domain conversational systems *viz.* XiaoIce (216) and GPT-3-based systems (217) generate semantically and grammatically rich responses. However, these open-domain counseling systems lack contextual understanding in the response generation process, which includes being unable to respond with the intended dialogue-act. Therefore, modeling this problem using open-domain dialogue systems cannot suffice the task of goal-oriented dialogue systems. To mitigate the issue in the mental healthcare domain, there is a need to harmonize VMHA’s dialogue with individual intentions to be useful for clinical practice. A very generic solution to this is to design a model that gauges the dialogue context and predicts the next dialogue-act (*aka* response-act), which collectively helps generate the next utterance.

Our approach. Our work focuses on response generation by exploiting response-acts. To understand the problem better, Figure 6.1 shows an example of a counseling conversation. As we observe, the dialogue contains both therapist’s and client’s utterances, each possessing a dialogue-act that is critical in maintaining the flow of the conversation. Evidently, the dialogue-acts of the utterances generally form a pattern. For instance, the dialogue-act of the third utterance from the therapist is *information-request*, which in succession is followed by *information-delivery* in the fourth utterance by the client.

Earlier approaches (218; 219; 111) exploited dialogue-act and context to build rich representation for several tasks on dialogue system. Another work proposes a dialogue management strategy in order to improvise on response generation task exploiting fine-grained belief states (220). These fine-grained belief states are task-specific, and their proposed system, SimpleTOD, needs relatively more information (specific slots) in order to generate a response. At the same time, exploiting the slot-filling task to generate responses supports the dialogue system in most goal-oriented cases. However, counseling conversations cannot be categorized either into open-ended or goal-oriented dialogue and hence need a separate focus on the hybrid conversational pattern. On the other hand, several studies utilize other guiding factors, such as keyword, target, etc., for the response generation task (221; 222). In our work, we predict the dialogue-act of the next utterance *aka* response-act and take advantage of state-of-the-art language models to generate relevant responses. At the same time, contextual information in the conversation plays an essential role in developing a full-fledged conversational system. To this end, we propose a response-act guided dialogue generation model named *READER*. It comprises a foundation language model, on top of which we deploy three unique heads, namely, the response-act head (RAC-Head), the language model head (LM-Head), and the value head (V-Head). These three heads jointly learn to optimize the reinforced loss and primarily perform the response generation task. *READER* learns by optimizing Proximal Policy Optimization (PPO), for which we curate a unique reward function.

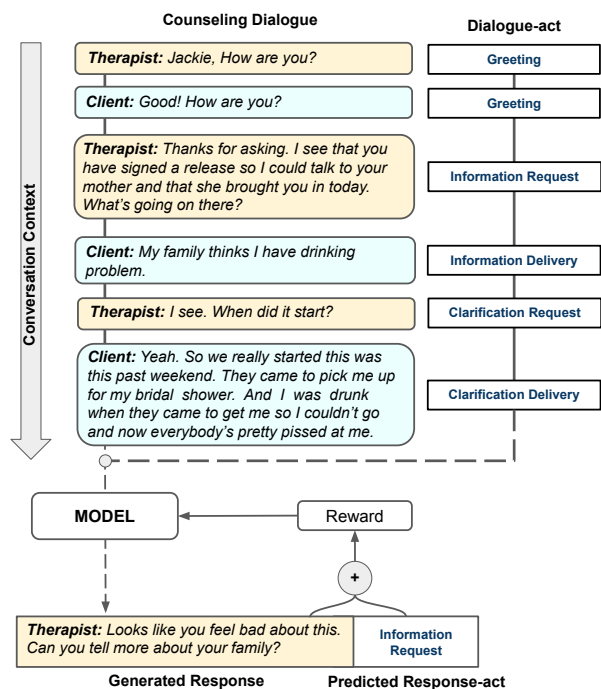


Figure 6.1: A sample counseling conversation along with associated dialogue-acts. The proposed model – *READER* takes utterance- and dialogue-act context to predict response-act and subsequently generate a response.

Evaluation. We benchmark `READER` on the HOPE dataset (1), which is a dyadic counseling conversation dataset containing 13k utterances from therapist and client. We observe that `READER` outperforms several baselines across three relevant quantitative metrics -- METEOR, ROUGE, and BERTScore, with improvements in the range of 0.82 – 11.53%. In addition, we also present an extensive qualitative and quantitative analyses of the performance, error analysis and human evaluation. Furthermore, to evaluate the generalizability of the `READER`, we benchmark it over the Switchboard Dialogue-act corpus (SWDA) (223) and obtain better results than baselines by 0.1 – 9.4%. Below, we summarize the contributions:

- We exploit future dialogue-acts (*aka* response-acts) in guiding the response generation model to generate the intended response and maintain the flow of counselling conversation in the mental-health domain. To the best of our knowledge, ours is one of the first attempts that exploits response-acts to generate precise responses in VMHAs or any other dialogue systems.
- We propose a novel transformer-reinforcement-learning (TRL) driven response-act guided model, `READER` to generate response in mental health counseling conversations.
- Our evaluation on the HOPE dataset shows significant improvements in the performance of response generation over several competing baselines. We also perform extensive ablation analysis and justify the choice of various components of `READER`.
- We conduct a thorough and qualitative human evaluation on the generated responses and establish that the proposed approach is qualitatively efficient as well.
- We also show the effective generalizability of `READER` on another dataset, i.e., the Switchboard Dialogue-act dataset.

The code of `READER` is open sourced for research purposes⁴.

6.2 Related Work

To bring more clarity in understanding the role of dialogue-acts, we present relevant studies in two broad areas – (i) dialogue-act classification, and (ii) dialogue/response generation. We intend to comprehend how dialogue-acts could bring effective innovation in building a conversational system for a dedicated task.

Dialogue-act Classification. Earlier studies by (224) and (225) employs a sparse representation of each dialogue-act in the form of triple vectors (domain-action-slot); this triple vector is represented as a one-hot encoding. However, acts become very large with the use of such sparse representations. Later, (226) addressed the issue by considering dialogue-act structures. Further, the authors represented dialogue-acts considering the act structures with level-wise vectorization on a one-hot scale where a binary classifier predicts each dimension of vectors. Their methods are further improved in a recent work by (227). The authors exploited a separate expert decoder for different areas and dialogue-acts to fuse them with a main *chair* decoder. A recent work applies a fusion approach to fuse their language model with a next utterance generation decoder (228). Several other studies use reinforcement learning to generate dialogue responses (229).

Response Generation. Studies on dialogue generation (230; 124) showed improved performance by leveraging data corpus size, which in turn resulted in learning better context-sensitive features from

⁴Code: <https://github.com/LCS2-IIITD/READER>

large language models. (231) extended this idea further by deploying models with large parameters. They used a similar idea on XLNet, a generalized autoregressive pretrained model, in order to (i) maximizing the expected likelihood over all permutations of the factorization order allowing learning of bidirectional contexts, and (ii) coping with the drawbacks of BERT by leveraging the proposed approach’s autoregressive formulation.

Later (49) explored their hypotheses of the zero-shot learning capacity of large language models as multi-task learners on the task of response generation using GPT-2. The authors also showed an intuitive qualitative analysis of a sample to fetch quality insights. The analysis shows the reflection of coherent responses to prompts. The result presents a better path toward building a response generation system that learns to perform the task from their naturally occurring demonstrations. A recent study on transformer-based models has been fine-tuned for dialogue modeling through various data modification techniques. This includes methods such as adding information about the user’s persona, masking, etc. (232).

At the same time, studies by (233) control responses using meta-words and manually controlled features (*viz.* length of response, specificity, etc.). They defined a meta-word as an organized record. The authors further described the response attributes. This allows them to model the relationship (one-to-many) within task-independent conversations and execute the problem of generating a response in an explainable and controllable manner. (220) proposed a simpler architecture which relies on the belief-states generated by the dialogue management module. These belief states are similar to fine-grained intents and slots, exploiting which the authors aimed for the dialogue-generation task. Further, (234) proposed a distributional method for handling text generation in a controllable manner by exploiting language models. This method allows point-wise specification of details and distributional constraints on the target language model in one standard framework. Their work is the first effort into this concept while minimizing relative entropy from the earlier proposed language model distribution. They uniquely defined the optimal target distribution as an explicit EBM (Energy-Based Model) representation. Moreover, using those optimal representations, we train a target-controlled autoregressive language model through an adaptive distributional variant of the policy gradient. They conducted experiments on point-wise constraints and showed the advantages of their method over traditional fine-tuning methods. Furthermore, one of the studies on dialogue modeling (235) propose to combine the merits of template-based and corpus-based DRGs by introducing a prototype-based, paraphrasing neural network, called P2-Net, which aims to enhance quality of the responses in terms of both precision and diversity. Instead of generating a response from scratch, they generate system responses by paraphrasing template-based responses. Their approach learns to separate a response into its semantics, context influence, and paraphrasing noise, and to keep the semantics unchanged during paraphrasing.

6.3 Methodology

In a regular conversation, dialogue-acts of the interlocutors tend to form a pattern. For instance, if person *A* seeks some clarification from person *B*, the most probable response from *B* would be to elucidate the clarification raised by *A*. Therefore, leveraging the above connotation, we propose to utilize the next dialog-act (or response-act) in the response generation task. Formally, we formulate the problem as follows:

Given a counseling dialogue containing utterances and their corresponding dialogue-acts as $U \in \{u_0, u_1, \dots, u_{t-1}, u_t\}$ and $DA \in \{d_0, d_1, \dots, d_{t-1}, d_t\}$ respectively, where t is the time step, our twofold jointly-learned tasks are – (a) to predict the response-act d_{t+1} (auxiliary), and (b) to generate a response u_{t+1} in the dialogue abiding by the predicted response-act d_{t+1} (primary).

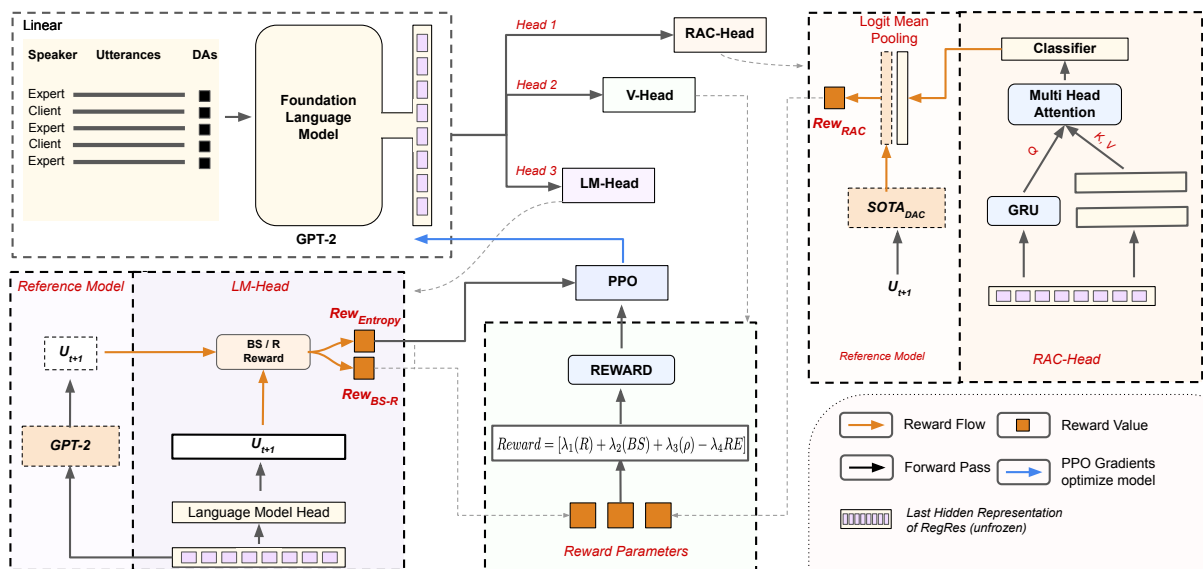


Figure 6.2: Architecture of READER. It contains three heads on top of the foundation language model, GPT-2: (a) RAC-Head classifies the response-act trained on context-aware representations, (b) LM-Head generates the response, and (c) V-Head calculates the final reward and initiates Proximal Policy Optimization (PPO).

To this end, we propose READER, a novel response-act guided reinforced response generation model. The architecture of READER is presented in Figure 7.3. READER leans on the joint transfer-reinforcement-learning (TRL) paradigm for generating response-acts and responses. Our method of transformer reinforcement learning takes inspiration from an earlier work (229). Moreover, we train the foundation language model with Proximal Policy Optimization (PPO) (236). We define a vocabulary Σ and the foundation language model θ (in our case, GPT2) that defines a probability distribution over sequences of tokens.

On top of the foundation language model, we place three task-cum-learning specific heads. First, the language model head (LM-Head) is generalized for text generation tasks. Secondly, we introduce a response-act classification head (RAC-Head), an encoder-only model to classify response-acts. At last, we have a value head (V-Head) to compute the reward to send back to the foundation model. Next, we train the model jointly to generate responses from LM-Head and predict response-acts from RAC-Head simultaneously. Subsequently, V-Head computes the reward considering the scores of LM-Head and RAC-Head, which in turn is optimized via PPO. We furnish details related to each head and the reward computation in subsequent sections.

RAC-Head. Dialogue-acts play an essential role in articulating dialogue flow. RAC-Head is a transformer-based encoder-only module on top of the foundation language model that learns to predict the future response-act. The head exploits the last hidden representations of the foundation language model. We feed the hidden representations to a GRU to exploit the contextual pattern of the dialog. In parallel, we obtain linear projections of the hidden representation. Next, these contextually rich representations are passed through a multi-head attention module in which we treat the GRU representations as the *query* and the linear projections as the *key* and *value*. Finally, we apply softmax to classify a response-act. The prediction calibrates READER to adapt the PPO optimization through the RAC-Head’s logits, thus allowing LM-Head to generate an appropriate response.

LM-Head. We use GPT-2⁵ as our foundation language model. It has been established as one of the preferred models for a variety of generative tasks (237; 238; 239; 240).

6.3.1 Reference Models for Reward Computation

We aim to augment the response by inheriting adequate semantics and response-acts. To maintain the stability of the reward function, we deploy state-of-the-art reference models to compare the outcomes for both tasks. For the language model head (LM-Head), we employ the pre-trained GPT-2 model as the reference model, whereas, SPARTA (I) is used for the response-act head (RAC-Head). Subsequently, we compute ROUGE (R), BERTScore (BS), and relative entropy (RE) between the proposed and the reference language model’s outputs. Unlike primitive methods of RL-training with a standard reward function where the model deviates to learn biased features in order to maximize the reward, where for instance, the model may start copying text from reference text to maximize ROUGE, Our approach employs each metric to calculate the reward function and tracks the relative entropy with the performance of reference model in parallel. It ensures that READER’s prediction does not deviate significantly and leverages the semantic richness of the pre-trained reference language model. We calculate the relative entropy as follows:

$$RE = \mathbb{E}_{z \sim P_{lm}} [\log P_{lm_p}(z) - \log P_{lm_{ref}}(z)] \quad (6.1)$$

where z is sampled from P_{lm} , and lm_p refers to the proposed language model; whereas lm_{ref} refers to the reference model. A lower RE score demonstrates better generations; therefore, we employ RE as a direct parameter in the reward computation.

Similarly, we utilize SPARTA (I) to compute the logit values for the predicted dialogue-act and apply mean-pooling for the reward computation.

V-Head. The value head (V-Head) is responsible for accumulating the reward parameters from other heads to yield the reward and subsequently, use it to reinforce the READER.

For our reward function, we use metrics including BERTScore and ROGUE Score. Along with these, we use known response-acts to train a reward model (SPARTA), and then optimize that reward model. Our proposed reward function accumulates the weighted Rouge score (R), BERTScore (BS), the relative entropy (RE), and the DAC-Head’s logit value (ρ). The former three components (R, BS, RE) offer feedback on the semantic and syntactic richness of the current state, while the last component (ρ) guides the model towards the desired response, exhibiting the predicted response-act. We compute the reward as follows:

$$Reward = [\lambda_1(R) + \lambda_2(BS) + \lambda_3(\rho) - \lambda_4 RE] \quad (6.2)$$

where $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are hyperparameters and tuned to optimize and maximize the reward. Subsequently, we reinforce the yielded reward to optimize the current state using PPO.

6.3.2 Training and Proximal Policy Optimization

Similar to the optimization policy explored by (229) on a general-purpose task, the training of GPT-2 with PPO in READER is a three-step process:

Initiate RAC-Head and LM-Head: Given u_t, d_t along with the context $\{ \langle u_{t-k}, d_{t-k} \rangle, \dots, \langle u_{t-1}, d_{t-1} \rangle \}$, where k is the context size, READER generates a response-act and the response.

⁵<https://openai.com/blog/better-language-models/>

Evaluate Outcomes: In this step, we calculate the log-probability distribution of logits from the active component (trainable model) of LM-Head and RAC-Head. Subsequently, we obtain the difference in the log probability distribution of the reference model, which is used to impose a penalty and ensure the coherence of the outputs.

PPO Optimization: We choose ROUGE, BERTScore, and Relative entropy to assess the quality of the generated response and the max-logit scores in case of the response-act classification task. We receive the reward score from V-Head (c.f. Equation 6.2). In order to optimize READER, we first compute the relative entropy from LM-Head. At the same time, LM-Head and RAC-Head yield the remaining reward parameters. At last, V-Head accumulates and computes the reward from each head as per Equation 6.2. We perform optimization that subsequently allows READER to learn to penalize/reward the foundation language model.

This approach is extended from an earlier work (229). The authors initialized a policy $\pi = \theta$, and then fine-tuned π to operate on downstream tasks using PPO. If the task is defined by a reward function ($r : X \times Y \rightarrow R$), then the authors used PPO to optimize the expected reward. In their algorithm, PPO utilizes clipped surrogate objective, and the model maximizes a surrogate objective. Another study (229) exploited the usage of PPO algorithm to further define the downstream task to optimize the main objective function. The authors opted for minimum of the clipped and unclipped objective. Hence the final objective is lower bound (i.e., a pessimistic bound) on the unclipped objective. With this scheme, we observe that only probability ratio is ignored when it improves the objective.

As a result, we then exploit the above mentioned PPO method to optimize our PPO algorithm (229) with the following equation.

$$R(x, y) = r(x, y) - \beta \cdot \log \pi(y|x) \cdot \theta(y|x) \quad (6.3)$$

where r and θ represent reward function and foundation language model, respectively. In our case, we experiment with a constant as well as dynamic β to achieve a favorable value of $RE(\pi, \theta)$. The relative entropy plays the role of an entropy bonus; it prevents the policy from moving too far from the range where r is valid. We rely on the relative entropy to sync with the fine-tuned reference model’s coherent responses.

6.4 Experiments

In this section, we first discuss the counseling dataset. We then define the baseline systems and evaluation metrics which we use to compare the performance of the proposed model and baselines.

6.4.1 Dataset

We use HOPE (1), a mental health counseling conversation dataset. It contains 12.8K utterances from 212 dyadic counseling sessions between therapists and clients, publicly available on a video sharing platform. The conversation encompasses diverse demographic groups with distinct mental health discussions. (1) extracted transcriptions of the utterances and processed them to remove any noise and/or transcription issues. The collected dialogues are dyadic in nature, i.e., clients and therapists are the only interlocutors. Each utterance in the HOPE dataset is annotated with one of the twelve dialogue-acts – *information-delivery* (ID), *information-request* (IRQ), *yes/no-question* (YNQ), *clarification-request* (CRQ), *opinion-request* (ORQ), *clarification-delivery* (CD), *positive-answer* (PA), *negative-*

		R1			R2			RL			BS	METEOR
		P	R	F1	P	R	F1	P	R	F1		
Baselines	DialoGPT (241)	12.34	40.48	15.72	2.92	11.83	4.42	12.23	38.60	15.76	0.7603	0.2021
	GPT2 (49)	12.70	32.63	14.98	3.08	7.92	3.51	13.74	32.05	15.87	0.7445	0.1754
	DialogVED (242)	12.48	31.74	12.8	0.98	2.45	1.22	12.45	31.11	14.46	0.7189	0.2000
	ProphetNet (243)	12.15	34.29	14.48	3.30	10.41	4.17	12.24	33.12	15.18	0.6707	0.1901
	VHCR (246)	11.29	21.33	11.81	2.66	3.49	3.00	10.01	19.72	10.99	0.5953	0.1041
	HRED (244)	11.52	21.51	10.72	1.89	6.42	2.92	12.12	24.36	13.56	0.6259	0.1425
	HRED w/ Sp. Utt. Encoder (245)	11.77	28.63	10.08	1.29	4.19	2.06	12.25	21.27	12.72	0.6171	0.1801
Ours	READER w/ DialoGPT	12.41	43.91	16.12	3.70	13.72	4.98	11.92	41.02	16.30	0.7656	0.2098
	READER – RAC-Head	12.64	41.48	15.78	3.60	11.83	4.58	12.3	38.64	15.90	0.7628	0.2039
	READER	12.82	43.93	16.15	3.77	13.67	4.93	12.51	40.82	16.32	0.7666	0.2103
Ablations	– Rew(R)	11.73	38.82	14.65	2.28	8.45	2.96	11.21	35.76	14.53	0.7561	0.1840
	– Rew(RAC)	12.36	40.71	15.43	3.13	11.12	4.06	11.91	37.63	15.40	0.7609	0.2000
	– Rew(RAC + R)	11.92	38.06	14.70	2.43	8.26	3.11	11.40	34.98	14.58	0.7530	0.1874
	– Rew(R + BS)	12.48	41.13	15.57	3.52	11.85	4.47	12.22	38.29	15.77	0.7527	0.2092
	– Rew(RAC + BS)	12.01	40.45	15.18	2.72	9.93	3.52	11.46	37.05	14.97	0.7577	0.1908
$\Delta_{\text{READER}-\text{BEST}}(\%)$		$\uparrow 0.94$	$\uparrow 8.5$	$\uparrow 2.73$	$\uparrow 14.24$	$\uparrow 15.50$	$\uparrow 11.53$	$\downarrow 8.90$	$\uparrow 5.69$	$\uparrow 2.83$	$\uparrow 0.82$	$\uparrow 4.05$

Table 6.2: Results obtained on the HOPE dataset. We show ROUGE (1, 2, L), BERTScore (BS), and METEOR to assess the performance of the READER. $Rew(x)$ is the reward function, where x is the parameter.

answer (NA), opinion-delivery (OD), greeting (GT), acknowledgment (ACK), general chit-chat (GC).

A detailed statistics of the HOPE dataset is presented in Table 7.1. The dataset shows a high correlation between certain pairs of dialogue- and response-acts. For instance, an utterance requesting information (labeled IRQ) is mostly followed by an utterance delivering the information (labeled IRD).

HOPE	Train	Validation	Test	Total
Dialogue Sessions	149	21	43	212
Client Utterances	4668	595	1119	6382
Therapist Utterances	4751	599	1122	6472
#Total Utterances	9419	1194	2241	12854

Table 6.1: Statistics of the HOPE dataset (1). The dyadic counseling conversational dataset contains a total of 12.8k utterances, each associated with one of the 12 dialogue-act labels.

6.4.2 Baselines and Evaluation Metrics

We compare READER’s performance with various competitive baselines in the domain of dialogue generation. Moreover, to have a fair comparison with READER, we choose systems which leverage and exploit the dialogue context for the response generation. To the best of our knowledge, no other systems have reinforced the response-act for the dialog generation. We choose the following baselines in this work. **DialoGPT** (241) is a pretrained transformer model dedicated for response-generation task. **GPT-2** (49) is a decoder only model trained on a large corpora. A vanilla finetuned version of GPT-2 works well in our use case. **DialogVED** (242) introduces continuous latent variables into the encoder-decoder pre-training framework to increase the relevance and diversity of responses. **ProphetNet-Dialog** (243) focuses on pretraining of dialogue specific corpus to generate coherent response. **HRED** (244) is based on generative modeling to develop conversational response containing hierarchical encoder-decoder paradigm. It is trained on a large dialogue corpus for the utterance generation task. **HRED with Speaker and Utterance Encoder** (245) adds speaker and utterance level information to the hierarchical encoder-decoder (HRED) setup. It leverages the personalization parameters in a dialogue system. It is also trained on a large dialogue corpus. **VHCR** (246) uses variational hierarchical RNNs for the conversation-only setup. It is trained on a large conversation corpus for the dialogue modeling task.

For evaluating the performances of `READER` and other comparative systems, we employ **ROUGE**, **METEOR**, and **BERTScore** as evaluation metrics. We use `py-rouge`⁶, `nltk-meteor`⁷ and `Hugging Face - BERTScore`⁸ for computing the scores.

6.5 Results and Analysis

In this section, we discuss the results obtained from the `READER` model and the aforementioned baselines. Table 7.2 summarizes the comparative and ablation results on the HOPE dataset.

6.5.1 Performance Comparison

Our evaluation shows superior performance of `READER` across a majority of the metrics. Evidently, there is a significant increase in the recall of the ROUGE-2 score – our model receives a ROUGE-2 recall of 13.67, which is +15.50% as compared to the second best performer, DialoGPT (11.83). At the same time, our model yields 43.93, 40.82, and 76.66 scores of ROUGE-1 recall, ROUGE-L recall, and METEOR, respectively, with an increase of +3.45, +2.22, and + 0.63 points as compared to the best baseline, i.e. DialoGPT. On the other hand, to evaluate the linguistic properties in the generated utterances, we calculate METEOR on `READER`'s generations. Similar to the earlier cases, `READER` reports an improved BERTScore of 0.2103, +4.05% points better than DialoGPT.

Ablation on Foundation Model. Among all baselines, DialoGPT performs the best on average with GPT-2 closely competing with it. However, in our case, we choose GPT-2 as the foundation language model due to the marginally better performance in `READER`. We also experiment by swapping it with DialoGPT and report the results in Table 7.2. We observe that `READER` with GPT-2 performs better on recall scores of R1 (+0.41), R2 (+0.02), RL (+0.2), BS (+0.0139) as compared to `READER` with DialoGPT.

Model Component Ablation Study. One of the prime contributions of this work involves the role of the RAC-Head. `READER` relies on the RAC-Head to determine the response-act of the generated utterances. Further, RAC-Head gradually allows the response generation to adapt according to the predicted response-acts during the PPO optimization. We perform ablation on RAC-Head and present the results in Table 7.2. We observe that `READER` without RAC-Head scores 41.48, 11.83, 38.64 on ROUGE-1, 2, L (recall scores), respectively. The final model performs relatively better with an increased score (+2.45, +1.84, +2.18) on the same metrics as compared to the `READER` without RAC-Head.

Discussion on Reward Selection and Reward Ablation. We meticulously conduct experiments on several hypotheses to design a reward function that optimizes the PPO policy and penalizes the model for every shortcoming. While experimentation, we consider several metric scores as a parameter to the reward function (c.f. Equation 6.2). However, most of the parametric configurations deteriorate the results. We show various possible ablations on the final set of parameters, i.e., ROUGE, BERTScore, and RAC-Head's logits in the lower half of Table 7.2. We observe that a combination of RAC-Head along with BERT Score, ROUGE scores, and relative entropy yields the best result. In addition, using only BERTScore or ROUGE scores in the reward function deteriorates the results significantly. We also observe a decrease in the metrics after detaching RAC-Head. This justifies the contribution of both the heads of `READER` toward the generation task. Moreover, Figure 6.3 shows the incremental graph of the reward function

⁶<https://pypi.org/project/py-rouge/>

⁷https://www.nltk.org/api/nltk.translate.meteor_score.html

⁸<https://huggingface.co/spaces/evaluate-metric/bertscore>

Context	Next utterance	RAC
Therapist: Oh great, which is a little different. But yeah, that's great. That's a really good goal.	Gold: Okay, so you're getting A's and B's in really good numbers. So what did you discover about this?	CRQ
Client: Yeah, thanks.	READER: Okay. Okay. What classes are you not getting A's and B's in?	CRQ
Therapist: It's hard work to get into law school. Are you doing well in school? What kind of grades you get?	GPT2: Anyway, I should mention I'm probably going to get through that anyway. That might not be fun.	-
Client: I'm doing okay. I'm getting mostly A's and B's so I'm pretty good.	DialoGPT: Are you going to law school?	-
Client: My employer just said I wasn't a good fit.	Gold: My anxiety is really bad.	ID
Therapist: wasn't a good fit. So this so these symptoms have caused you some real problems.	READER: Oh, I get really anxious. just scared that something's going to happen and your anxiety is going to be costly.	ID
Client: Yeah.	GPT2: That's it. Lerner. It's a honest fact and it is one piece of information about your mind. Lerner	-
Therapist: With employment and how about with your anxiety level?	DialoGPT: I'm not sure if you're joking or not.	-

Table 6.3: An example showing the generated response along with the reference response.

during the PPO optimization. Evidently, the plot demonstrates the progressive learning curve considering the mean reward score, and supports our claim that the model improves with the increasing step.

6.5.2 Qualitative Analysis

To further assess the quality of the generated responses of `READER`, we present a qualitative analysis in Table 7.3. In comparison with the reference utterance, we observe that semantics and grammar are well established in the generated response. Additionally, we observe that the response-act of the generated response is in accordance to the intended act. It suggests the effectiveness of `RAC`-head in reward computation. We also show generated utterances for the two best performing baselines (`DialoGPT` and `GPT`). Though the outputs are syntactically correct, they are mostly incoherent with the dialogue context.

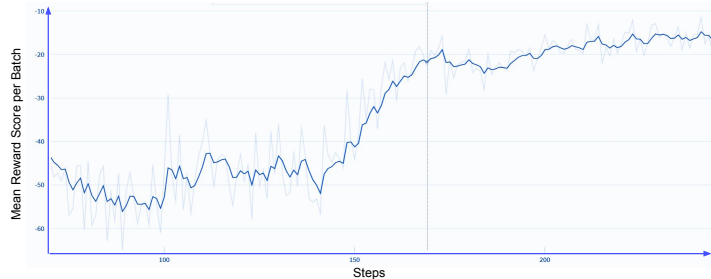


Figure 6.3: The increment flow of reward in the PPO of `READER`.

Human Evaluation We also perform human evaluation on a subset of model outcomes on linguistic ground. We use four linguistic parameters, namely, *relevance*, *consistency*, *fluency*, and *coherence*, to perform the human evaluation (247). We define these parameters as follows: **Fluency** demonstrates the linguistic quality of the generated responses; **Coherence** shows the structure and organization of the generated responses; **Relevance** shows the selection of relevant content in the generated response considering the reference utterance; and **Consistency** evaluates the factual alignment between the generated response and the source utterance. Evaluators are blind to the source of each dialogue, with responses shown in random order and no indication of their origin; only the source utterance and corresponding response were provided.

In total, we take 50 randomly-selected instances and ask 10 human evaluators to assign a score on a scale of [1, 5] to each of the four parameters, where 5 represents the best outcome. All human evaluators are linguistic experts, aged between 20 to 35. For comparison, we repeat the exercise for `DialoGPT` and `GPT2` as well. Finally, we compute the average score and report the findings in Table 6.4. Our analysis

shows that `READER`'s outputs are also qualitatively better than baselines in each dimension.

6.5.3 Application of `READER`: Dialogue Generation

In this section, we present the application of `READER` for generating counseling dialogues. To do so, we adopt two setups: **a) Natural:** an end-to-end conversation between a client and an agent (`READERTherapist`); and **b) Synthetic:** an end-to-end conversation between two agents, i.e., `READERTherapist` and `READERClient`. The first setup is a natural configuration for VMHAs, when deployed at the application stage, it generates therapist utterances to interact with real-time clients having mental health issues. To

do so, at every step i of the response generation, we provide actual client inputs and the previously-generated `READER`'s outputs for therapist ($\{i - n, \dots, i - 2, i - 1\}$) as the recurring context to `READER`. On the other hand, the second setup is an analysis configuration to assess the effectiveness of `READER` in handling diverse inputs (e.g., generated by an agent). Moreover, this can also be viewed as a data augmentation technique to generate synthetic dialogues. In this setup, we provide `READER`'s generated outputs for both client and therapist as context.

Furthermore, in both setups, we assume that a context is present to instigate the conversation, such that the agent (or `READER` in our case) understands the dynamics of the conversation and starts generating responses that are aligned with the conversation. This approach is similar to existing VMHAs, like `WoeBot`, where the agent collects initial information from the client in terms of template-based questions and propels the conversation further with the provided details. Evidently, we observe that the proposed model is able to comprehend the context of the conversation in both setups and generates aligned responses as shown in Figure 6.4 and Figure 6.5.

6.5.4 Generalizability

`READER` outperforms several baselines across most of the metrics on the HOPE dataset. Further, to assess the model's generalizability, we extend our experiments and evaluate `READER` on the Switchboard Dialog-act corpus (223). We observe that `READER` improves the performances of two best performing baselines (`DialoGPT` and `GPT2`) by 0.1% – 9.4% in 10 out of 11 metrics. In particular, we observe a significant improvement of 9.4% in `BERTScore`; thus suggesting that `READER`'s outputs are semantically richer than other baselines along with the marginal improvements in textual similarity. We argue that in the presence of the information of dialogue-acts, `READER` harnesses the context in an efficient way for generating semantically-rich responses.

Model	Relevance	Consistency	Fluency	Coherence
DialoGPT	2.11	2.42	2.90	2.30
GPT2	2.70	3.00	3.01	2.44
READER	2.85	3.05	3.05	2.95

Table 6.4: Human evaluation on the responses generated from `READER` when compared to the top two best-performing baselines. We observe that the performance of `READER` across all metrics is up to the mark and slightly better than the best-performing dialogue models.

6.6 Discussion

Societal Impact and Deployment. Our work acts as a support to the mental health community and ongoing research by leveraging the advancements in AI-based dialogue systems for counseling. Such advancements in the mental health domain are likely to bring a high social impact by directly assisting experts and providing on-demand availability to help-seekers.

Ethical Considerations and Future Work. Considering the severity of the research area, we make sure that at each step, we maintain the privacy of the personal data of clients. In the future, we plan to extend our work in the expansion of Virtual Mental Health Assistants (VMHAs) modules and scale the idea of including dialogue components such as empathetic understanding.

6.7 Summary

The continuous need to face the shortage in the number of mental health experts is becoming a significant challenge every coming year. With new AI-based therapist-bots coming into the picture, clients receive much support with ease of access. However, generation of an accurate yet effective response for the clients with intended dialogue-act remains unexplored. To this end, we proposed a novel response-act guided dialogue generation model, *READER*. We designed a unique reward function that exploits several linguistic properties to train the model using transformer-reinforcement learning (TRL) and further improvised the PPO optimization. We added three heads on top of the foundation language model: RAC-Head, Value-Head, and LM-Head, which collectively curate the reward. We compared the performance of *READER* with several baselines. Our model outperformed several baselines. At last, we demonstrated an extensive ablation study and concluded the paper with a discussion on ethical considerations and generalizability.

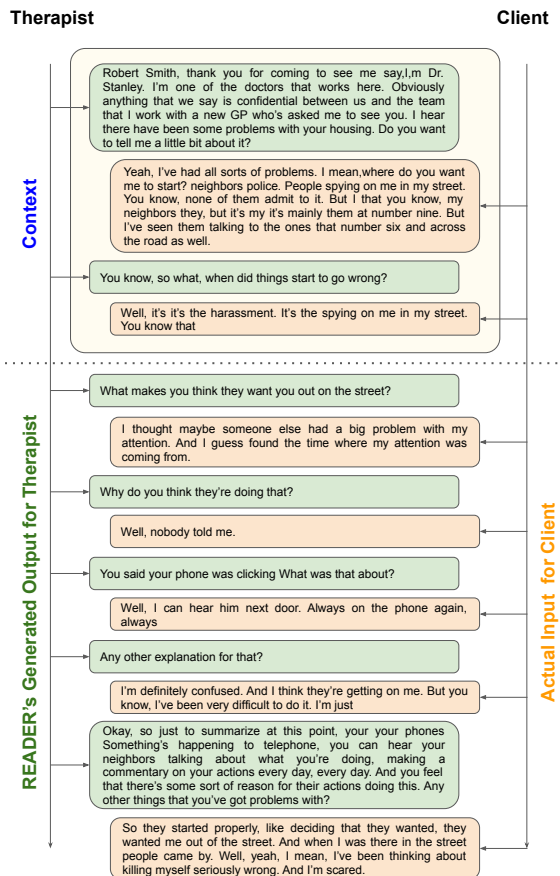


Figure 6.4: Application of READER – Natural setup. Given a context, at each step, READER generates an output for the therapist. This aligns with the natural configuration of VMHAs, where a client seeks help from a bot or a virtual agent.

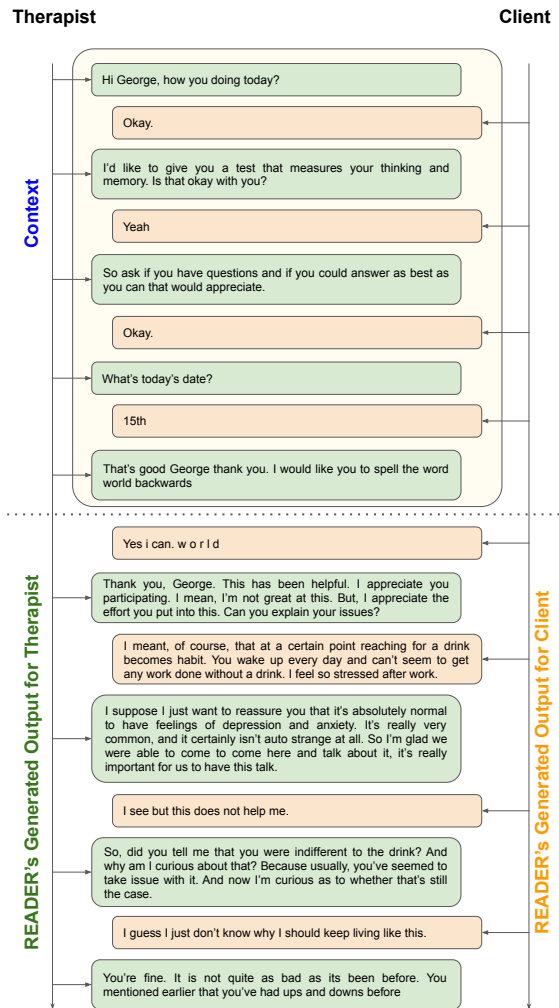


Figure 6.5: Application of READER – Synthetic setup. Given a context, at each step, READER takes turns to generate outputs for the therapist and the client. This setup is an analysis configuration to assess the effectiveness of READER in handling diverse inputs (e.g., generated by an agent).

7. Commonsense Guided Response Generation

The crisis of mental health issues is escalating alarmingly. Effective counseling serves as a critical lifeline for individuals suffering from conditions like PTSD, stress, depression, etc. Therapists forge a crucial therapeutic bond with clients, steering them towards positivity. Unfortunately, the massive shortage of professionals, high costs, and mental health stigma pose significant barriers to consulting therapists. As a substitute, Virtual Mental Health Assistants (VMHAs) have emerged in the digital healthcare space. However, most existing VMHAs lack the commonsense to understand the nuanced sentiments of clients to generate effective responses. To this end, we propose EmpRes, a novel sentiment-guided mechanism incorporating commonsense awareness for generating responses. By leveraging foundation models and harnessing commonsense knowledge, EmpRes aims to generate responses that effectively shape the client's sentiment towards positivity. We evaluate the performance of EmpRes on HOPE, a benchmark counseling dataset, and observe a remarkable performance improvement compared to the existing baselines across a suite of qualitative and quantitative metrics. Moreover, our extensive empirical analysis and human evaluation show that the generation ability of EmpRes is well-suited and, in some cases, surpasses the gold standard. Further, we deploy EmpRes as a chat interface for users seeking mental health support. We address the deployed system's effectiveness through an exhaustive user study with a significant positive response. Our findings show that 91% of users find the system effective, 80% express satisfaction, and over 85.45% convey a willingness to continue using the interface and recommend it to others, demonstrating the practical applicability of EmpRes in addressing the pressing challenges of mental health support, emphasizing user feedback, and ethical considerations in a real-world context.

7.1 Introduction

Mental health issues are on the incline, demanding immediate attention. However, the accessibility to mental health professionals remains limited, exacerbating the challenges faced by individuals seeking support. In fact, according to a recent report by Mental Health America 2023, out of the staggering 50 million people affected by mental health issues, over 55% did not receive the necessary treatment, highlighting the magnitude of the problem¹. To address the gap of experts and meet the requirements of support-seekers, virtual mental health assistants (VMHAs) come to the rescue.

The VMHA's ultimate goal is to provide effective responses that positively influence clients' sentiments to lead toward healthy counseling practices and ensure client stability throughout sessions. Studies suggest prioritizing control over clients' sentimental feelings as the key focus for counselors working with new clients displaying negativity (248). In a real-world setting, mental health experts proficiently employ their commonsense knowledge, along with active listening and validation skills, to understand clients'

¹<https://mhanational.org/issues/state-mental-health-america>

experiences and subsequently provide effective responses during counseling. To mimic such a counseling approach, our research aims to explore the benefit of commonsense knowledge as a means to aid VMHAs in understanding clients. To explain this better, a sample of the counseling session is shown in Figure 8.3. As we observe, each counseling utterance is tagged with a sentiment label, which is utilized by the knowledge extractor in a way such that the mental health assistant generates a response that influences the client positively. Several VMHAs, including notable ones like Wysa² and Woebot³, are designed to function as mental health professionals. However, they currently lack the capability to primarily address clients’ sentimental states and respond using commonsense knowledge within their conversational methodologies.

Given the long history of NLP and mental health studies, such as on social media (249; 250; 251), linguistic behavioral coding (252), and virtual counseling agents (253; 254; 255), there has been a significant leap in the conversational research space for multiple research applications. Presently, open-domain conversational systems such as ChatGPT (256), XiaoIce (216), and GPT-3-based variants (217) excel in generating responses that are linguistically accurate and semantically coherent. However, in the sensitive context of counseling, these open-domain systems lack the capability to sufficiently meet the requirements (257). On the other hand, certain approaches have focused on utilizing conventional methods to generate responses for mental health counseling (258; 259; 260). More recent works have researched into controlling the response generation methodologies with factors like dialogue-acts (261), empathy (176; 175), target (222), and related aspects. These conventional methods were seen to be helpful only for a handful of specific cases, leaving room for dedicated research in the space of mental health counseling to tackle the nuanced demands of VMHAs. To address these gaps, we propose EmpRes, a sentiment-guided commonsense-aware response generation mechanism for mental health counseling. We use a commonsense transformer to exploit commonsense knowledge and subsequently influence the client’s sentiments in such a way that – (i) if the client possesses a negative sentiment, EmpRes’s response influences the client toward the positive sentiment, and (ii) if the client is already in a positive state, EmpRes’s response should keep the client in the same sentiment state. This approach is motivated by the analysis of psychiatric conversations (262), where

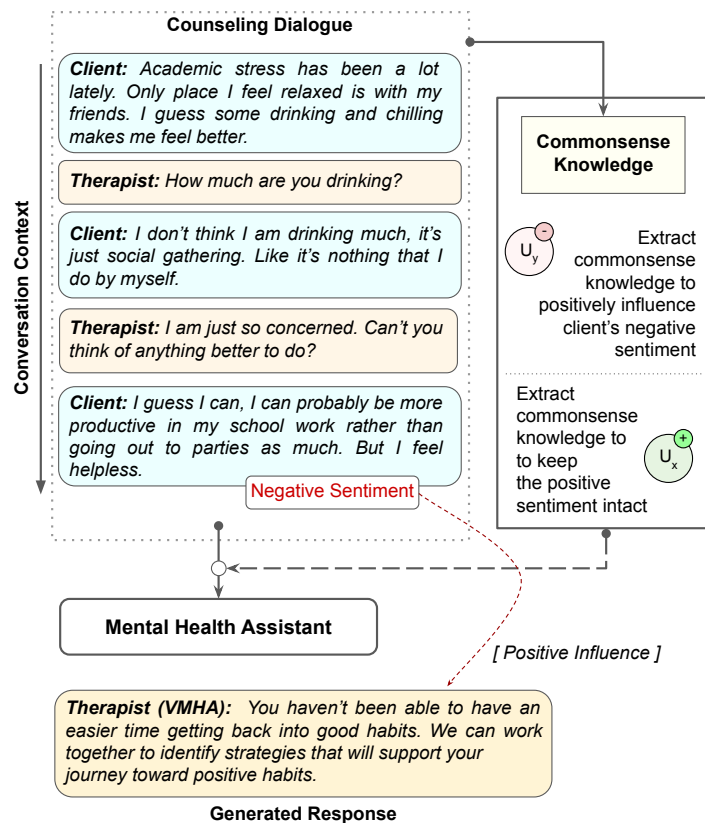


Figure 7.1: A sample counseling conversation in which the client expresses negative sentiment at last; the job of a VMHA is to generate a response to positively influence the client and contains rich commonsense knowledge.

²<https://wysa.com>

³<https://woebothealth.com>

clients express complex (mostly negative) sentiments, and the therapeutic process aims to guide them toward a more positive state (263). At last, we employ GPT-2 with modified knowledge-aware attention to learn commonsense and sentimental influence. As a result, the proposed system helps stabilize a client’s sentimental state, making them more receptive toward subsequent treatment.

We evaluate EmpRes using the HOPE dataset, a recognized benchmark in counseling (1), and compare its performance against 12 potential baselines. Our approach demonstrates significant superiority over all the baselines across five key metrics. Furthermore, through empirical analysis and human evaluation, we ascertain that EmpRes’s generation capability is well-aligned with expectations, even surpassing the gold standard in certain instances. Additionally, we deploy EmpRes as a chat interface for individuals seeking mental health support, evaluating its effectiveness through a comprehensive user study in a constrained setting. Findings indicate a notable positive response, with 91% of users acknowledging the system’s effectiveness, 80.25% expressing satisfaction, and over 85% expressing a willingness to continue using and recommending the interface, thereby underscoring the practical utility of EmpRes. We summarize our contributions below:

- We propose EmpRes, a novel transformer-based model for therapist response generation. It prioritizes the central objective of enhancing the therapeutic experience by influencing the client’s sentiment toward the positive side.
- We exploit commonsense knowledge to understand clients better and construct sentiment-guided and commonsense-aware graph representations for counseling dialogue.
- Our evaluation on the HOPE dataset shows EmpRes’s improved performance in generating therapist responses across 12 competing baselines on five text-generation metrics.
- Our thorough qualitative analysis and human evaluation demonstrate EmpRes’s efficacy, which in some cases, surpasses the gold standard.
- We deploy EmpRes-assistant as a chat interface for individuals seeking support. We conduct an exhaustive user study to evaluate the deployed system’s effectiveness.

7.2 Related Work

Commonsense Knowledge for Dialogue Generation. The positive outcomes of exploiting commonsense knowledge in AI and, more precisely, in NLP have been well studied (264). Most of the early research primarily focused on incorporating commonsense knowledge in improving various aspects of text generation, including textual inference, story comprehension, and generation (265; 266; 267). Commonsense is vital for generating coherent, contextually appropriate, and plausible text by providing background information and logical connections between ideas. As research advanced, there was a growing interest in conversational systems and domain-specific assistants, and therefore, there was a need for applying commonsense to dialogue generation for more relevant responses. (268) highlighted the problem of neural text degeneration, in which dialogue models tend to generate generic and uninformative responses. It supports the argument that there is a need to incorporate commonsense knowledge to avoid such degeneration. Recent advancements in dialogue generation leveraging commonsense knowledge are using commonsense transformer (COMET) (269). COMET is a generative model that possesses rich commonsense understanding by using deep pretrained language models. Using ATOMIC and Conceptnet, which are external knowledge graphs, authors capture different aspects of commonsense knowledge. Several studies (39; 270; 271) highlighted the use of the COMET model in different contexts, such as counseling, Chinese dialogue and emotional support. These models seek to enhance dialogue responses by

	Utterances	Positive	Negative
Therapist Utterances	6472	4569	1903
Client Utterances	6382	2540	3842
#Total Utterances	12854	7109	5745

Table 7.1: Statistics of the HOPE dataset (I). The dyadic counseling conversational dataset contains a total of 12.8k utterances, each tagged with positive/negative sentiment.

integrating external commonsense knowledge, but exploiting commonsense knowledge for text generation is still relatively nascent. Challenges include acquiring, representing, and effectively integrating external commonsense knowledge, making the generation process more complex.

Response Generation. Response generation is a fundamental aspect of dialogue systems, in which the system generates appropriate and contextually relevant responses based on user input. Several techniques, along with commonsense to control response generation, are available. A few earlier works have highlighted the importance of incorporating emotion into response generation for empathetic and personalized dialogue systems (272; 273; 274). By incorporating techniques such as selecting appropriate exemplars based on emotional content, recognizing emotional causes in conversations, and considering both user sentiment and desired emotional tone, they can use emotion in their response. Additional research works considered response generation as a sequential decision-making problem (275; 276). This approach consists of a dialogue agent that learns to generate responses by interacting with an environment, receiving rewards based on the quality of its responses, and updating its policy through trial and error. By leveraging reinforcement learning algorithms, dialogue agents can optimize their response generation strategies to improve dialogue quality and achieve specific objectives, such as coherence, informativeness, or engaging conversations. On the other hand, there exist works that employ graph structure to capture crucial aspects of conversation context, including self-other awareness, causality reasoning, and multi-party dependencies, which account for participant relationships, causal dependencies, and diverse contextual information, leading to a contextually relevant and engaging dialogue (277; 278; 279). This motivated us to design a sentiment-guided commonsense relation graph.

7.3 Dataset

We use HOPE (I), a benchmark counseling conversation dataset. It contains 12.8K utterances from 212 dyadic counseling sessions between therapists and clients, publicly available on social platforms. The conversation in the dataset encompasses a wide range of demographic groups, each engaging in unique discussions related to mental health. To ensure data quality, earlier work was conducted to thoroughly process the transcriptions, eliminating any instances of noise or transcription errors (I). The dialogues collected for further analysis adopt a dyadic structure involving only two participants – clients and therapists. To enhance the dataset, we apply a pseudo-labeling approach, tagging each utterance with its corresponding sentiment label. A detailed statistic of the HOPE dataset is presented in Table 7.1. We exploit the popular BERT-based fused with commonsense knowledge for the sentiment classification model to perform pseudo-labeling⁴. This auxiliary binary classification task enables us to tag 12.8K utterances with sentiment labels. In our analysis, we observe an uneven distribution of emotion labels (as it should be), particularly between positive and negative sentiments. While therapists predominantly maintain a neutral stance in most cases, the binary nature of emotion labels leads to a considerable presence of negative sentiment labels as well. We discuss the additional details about pseudo labeling in the next section.

⁴nlptown/bert-base-multilingual-uncased-sentiment

7.3.1 Dataset Extension

For our research, we utilized the HOPE benchmark counseling conversation dataset. This dataset comprises 212 counseling conversations, amounting to approximately ~ 13 utterances. To annotate the dataset with sentiment labels, we employ pseudo-labeling by utilizing both the commonsense transformer (COMET) and a BERT-based sentiment classification model. The central idea behind our approach is to utilize additional attributes derived from each utterance to enhance the performance of the sentiment classifier. Specifically, we used COMET to extract commonsense attributes related to the utterances, which we then integrated into the sentiment classification process. This integration aimed to provide the classifier with more contextual information, subsequently improving its accuracy. After conducting iterative experiments with various foundation models for sentiment classification, we observed that our approach performed marginally better than more sophisticated models. This led us to opt for a simpler and lighter foundation model, which balanced performance with efficiency.

The overall dataset extension pipeline is illustrated in Figure 7.2. In this pipeline, the utterance encoder first processes the input utterance. Subsequently, COMET retrieves the $xAttribute$ relations for the utterance, providing a set of attributes that describe likely properties or characteristics associated with the utterance. To ensure robustness, we selected the top-5 attributes from COMET for each utterance. These selected attributes were then concatenated with the original utterance using a $[SEP]$ tag, creating an enriched input that was passed through a feed-forward layer for final classification.

7.3.2 Discussion on COMET Relation Selection

The selection of commonsense relations plays a crucial role in our model’s understanding of the sentiment and intent of the clients. To assess the impact of different relations, we individually ran the commonsense transformer (COMET) on a sample of the HOPE dataset. For this purpose, we employ all nine COMET’s relations – $oEffect$, $oReact$, $oWant$, $xAttr$, $xEffect$, $xIntent$, $xNeed$, $xReact$, and $xWant$. We observe that the $xAttr$ relation provides valuable insights into the specific attributes or characteristics related to the client’s utterance. By analyzing this relation, we can determine whether the client is experiencing positive or negative emotions, which aids in sentiment classification. Hence, we employ such relation in the sentiment classification task as discussed in the Section 7.3.1

At a moment, a client can be found in one out of two possible states – (i) positive or (ii) negative. In order to positively influence the client, we formulate the problem to fetch the commonsense knowledge in such a way that it conditions the knowledge extractor based on the client’s last sentimental state. In general, if a client falls into a negative sentimental state, others (in our case, the therapist) consoles the client and influences them toward a positive side. Hence, we employ $oReact$ and $oWant$ relations effectively to capture what others would want for the client and how they feel about the client, respectively. These relations demonstrate a positive, helpful, and sentimental understanding of the client’s

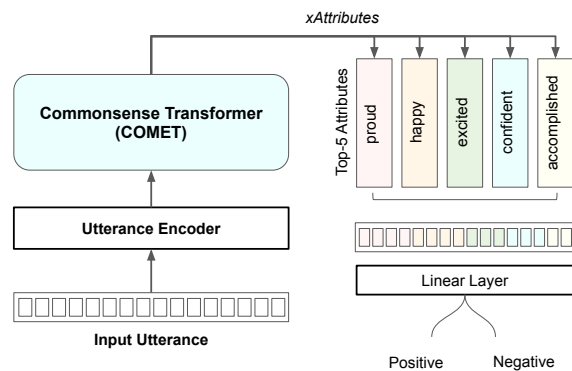


Figure 7.2: A schematic diagram of pseudo labeling for sentiment labels for each utterance in the HOPE dataset. We augment top-5 additional attributes using COMET to assist state-of-the-art sentiment classifier in predicting sentiment label.

negative sentiments. On the other hand, if a client feels positive, we intend to keep the client in the same state only. In addition, whatever client’s thought is bringing positivity, we further exploit it in our response via three specific and meticulously analyzed relations – $xReact$, $xWant$, and $xIntent$. These relations augment the self-feelings and self-thoughts, which further assists the VMHAs in generating responses capable of embedding similar thoughts and feelings into the response.

Overall, the selection of commonsense relations plays a significant role in capturing the nuances of sentimental influence in our model. By leveraging the most relevant relations, we ensure a more accurate representation of clients’ sentimental states and enable our model to provide supportive responses.

7.4 Methodology

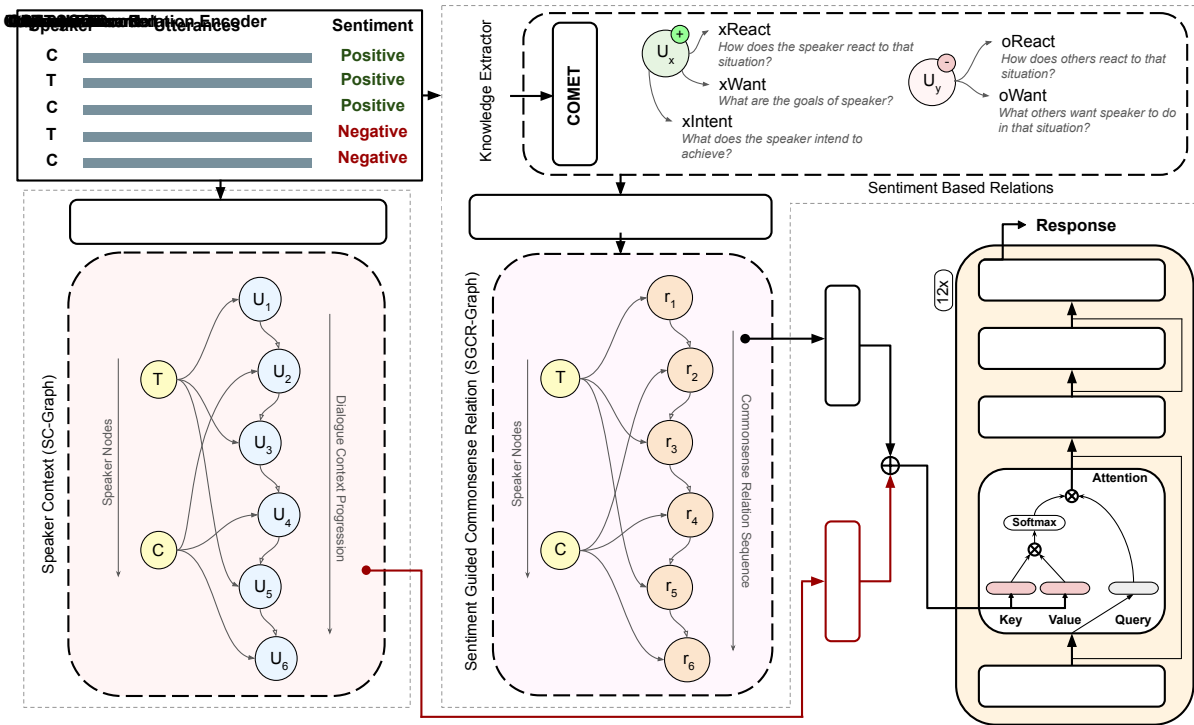


Figure 7.3: A schematic diagram of EmpRes. *Knowledge Extractor* exploits conditional commonsense relations (r_i) to construct Sentiment-Guided Commonsense Relation Graph (SGCR-Graph). A graph attention layer (GAT-SGCR), on top of SGCR-Graph, formulates sentiment-guided commonsense-aware representations. A similar layer on Speaker-Context Graph (SC-Graph), called GAT-SC, is responsible for formulating dialogue representations. Fused graph representations act as key and value for GPT-2’s attention block, responsible for therapist response generation.

In a de facto setting, clients and virtual assistant exchange utterances. During this conversation, clients tend to express their sentiments, and VMHAs are expected to utilize commonsense to gain a clear understanding and subsequently generate responses.

Problem Formulation. Suppose there is a counseling dialogue between a therapist and a client containing utterances and their sentiment labels as $U \in \{u_0, u_1, \dots, u_t\}$ and $S \in \{s_0, s_1, \dots, s_t\}$ respectively, with t being the time step. Here, sentiment labels are either *positive* or *negative* along speaker labels, $S_p \in \{T, C\}$, corresponding to each utterance. Our twofold jointly-learned tasks include (i) exploiting

COMET to generate sentiment-guided commonsense-aware dialogue representations (auxiliary), and (ii) generating an effective therapist response u_{t+1} against a client’s utterance with an intent to keep the client with positive sentimental feeling (primary).

To this end, we propose `EmpRes`, a novel commonsense aware sentiment-guided response generation mechanism for VMHAs. `EmpRes` integrates the effectiveness of commonsense knowledge and sentiment guidance to place the utmost importance on fostering positive sentimental states in clients. Figure 7.3 illustrates the schematic framework of `EmpRes`. `EmpRes` generates the therapist’s response against the client’s last utterance. Initially, we construct a Speaker-Context Graph (SC-Graph) utilizing BERT to capture comprehensive representations of utterances and their relations with speaker information. On the other hand, the *Knowledge Extractor* module leverages a commonsense transformer (in our case, COMET) to extract relevant commonsense knowledge, taking into account the associated sentiment labels. To maintain a positive sentiment within the client, we conditionally employ distinct commonsense relations for utterances with positive and negative sentiments. This conceptual framework enables us to incorporate sentiment-guided knowledge effectively, thereby assisting in the construction of the Sentiment-Guided Commonsense Relation Graph (SGCR-Graph). A collective information from both SC-Graph and SGCR-Graph helps to train the foundation model (in our case, GPT-2) and subsequently generates responses that elicit positive sentiment labels throughout conversations. We provide details of each module below.

7.4.1 Speaker Context Graph (SC-Graph)

Dialogue structure is maintained by the utterance and the speaker’s information. To effectively capture and maintain the dialogue structure, we utilize the Speaker Context Graph (SC-Graph). Such an approach models the dialogue by treating each utterance and its corresponding speaker information as nodes within the graph, thus encapsulating the relevant context for each part of the conversation.

In the SC-Graph, each utterance node is connected via an edge to its respective speaker label, $SC\text{-}Graph(S_p \rightarrow u_t)$. This connection ensures that the speaker’s identity is directly associated with their respective utterance, preserving the speaker-specific context. Furthermore, the sequential nature of the dialogue is maintained by connecting consecutive utterances through edges, represented as $SC\text{-}Graph(u_t \rightarrow u_{t+1})$. This structure captures the chronological progression of the conversation, ensuring that the temporal context between utterances is preserved.

To learn the contextual graph representations derived from the SC-Graph, we employ a graph attention layer (GAT-SC). This layer learns attention weights allowing the model to assess the importance of different nodes and edges selectively and generate richer contextual representations.

7.4.2 Commonsense Knowledge

We employ a commonsense transformer, COMET, to extract sentiment-specific knowledge pertaining to each utterance. This extraction process involves the utilization of two distinct modules within its pipeline – (i) Knowledge Extractor, and further, this acquired knowledge is used to construct (ii) Sentiment-Guided Commonsense Relation Graph (SGCR-Graph).

Knowledge Extractor. For each utterance-sentiment tuple $\langle u_i, s_i \rangle$, we define a conditional commonsense knowledge extraction rule. We feed a pretrained commonsense transformer with u_i as input by retrieving relevant knowledge aligned with sentiment-guided relations. For utterances expressing positive sentiments, we select COMET relations such as $xReact$, $xWant$, and $xIntent$. This selection is supported not only by empirical observations but also by psychological theories, indicating that individuals tend to experience

a psychologically happier state of mind when expressing positive sentiments and primarily focusing on themselves (280; 281). Conversely, for utterances reflecting negative sentiments, we employ COMET relations such as *oReact* and *oWant*. In these cases, clients typically perceive themselves in a negative state of mind, prompting others (in our case, the therapist) to offer consolation and encourage a shift toward positivity. Considering the robustness of the model, we fetch top-5 knowledge inferences for each utterance. Once knowledge from the commonsense transformer is retrieved, we construct an SGCR-Graph associating the utterance-specific knowledge and speaker labels. We discussed the commonsense relation selection in detail in Section 7.3.2.

Sentiment Guided Commonsense Relation Graph (SGCR-Graph). The process begins with encoding the knowledge extracted by the knowledge extractor using *Knowledge-Encoder*. This module is responsible for generating meaningful representations of the input knowledge. The input to the Knowledge-Encoder consists of a concatenation of a sentiment label prompt and the top-5 outputs of extracted knowledge, separated by a special separator tag used by the BERT transformer. This concatenation ensures that both sentiment and relevant knowledge are considered simultaneously during encoding.

Graph Structure: The resultant representations (r_t) for each utterance (u_i) are then used to construct a new graph. In this graph, edges are established between the speaker labels (Counselor (T) or Client (C)) and the utterance knowledge representations. This structure allows the graph to encapsulate both the speaker’s identity and the associated commonsense knowledge for each utterance. To maintain the context and capture the dialogue progression, directed edges are also added between the knowledge representation nodes of consecutive utterances to ensure that the temporal flow of the dialogue is preserved within the graph structure.

To further capture the nuances of sentiment knowledge and commonsense, we apply an additional graph attention layer (GAT-SGCR). This layer selectively focuses on important nodes and edges, refining the contextual representations.

7.4.3 Response Generation Block

For generating responses, we utilize models like GPT-2, which are pretrained on extensive text corpora and capable of producing semantically rich text. However, these models typically lack mechanisms for controlled and customized text generation. To address this, we employ GPT-2, a decoder-only transformer, and enhance its capabilities by incorporating external knowledge using a novel approach called *Knowledge Aware Attention*. The parametric size of GPT-2 (and others) is particularly suitable for our approach, as it allows for effective control of external knowledge without introducing significant noise. Larger models, while more powerful, tend to dilute the impact of external knowledge.

Knowledge Aware Attention. To integrate external knowledge into the GPT-2 model, we modify its multi-head attention block (MHA-Block). This modifies the conventional attention mechanism into a new cross-attention block designed to incorporate enriched knowledge representations. In our approach, the *query* in the attention mechanism is derived from the hidden representation of the previous decoder layer, following the standard dot product attention method. However, the *key* and *value* components are sourced from enriched representations obtained through two graph attention layers: *GAT – SC* and *GAT – SGCR*. These layers provide contextual and sentiment-guided commonsense information, respectively.

The representations from GAT-SC and GAT-SGCR are concatenated and then passed into the *Knowledge Aware Attention* block. This fusion of external knowledge with the internal mechanisms of GPT-2 improves the model’s ability to generate responses that are not only semantically coherent but also

		R1			R2			RL			BS	METEOR
		P	R	F1	P	R	F1	P	R	F1		
Baselines	HRED (244)	7.86	6.29	5.73	0.65	0.61	0.39	7.01	8.00	7.87	0.2421	0.298
	VHCR (246)	7.28	5.66	5.28	0.44	0.04	0.41	8.2	8.96	8.32	0.5030	0.3266
	HRED w/ Sp-Utt. Enc. (245)	8.26	7.55	7.08	0.69	0.83	0.653	8.38	7.53	7.07	0.4169	0.0465
	GPT2 (49)	11.13	9.23	9.23	1.02	1.52	1.13	11.49	12.12	9.26	0.4727	0.3494
	ConKADI (282)	8.43	6.34	6.38	1.24	0.98	0.98	9.48	9.13	7.83	0.2745	0.2588
	DialogGPT (241)	11.00	10.27	10.27	1.05	1.74	1.30	12.04	11.73	10.10	0.5819	0.3623
	CoMAE (283)	0.15	0.27	0.16	0.01	0.03	0.01	0.10	0.20	0.11	0.2930	0.4030
	ProphetNet (243)	8.60	8.31	6.63	1.05	0.97	1.01	8.51	8.83	7.22	0.5422	0.3998
	DialogVED (242)	9.88	10.70	8.75	1.11	1.29	1.46	10.02	11.5	9.05	0.5898	0.3711
	EmpHi (272)	6.45	4.48	4.49	1.14	1.02	0.93	11.52	11.23	8.57	0.4658	0.2758
	CEM (284)	8.25	5.58	6.78	0.75	0.91	0.81	9.54	8.59	7.07	0.3910	0.1232
	READER (261)	10.59	10.85	10.85	0.58	1.23	0.71	10.85	9.32	9.46	0.7600	0.2103
		13.23	13.33	12.93	1.82	1.99	1.92	13.23	13.5	11.80	0.8532	0.5984
Ablations	- SC-Graph - SGCR-Graph	11.13	9.23	9.23	1.02	1.52	1.13	11.49	12.12	9.26	0.4727	0.3494
	- SC-Graph	12.94	12.84	11.74	1.68	1.73	1.82	12.92	13.02	11.47	0.7654	0.5000
	- SGCR-Graph	12.12	12.83	12.83	1.80	1.52	1.54	12.54	12.83	10.73	0.6363	0.4240
$\Delta_{-BEST}(\%)$		$\uparrow 18.86$	$\uparrow 22.85$	$\uparrow 19.17$	$\uparrow 46.77$	$\uparrow 14.36$	$\uparrow 47.69$	$\uparrow 9.88$	$\uparrow 11.38$	$\uparrow 16.83$	$\uparrow 12.26$	$\uparrow 51.28$

Table 7.2: The upper half shows the results obtained on the HOPE dataset. We show ROUGE (1, 2, L), BERTScore (BS), and METEOR to assess the performance of the proposed model. The lower half of the table shows the ablation study in which we present individual contributions of two essential graph components in . Here, *SC-Graph* represents speaker-context graph, and *SGCR* denotes sentiment-guided commonsense relation graph. $\Delta_{-BEST}(\%)$ shows the percentage increment in the final model’s metric when compared with the baselines.

contextually and sentimentally informed. By integrating these enriched representations, we improve the model’s capacity to control text generation, leading to more relevant responses.

7.5 Experiments and Results

We perform numerous experiments to evaluate the performance of *EmpRes*, on the HOPE dataset by comparing it against twelve potential baselines. In this section, we provide comprehensive information about the HOPE dataset, the selection of baselines, and the metrics chosen for evaluation.

7.5.1 Comparable Baselines

We compare *EmpRes* with 12 competing baselines, focusing on systems that leverage dialogue context, commonsense, and empathy for response generation. Notably, the selected baselines cover all possible aspects of comparison. However, to the best of our knowledge, no other system has addressed the sentimental influence of any one interlocutor. In this study, we consider the following baselines as part of our analysis.

- **HRED** is based on a hierarchical encoder-decoder and trained for the dialogue modeling (244).
- **VHCR** is based on variational hierarchical RNNs for dialogue modeling (246).
- **HRED with Speaker and Utterance Encoder** enhances the HRED framework by incorporating speaker and utterance level information (245).
- **GPT-2** is a decoder-only transformer model. Our model, *EmpRes*, is also developed on GPT-2 (49).

Context	Therapist's Response
Therapist: How much are you drinking?	Gold: Yeah, are your friend's the problem? Maybe you just start hanging out with more kids in the youth group and not with your party kids? : You haven't been able to have an easier time getting back into good habits. <i>We can work together to identify strategies that will support your journey toward positive habits. [Positive Influence]</i>
Client: I don't think I am drinking that much. I mean, it's just social gathering. Like its nothing that I do like by myself or whatever.	GPT2: You are going through something?
Therapist: I am just so concerned. Can't you think of anything better to do?	ChatGPT: Exploring healthier coping mechanisms and finding fulfilling activities can greatly improve your overall well-being.
Client: I guess I can probably be more productive in my school work rather than going out to parties as much. But I feel helpless. <i>[Negative Sentiment]</i>	Gold: Tell me more about your new life.
Therapist: You were in asylum for two weeks. How are you feeling now?	: <i>That is an incredible improvement. Do you want to discuss about what other improvements you see?[Positive Influence]</i>
Client: Yes, initially it was good then I got bored of daily stuff. I think, you know, it helped my craving for smokes though.	GPT2: It is hard to live with such habits and no self control.
Therapist: Can you tell me if you have smoked after getting discharged?	ChatGPT: Acknowledging cravings without acting on them is a crucial step. Just track your progress, and find healthier coping mechanisms.
Client: I think sometimes cravings can go high but my self control over my thoughts has improved so thanks to you. <i>[Positive Sentiment]</i>	
Error Analysis	
Therapist: I think that's right. I see you crawled up over there. Also, you have a worried face and you're having hard time approaching it. You're feeling fear, but do not worry. When was the last time you had cheesecake?	Gold: I know, and so you've been avoiding food like this for so long that it totally makes sense that you're feeling fear. : You are so much more than just being scared about this.
Client: I honestly can't remember. I don't know.	GPT2: When was the last time you had cheesecake?
Therapist: I know it's been a really long time. You certainly didn't need it all through when you were really stuck in your eating disorder.	ChatGPT: It sounds cheesecake holds some nostalgic memory for you, but it's also connected to a difficult time with your eating disorder.
Client: I used to, I used to like it. I remember that, but it was a long time ago. Yet controlling is hitting hard on me. <i>[Negative Sentiment]</i>	

Table 7.3: A comparative analysis of therapist's response by the proposed model, , in contrast with the best-performing baseline (GPT-2), gold responses, and a state-of-the-art model, ChatGPT. Unlike other models, the responses generated by possess a positive influence on the client's negative sentiments. This is evident via the model's ability to control the response generation by exploiting *sentimental control* and *commonsense awareness*.

- **ConKADI** is an early effort in commonsense aware response generation (282).
- **DialoGPT** is a pretrained language model trained specifically on dialogue corpus (241).
- **CoMAE** is a multi-factor hierarchical framework trained for empathetic response generation (283).
- **ProphetNet-Dialog** is trained to chat naturally and generate coherent responses (243).
- **DialogVED** introduces continuous latent variables into the pretraining framework to enhance the relevance and diversity (242).
- **EmpHi** focuses to generate empathetic dialogue responses (272).
- **CEM** leverages the user's cognitive understanding for dialogue generation (284).
- **READER** is a transformer-reinforcement-learning approach for dialogue-act guided response generation. It is also trained on counseling conversational dataset (261).

To assess the comparative performance of all baseline approaches and EmpRes, we evaluate them using three widely-used evaluation metrics, namely ROUGE, METEOR, and BERTScore. The computation of these metrics is facilitated by specific libraries – *py-rouge*⁵ for ROUGE scores, *nlk-meteor*⁶ for METEOR, and *Hugging Face - BERTScore*⁷ for BERTScore.

⁵<https://pypi.org/project/py-rouge/>

⁶https://www.nltk.org/api/nltk.translate.meteor_score.html

⁷<https://huggingface.co/spaces/evaluate-metric/bertscore>

7.5.2 Performance Comparison

Table 7.2 shows the performance of the competing models. The results demonstrate the superiority of EmpRes across all evaluation metrics. Notably, our model excels in capturing both semantic and syntactic structures, as evidenced by the significant improvements of +12.26% and +51.28% in BERTScore and METEOR metrics, respectively. Interestingly, none of the baselines were able to capture the n -gram overlap ($n \geq 2$) efficiently, whereas EmpRes receives a significant increase of 47.69% in F1 and 46.77% in ROUGE-2 precision. In addition, EmpRes consistently outperforms the baselines in ROUGE-1 with an improvement of 18.86%, +22.85%, and +19.17% across precision, recall, and F1, respectively.

7.5.3 Ablation Study

EmpRes comprises two crucial modules, namely the Speaker-Context Graph composed of (*SC-Graph*) and the Sentiment-Guided Commonsense Relation Graph and (*SGCR-Graph*), both of which are important in capturing the dialogue structure and fostering commonsense awareness.

To thoroughly assess their impact, we conducted an ablation study as outlined in the latter portion of Table 7.2. The results of this study underscore the importance of both components in enhancing EmpRes’s performance. Removing either the *SC-Graph* or the *SGCR-Graph* significantly influences the model’s ability to generate coherent and contextually relevant responses. Interestingly, when the *SC-Graph* is eliminated, leaving only the influence of the *SGCR-Graph*, the model still demonstrates a commendable capability to produce responses that retain semantic richness and effectiveness.

This observation highlights the substantial contribution of sentimental influence and commonsense awareness to the dialogue generation process. Even though there is a marginal decrease in performance compared to the complete EmpRes, the responses generated in the absence of the *SC-Graph* remain remarkably close in quality. Thus, these findings reinforce the pivotal role played by sentimental cues and commonsense understanding in facilitating more natural and engaging conversational interactions.

7.5.4 Qualitative Analysis

To provide comprehensive support for our findings, we perform an extensive qualitative analysis. This analysis probes into the model’s generation capability, followed by a detailed examination of the human evaluation scores.

Generation Analysis. We investigate the quality of generations obtained from EmpRes. We observe that the responses generated by EmpRes contain structured semantics as shown in Table 7.3. To provide an estimate of the generative capabilities of other LLMs, we support EmpRes’s generations with gold responses from the HOPE dataset along with generations from GPT-2 and ChatGPT⁸. Evidently, the EmpRes is able to compete with the semantic generation standard of ChatGPT and surpasses the ability to influence the client in the counseling conversation positively. Table 7.3 highlights the section of EmpRes’s responses in which we observe sentimental control and commonsense awareness. We further observe that the responses received from EmpRes are comparable to the gold responses and, in some cases, superior and influential.

Error Analysis. EmpRes performs well in most of the cases. However, we uncover a few instances where a disproportionate emphasis on one of the integral components is observed. For instance, we

⁸<https://chat.openai.com/>

scrutinize a counseling scenario in the *error analysis* section of Table 7.3, which carries discussion related to eating disorders. Although EmpRes’s response adheres to the gold standard in terms of positive influence and sentiment, it falls short in terms of generating responses embedded with considerable knowledge. Our thorough analysis reveals that approximately 10% of randomly sampled responses exhibit deficiencies in either sentimental influence or the incorporation of relevant commonsense knowledge. Not to ignore that such generations are also possible due to the weak gold standard as in many cases, our human evaluation further claims that there exists a significant number of generations surpassing the quality of the gold standard. These evaluations are discussed next.

Human Evaluation. To ensure the quality and effectiveness of EmpRes, we conduct a comprehensive human evaluation on linguistic grounds (247). We employ four dedicated metrics:

- **Relevance** shows the selection of relevant content in the generated response considering the reference utterance.
- **Consistency** evaluates the factual alignment between the generated response and the source utterance.
- **Fluency** demonstrates the linguistic quality of the generated responses.
- **Coherence** shows the structure and organization of the generated responses.

We conducted an assessment of 70 randomly chosen instances, with 15 linguistic evaluators aged between 25 and 45. Each evaluator was tasked with assigning a score between 1 and 5 to four distinct parameters, where 5 represents the highest quality. This evaluation was repeated for both ChatGPT and GPT-2 for comparison purposes. By calculating the average scores obtained, we present the results in Table 7.4. We observe

Model	Relevance	Consistency	Fluency	Coherence
ChatGPT	2.13	1.89	3.41	3.15
GPT2	2.08	2.46	2.27	3.06
EmpRes	3.18 ($\rho = 0.0048$)	2.79 ($\rho = 0.0032$)	3.36 ($\rho = 0.0044$)	3.21 ($\rho = 0.0052$)

Table 7.4: Human evaluation on the responses generated by EmpRes when compared to the best-performing baseline and ChatGPT, along with their statistical significance test (p-value, ρ). The performance of EmpRes across all metrics is up to the mark and slightly better than comparing systems.

that EmpRes consistently demonstrates its superior qualitative performance except *fluency*, where ChatGPT has surpassed due to its exceptional ability to formulate human-level generation. Evaluators further stated that the model, in some cases, is able to generate better than the gold standard.

Relevance to Mental Health Experts. To further validate the work’s relevance, we collaborated with a licensed clinical psychologist and professor from a prominent organization. The effectiveness of EmpRes is validated through an exhaustive expert evaluation using three parameters on the scale of 1 (absent) to 5 (present).

- **Positive Impact** shows responses’ positive contribution to the user’s sentimental well-being.
- **Commonsense Understanding** shows the incorporation of contextually relevant responses that reflect a commonsense from context.
- **Therapeutic Appropriateness** ensures advice and guidance align with a therapeutic approach.

We observe an average rating of 2.70, 2.80, and 2.60 on the EmpRes’s responses for positive impact, commonsense understanding, and therapeutic appropriateness parameters, respectively. In addition,

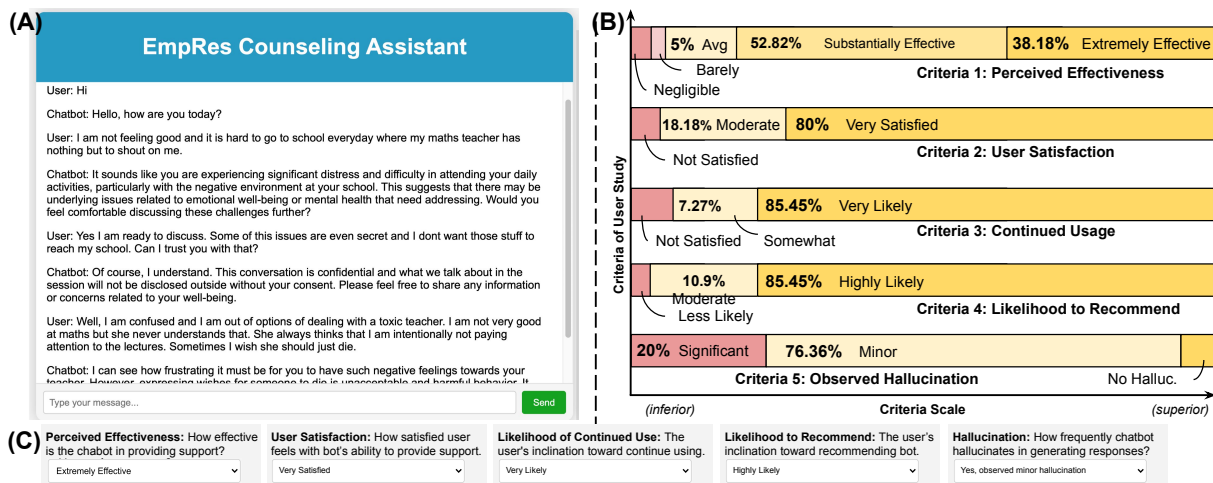


Figure 7.4: The figure illustrates (A) the prototype deployment of the chat interface and (B) findings of the user study conducted to evaluate the real-world effectiveness of the deployed EmpRes. (C) The user study comprises user feedback of conversation with the deployed system across five major criteria – (i) perceived effectiveness, (ii) user satisfaction, (iii) continued usage, (iv) likelihood to recommend, and (v) observed hallucination.

experts have confirmed that our model’s average rating surpasses that of competitive responses, further affirming the model’s capabilities. On the other hand, experts also claim that the proposed system sometimes fails to reflect certain crucial elements, showing a scope for improvement for future studies.

7.5.5 Experimental Setup

We perform numerous experiments using various combinations of the relevant commonsense relations, modeling loss. Further, we extensively experimented with graph construction. The process of graph construction for each batch is executed beforehand to keep the learning efficient. With this process, we significantly dropped the training time, making the model 10x more efficient. Moreover, we conduct extensive hyperparameter tuning, using a held-out validation set to explore a range of settings for key parameters (e.g., learning rate, batch size, and training epochs) to accurately learn the joint learning framework and observe a significant boost in the BERTScore and METEOR metrics. We perform all experiments on an Nvidia A6000 GPU. We tune our hyperparameters to find the optimal configurations. Our final version utilizes the learning rate of 2×10^{-6} , batch size of 8, which we run for 20 epochs. We observe that the model provides a significant decrease in the loss for the first ten epochs and then stabilizes itself in the next ten epochs.

7.6 User Study

We have implemented EmpRes in a controlled yet natural setting to assess its efficacy, allowing users to interact with the assistant. The primary objective is to introduce this system to a group of users, providing them access to the interface and encouraging engagement with the assistant. This deployment serves as the foundation for a comprehensive user study aimed at evaluating the real-world effectiveness of EmpRes. After gathering initial feedback from this study (stage 0), we advanced to stage 1, conducting a pilot study under expert supervision to explore usability aspects more thoroughly. The insights obtained from this

pilot study will inform refinements to EmpRes, enhancing its functionality in preparation for wider public usage in stage 2. Next, we present the findings from the initial user study conducted during stage 0.

We conducted a comprehensive user study to evaluate the real-world effectiveness and user satisfaction of EmpRes as a well-being assistant. A diverse group of participants ($N = 55$) were invited to engage with the assistant within a controlled setting and provide feedback on various interaction aspects. Invited participants represented a wide demographic range, including individuals from varying age groups (20-40), backgrounds, and geographic locations, to ensure a fair study and feedback mechanism. After each interaction, participants were prompted to provide feedback through a structured survey. The study focused on several key research questions to assess EmpRes’s performance pertaining to the perceived effectiveness, satisfaction, likelihood of continued usage, recommendation to others, and user’s perceptions regarding hallucinations. Based on these aspects, we discuss the findings of the study. Figure 7.4 summarises the key findings from the user study.

Firstly, regarding effectiveness and satisfaction, 91% of users rated EmpRes as *significantly or substantially effective* in providing mental health support. This high effectiveness rating indicates that EmpRes has the potential to positively impact users’ well-being. Additionally, 80% of participants expressed satisfaction with their interactions, demonstrating a generally positive user experience.

Evidently, the results indicate a significant positive response, with 91% of users rating EmpRes to be *significantly or substantially effective* in providing mental health support. Additionally, 80% expressed satisfaction, and over 85.45% conveyed their likelihood of continued usage and recommendation to others. On the other hand, a minute portion of 20% of users observed potential hallucinations in the model’s responses, highlighting the importance of controlling the model’s abilities to minimize such occurrences in future research.

7.6.1 Ethical Considerations for User Study

While deploying EmpRes as an assistant, we adhere to stringent ethical standards to safeguard user privacy, safety, and overall well-being. Prior to engaging with the assistant, we present participants in the user study with comprehensive guidelines and declarations. These declarations ensure that no personally identifiable information is collected within the deployed system, granting users the freedom to express themselves anonymously. In addition, participation in the study is entirely voluntary. The system adheres to ethical standards, emphasizing transparency and cultivating a positive user experience.

7.6.2 Societal Impact

Our proposed method serves as a valuable contribution to the mental health community and ongoing scientific investigations by harnessing the advancements in current-age dialogue systems specifically meant for counseling purposes. Such novel contributions within the realm of mental health hold significant potential for delivering substantial societal benefits. By directly aiding mental health professionals and industries in this space and offering accessible assistance to individuals in need, our work aims to foster a positive social impact in this domain.

7.7 Discussion

The design of commonsense relation selection is grounded in empirical modeling choices and evidence-based psychotherapeutic principles. While our earlier framing emphasized self-focus and positive thinking as drivers of adaptive conversational flow, it is important to note that our approach draws heavily from well-established therapeutic frameworks. Mindfulness-based interventions, as demonstrated in prior clinical studies (280), present acceptance. Within therapy, this typically follows a two-step process: first, validation of the client’s expressed emotional state; second, gentle guidance toward adaptive perspectives. Our method follows this same sequence. For utterances with negative sentiments, we retrieve and generate commonsense relations (e.g., oReact, oWant, causal scaffolds) that acknowledge and validate the client’s perspective before suggesting adaptive response generations. For positive sentiments, we emphasize xReact, xWant, and xIntent relations to reinforce beneficial mental states and sustain therapeutic momentum.

7.8 Ethics and Limitations

The research conducted in this sensitive area necessitates a comprehensive consideration of ethical implications and limitations. Given the gravity of the research, we prioritize the preservation of client privacy by implementing rigorous safeguards at every stage of our study. The study is backed with promising quantitative and qualitative results along with human feedback, who were adequately compensated. While the significance of this work to society cannot be overstated, it is important to acknowledge the scope of improvement in certain aspects, such as the model’s ability to handle extreme cases such as suicidal thoughts and accommodate linguistic diversity such as languages other than English. Furthermore, it is crucial to recognize that our approach represents only one facet of the broader efforts in this space. To illustrate this further, consider the typical trajectory of counseling. In the initial stages, clients often experience emotional distress, making it challenging for them to assess their situation or engage in formal treatment. Our primary goal is to develop a system that can effectively stabilize a client’s mental state, thereby making them more receptive to subsequent treatment. Therefore, its primary purpose is not posed to be an end-to-end treatment framework; instead, it serves as a crucial preliminary module within an end-to-end VMHA framework.

7.9 Summary

The perpetual challenge of addressing the scarcity of mental health professionals has become increasingly pronounced with each passing year. While the emergence of Virtual Mental Health Assistants (VMHAs) has provided clients with convenient access to support, the generation of commonsense-grounded effective responses remains an underexplored area. In this paper, we presented EmpRes, a novel model that leverages two fundamental graphs: the speaker-context graph and the sentiment-guided commonsense relation graph—to preserve dialogue structure and facilitate the extraction of relevant knowledge. Additionally, we employed GPT-2 with a modified knowledge-aware attention mechanism to incorporate augmented commonsense knowledge and sentimental influence into the response generation process. Our comprehensive evaluation against 12 competitive baselines demonstrated the superiority of EmpRes. Furthermore, we conducted an extensive ablation study, qualitative analysis, and human evaluation to highlight the effectiveness of EmpRes and identify its limitations. To assess the real-world applicability of our system, we deployed EmpRes as a chat interface and conducted a user study to evaluate its efficacy. The findings from the user study were promising, showing that 91.0% users found the assistant’s responses

to be *effective*. On the other hand, 80% users reported increased satisfaction with the interaction quality, noting that the responses felt more natural and supportive compared to other VMHAs they had previously used. The positive results from this deployment have laid the groundwork for further scalability and future enhancements. Our findings underscore the potential of EmpResto significantly improve the quality and responsiveness of VMHAs.

Part IV

Therapeutic Bond

8. Trust: A Therapeutic Bond Metric

In mental health counseling, a variety of earlier studies have focused on dialogue modeling. However, most of these studies give limited to no emphasis on the quality of interaction between a patient and a therapist. The therapeutic bond between a patient and a therapist directly correlates with effective mental health counseling. It involves developing the patient’s trust on the therapist over the course of counseling. To assess the therapeutic bond in counseling, we introduce *trust* as a therapist-assistive metric. Our definition of trust involves patients’ willingness and openness to express themselves and, consequently, receive better care. We conceptualize it as a dynamic trajectory observable through textual interactions during the counseling. To facilitate trust modeling, we present MENTAL-TRUST, a novel counseling dataset comprising manual annotation of 212 counseling sessions with first-of-its-kind seven expert-verified ordinal trust levels. We project our problem statement as an ordinal classification task for trust quantification and propose a new benchmark, TRUST-BENCH, comprising a suite of classical and state-of-the-art language models on MENTAL-TRUST. We evaluate the performance across a suite of metrics and lay out an exhaustive set of findings. Our study aims to unfold how trust evolves in therapeutic interactions.

8.1 Introduction

Mental health counseling relies on dynamic, evolving strategies tailored to the needs of each patient. For therapists, a successful session hinges on their ability to gauge the effectiveness of their approach, often relying on subtle cues to adjust their methods. This adaptability becomes especially crucial in text-based therapy, where the challenge of understanding patient engagement is even more complex. One of the many contributing factors responsible for effectiveness is the therapeutic bond. To assess this, we propose the need for a *therapist-assistive metric* to evaluate the success of ongoing strategies, one that could signal when recalibration is needed during counseling (285). In this work, we propose *trust* as a metric to assess the therapeutic bond. Trust, in a therapeutic context, reflects a patient’s willingness to disclose sensitive, personal matters; it unfolds as a dynamic trajectory rather than a static state. For instance, when a patient shifts from surface-level observations (e.g., “I’ve been feeling a little down”) to deeper self-exposure (e.g., “I’ve been scared of losing my job because I think it defines me”), it signifies growing trust. Conversely, hesitancy or disengagement might indicate faltering trust. The need for trust as an evaluative metric is particularly urgent in text-based therapy, which has seen exponential growth in recent years (286). However, they also place the burden of building and sustaining trust solely on the therapist’s linguistic and the patient’s reactions (287). Current research in mental health focuses primarily on tracking sentiment, emotion, or engagement but lacks mechanisms to assess the interaction quality (288; 289; 290).

Existing research has explored related areas like sentiments, user reactions, and moods in conversations (291; 292). Still, it often falls short of addressing trust as a measurable construct (293). Unlike existing attributes, trust possesses an evaluative component: it not only reflects the patient’s feelings but also signals the effectiveness of the ongoing therapy approach. Without a system to identify trust flow, therapists and the evaluative systems designed to assist them are left without actionable insights into the relational dynamics of their sessions.

In our work, we address this gap by proposing *trust*, a novel and expert-verified metric designed to assess therapeutic bonds. We conceptualize trust as a patient’s willingness to disclose sensitive, relevant matters, reflecting their engagement and confidence in the therapeutic process. To practically assess this, we propose `MENTAL-TRUST`, a large, manually annotated dataset capturing trust dynamics and topic-switch patterns across patient-therapist dialogues. Furthermore, we evaluate the potential of computational models for trust detection through `TRUST-BENCH`, a comprehensive benchmarking of classical and state-of-the-art language models. Finally, we conduct an in-depth analysis of trust trajectories, uncovering key insights into how trust evolves during counseling conversations. Our findings reveal that while large language models (LLMs) excel at capturing topical understanding, they often struggle with accurately detecting fluctuations in trust levels. Interestingly, smaller models such as BART, BERT and DeBERTa outperforms LLMs in this regard, showcasing a stronger ability to capture subtle variations in trust. Additionally, our in-depth analysis of trust trajectory paths, comparing increasing, decreasing, and neutral fluctuations, demonstrates that these smaller models effectively follow trust trajectories across all three cases. Our contributions are summarized below.

- We propose trust, a novel metric for evaluating therapeutic bonds.
- `MENTAL-TRUST`: We propose a novel and expert-backed annotated counseling dataset, capturing trust levels and topic-switch patterns in patient-therapist dialogues.
- `TRUST-BENCH`: We benchmark the performance of state-of-the-art language models, assessing their ability to detect trust.
- We analyze and criticize the findings from `TRUST-BENCH` to further understand trust trajectories, highlighting their potential to assist therapists in real-time strategy recalibration.

8.2 Related Work

Mental health has emerged as a profound global concern, necessitating innovative solutions beyond traditional approaches. The rise of digital interventions, particularly AI-powered tools, such as chatbots and diagnostic algorithms, have transformed the landscape of mental healthcare. Amidst growing demand and stretched resources, current research has focused on multiple verticals. Earlier works within dialogue space for counseling understanding include counseling understanding (294; 295), ethics (296), counseling note generation (170; 297), behavior assessment, generation (261). However, barely any work touches upon the quality of interaction between therapist and patient.

Study of bonds and psychology in digital interventions. The study of human bonds, especially in the context of psychological interactions (298; 299; 300), has gained increased attention in digital interventions. Research has shown that building a therapeutic bond between patients and healthcare providers is a critical factor in achieving alignment with treatment goals (301; 302; 303). In traditional settings, this bond is forged through direct human interaction. However, as current methods become more prevalent in counseling, understanding how bonds form between patients and these systems has become essential. Studies highlight the challenges of replicating the empathetic connection that a human therapist provides, pointing out that AI tools often struggle to instill a sense of trust and rapport, both crucial elements of therapeutic progress (304).

Trust. Plutchik’s wheel of emotions is one the early successful attempts to classify human emotions into eight basic categories, including trust (305). Here, trust is understood as the emotion that signals

#	Utterances	Trust	Topic Discussing
1	Therapist: In the other chair, let's have critic that's inside, that internal critic. That's advocating for settling down, for getting married, for having kids. In a way, it represents societal pressure . These two sides separate to you? Patient: Yeah.	1.5	Balancing career and settling down
2	Therapist: You feel like you can be your emotional self here and the internal critic there? Patient: I think so.	1.5	Balancing career and settling down
3	Therapist: Ok so, I think you remember how this works. You'll be talking to the internal critic, right? So, the internal critic is on the side of settling down and getting married. And then, when you're ready to have the critic respond, you'll switch to that chair and respond to the emotional, needs-driven side. Does that make sense? So whenever you're ready, go ahead and speak directly to the internal critic. Patient: I want to finish what I'm doing, and I want to finish what I started. I've worked so hard on my career path, and I'm so close to finishing. I want to finish for me. I want to accomplish those goals that I've set for myself. It's something I'm passionate about.	2	Balancing career and settling down
4	Therapist: So you've made your statement. If you are ready to switch then I use the internal critic. Patient: There's always time to pursue a career, but I need to focus on what's expected of me at this time, on the best age for having kids I have a loving boyfriend who is ready and wants to get married and settle down. And you are being selfish by just following a path that you can take at any time.	2	Balancing career and settling down
5	Therapist: Okay. So the internal critic is just called a selfish and one of the things is a response. Patient: I feel like I'm being attacked. And it's not wrong to think about myself and what I am passionate about.	2.5	Balancing career and settling down
6	Therapist: Let me interrupt you. So your hands are pointing back toward the career emotional self. What can you do with your hands that really embody that you want to communicate to the internal critic? Anything else? Patient: The pointing towards me to protect myself from what critic is saying. I wanna push that back to the critic.	3	Decision Making

Table 8.1: Dataset example. The table showcases a dialogue snippet from the dataset, accompanied by trust level annotations and the discussion topic. We highlight the key portions of the text contributing significantly to the constructive and incremental trust trajectory for better interpretability.

that ‘something is safe.’ Major theories in human psychosocial development emphasize the relevance of trust in building well-functioning relationships between individuals (306; 307), although there has been little research specifically on the patient’s level of trust toward psychotherapists (308). (36) attempted to define therapeutic trust (default therapeutic trust and overriding therapeutic trust), and differentiated it from ordinary trust. (35) aimed to develop and report preliminary psychometric analyses of a new brief measure to evaluate a patient’s level of trust and respect for their clinician. However, none of these model the trust dynamics of the patient during the progress of a dialogue.

In contrast to previous definitions that conceptualize trust as a static measure, our work focuses on trust as a dynamic trajectory within dialogue. Unlike psychometric measures that evaluate trust as an outcome, we define trust as a moment-by-moment phenomenon, capturing its fluctuations in response to the patient-therapist interaction.

8.3 Dataset

We propose MENTAL-TRUST, a trust-level rated counseling dataset specifically curated for trust modeling in counseling conversations. The dataset consists of 12.9K utterances extracted from 212 dialogues, where each patient utterance is annotated with seven expert-verified ordinal trust labels. These labels are designed to capture the nuanced progression of trust within therapeutic interactions. To create MENTAL-TRUST, we extend the publicly available HOPE counseling dataset, which includes a total of 12,912 utterances across multiple counseling sessions (294). Our subset focuses exclusively on patient utterances, providing a targeted framework for studying trust as a dynamic construct. The following subsections detail the data collection process, annotation scheme, trust labels, and key dataset statistics.

8.3.1 Data Annotation

Counseling conversations differ from standard dialogues, particularly in their nuanced focus on emotional depth, trust dynamics, and therapeutic context (309; 310; 311). These unique characteristics necessitate a specialized annotation approach to capture the intricate patterns. To address this, we collaborate with mental health experts to design a specialized annotation scheme. This scheme comprises seven ordinal trust levels, explicitly crafted to capture the evolving dynamics. Before detailing the annotation guidelines, it is essential to first establish an understanding of trust, which we discuss next.

Trust. Trust reflects the patient’s willingness to depend, or intent to depend, on the expert/therapist with a feeling of relative security in spite of a lack of control over the expert, even though negative consequences are possible. Specifically, the trust of a patient towards their therapist is characterized by (a) sharing personal, detailed, or sensitive information, (b) opening up about relevant concerns, and (c) maintaining focus on the topic of concern without unnecessary deviation. It is worth noting that our primary goal is to assess the patient’s trust in the therapist; however, the impact of the therapist’s responses and interventions would affect the client’s degree of opening up. Therefore, we annotate trust levels for every patient utterance from seven levels defined by experts. A trust level is an ordinal value assigned to each flip, determined by considering its conversational context and preceding trust levels. Finally, prior to annotation, the relevant topic of discussion is identified to ensure accurate context alignment. To make it more streamlined and precise to ongoing annotations, we mark the change of topic, both instances and the shifted topic. Annotators evaluate and update trust levels based on shifts in the conversation’s focus for better alignment with the identified topic. Detailed annotation guidelines are discussed below.

8.3.2 Annotation Guidelines

To effectively capture the dynamics of trust in counseling conversations, we develop a framework of seven ordinal trust levels. These levels represent a spectrum of a patient’s openness, ranging from complete refusal to engage with the therapist to fully opening up about their core issues. While there are seven levels in total, four major levels: least trust (L1), low trust (L2), building trust (L3), and achieved trust (L4), serve as the foundation. The intermediate levels (e.g., 1.5, 2.5, and 3.5) account for complex situations where a patient’s trust behavior does not fit neatly into one of the major categories. Below, we define each trust level:

1. **Least Trust.** This level is assigned when the client demonstrates an active refusal to open up. This could include non-aligning responses or outright rejection of the therapist’s attempts to engage.
2. **Low Trust.** This level represents a slight increase in trust from Level 1 or a slight decrease from Level 3. The patient shows hesitation when opening up but does not completely refuse. Indicators of low trust include:
 - Limited self-expression.
 - Use of discussion filler such as “um” or “hm”.
 - Expressions of insecurity or doubtfulness, such as “maybe” or “I guess”.
3. **Building Trust.** This level is scored when the patient demonstrates consistent engagement and openness, though they may still require prompting from the therapist or digress from the main topic of concern.

Models	#Sessions	#Utterances	Avg. Utterance Length	Avg. Utt. Len./Spk.		Avg. Tok/Spk./Utt.	
				T	P	T	P
Train	116	6902	56.56	28.51	28.05	26.15	29.75
Test	17	949	52.71	27.00	25.71	25.07	21.06
Val	34	2321	65.21	32.71	32.50	26.73	19.32

Table 8.2: Split-wise dialogue- and speaker-based analysis of the proposed dataset, MENTAL-TRUST.

4. **Achieved Trust.** This level is assigned when the patient fully opens up to discuss their core issues without any digression. At this point, they actively engage with the therapist on the point of concern, indicating complete trust.

Intermediate Trust Levels. In many conversations, patient utterances do not always fit neatly into one of the four primary trust levels. For instance, an utterance might exhibit elements of both Level 1 (least trust) and Level 2 (low trust), creating ambiguity for the annotators. To address this challenge, we experiment with intermediate levels, such as 1.5, 2.5, and 3.5, positioned between the major trust levels. These midpoints allow annotators to capture complex shifts in trust that might otherwise lead to inconsistencies. Not only annotators report that these additional levels significantly improved their ability to assign accurate trust scores, resulting in better alignment with expert judgment and inter-annotator agreement. One such importance is that the conversations are initialized with a trust level of 2.5, representing a neutral midpoint. This is a fair assumption to ensure consistency across annotations, as there is no prior context available to set a more specific initial trust level. From this point, annotators observe the patient’s trust trajectory and annotate based on the guidelines.

Annotating Topic Shifts. In addition to trust levels, we annotate topic shifts to capture changes in the focus of the conversation. A topic shift is marked whenever the conversation deviates from the current topic of discussion to introduce a new subject. This annotation helps track how trust dynamics interact with the flow of dialogue, as maintaining alignment. We ask annotators to record the new topic introduced and the point in the dialogue where the shift occurred, ensuring a detailed record of conversational transitions and *digressions*.

We present a dialogue snippet in Table 8.1 to showcase example annotations. Along with the dialogue, it includes the topic and trust level annotations, highlighting key factors contributing to incremental trust levels.

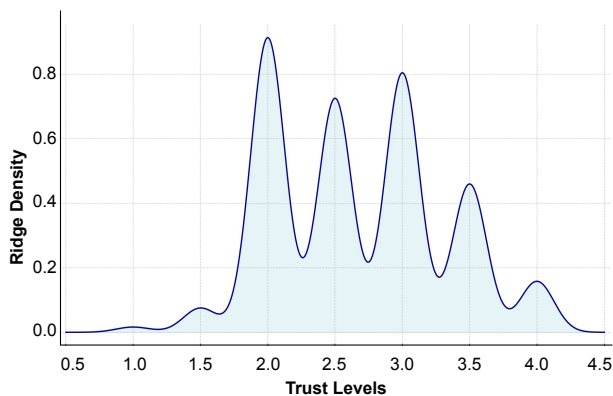


Figure 8.1: Distribution of ridge density for each trust level between 1 (min) – 4 (max) in MENTAL-TRUST.

8.3.3 Annotation Process

To ensure reliable annotations, we adopt an iterative approach involving three expert linguists and a team of domain experts. In Iteration 1, we begin with four primary trust levels only for a random sample of

10 dialogues and ask them to assign trust levels based on the initial guidelines. However, this iteration revealed significant disagreements, with an inter-annotator agreement (IAA) score of 0.22, as many utterances exhibited mixed characteristics that did not fit neatly into the defined levels. In Iteration 2, we introduced intermediate trust levels, and this refinement initially resolved much of the disagreement. However, a new challenge emerged: subjective interpretations of whether a patient’s utterance reflected engagement with the relevant topic of concern or constituted digression. Despite this, the IAA score improved to 0.43. To address this issue, Iteration 3 incorporated expert-provided criteria to clarify how the topic of discussion should influence trust levels. Annotators were instructed to consider whether utterances aligned with the core topic or diverged as digressions. With these refined guidelines, annotations became more consistent, and the final iteration achieved a Cohen’s Kappa score of 0.77, falling into the ‘substantial’ agreement category. This iterative process not only resolved key ambiguities but also ensured a robust and reliable annotation framework for the dataset.

8.4 Dataset Analysis

We conduct a detailed analysis examining both the structural composition and trust dynamics to understand the characteristics of MENTAL-TRUST.

Statistical Analysis. Table 8.2 presents the key statistics of our dataset split across training, testing, and validation sets. The dataset comprises a total of 167 counseling sessions, with 116 sessions (6,902 utterances) allocated to training, 17 sessions (949 utterances) to testing, and 34 sessions (2,321 utterances) to validation. The average utterance length shows consistent patterns across splits, with validation utterances being slightly longer (65.21 tokens) compared to training (56.56 tokens) and test (52.71 tokens) sets. This variation is natural in therapeutic conversations where utterance length often correlates with the depth of discussion.

We observe interesting patterns in speaker-specific assessment. The average utterance length per speaker shows therapists maintaining relatively consistent lengths across splits (28.51, 27.00, and 32.71 tokens), while patients demonstrate similar consistency (28.05, 25.71, and 32.50 tokens). This balance between therapist and patient utterance lengths suggests equitable dialogue participation. Likewise, the average tokens per speaker per utterance reveals that therapists typically use fewer tokens (26.15, 25.07, and 26.73) compared to patients in the training and test sets (29.75 and 21.06), except in the validation set where patient utterances are notably more concise (19.32). This pattern might reflect varying degrees of patient expressiveness across different sessions, while therapists maintain more consistent communication patterns.

Trust Distribution. Analysis of trust level distribution in our dataset reveals notable patterns in therapeutic trust dynamics. Figure 8.1 presents a ridge density plot of trust levels, demonstrating three distinct peaks. The primary peak occurs at trust level 2.0 with the highest density of 0.88, indicating that moderate trust levels are most prevalent. Two secondary peaks emerge at trust levels 2.5 and 3.0, suggesting that patients commonly transition through these intermediate trust states. On the other hand, the density tapers significantly at both extremes (below 1.0 and above 4.0), indicating that both very low and very high trust states are relatively rare. This aligns with clinical observations that establishing deep trust requires sustained therapeutic engagement, while complete distrust is uncommon once patients initiate therapy.

Models	Acc	$F1_m$	$F1_w$	P_m	R_m	P_w	R_w	OL_{CE}	Params
Encoder Only									
BERT (312)	85.04	68.93	84.90	69.02	69.16	85.02	85.04	3.67	109 M
Mental-BERT (313)	84.12	70.14	83.87	71.83	69.12	83.99	84.12	3.75	109 M
RoBERTa (314)	80.76	61.96	81.01	61.58	63.60	82.12	80.76	4.28	124 M
ALBERT (315)	84.48	72.77	84.34	83.88	70.21	85.19	84.48	3.73	11 M
DeBERTa (316)	85.30	72.01	85.23	82.82	72.91	85.81	85.30	3.46	184 M
XL-Net (317)	78.76	53.49	78.22	52.60	55.26	78.27	78.76	4.89	117 M
Encoder-Decoder									
BART (318)	87.39	81.15	87.34	84.82	79.35	87.52	87.39	2.92	407 M
Mental-BART (319)	89.03	81.63	88.96	89.48	78.22	89.15	89.03	2.53	407 M
T5 (320)	71.82	49.58	70.93	51.04	49.70	71.10	71.82	5.93	738 M
Decoder Only									
Llama 3.1 (321)	66.51	43.72	64.63	48.50	42.83	65.92	66.51	6.60	8000 M
Mistral-7b (213)	64.17	42.60	62.75	47.36	42.13	64.13	64.17	7.11	7000 M
Phi 3.5 (322)	74.43	50.24	73.10	52.98	49.52	73.27	74.43	5.27	3800 M
Closed-Source									
GPT 4o (323)	23.38	12.85	20.57	14.80	15.38	24.37	23.38	-	Large
Gemini 1.5 (324)	23.58	12.06	19.10	15.79	17.56	25.38	23.58	-	Large

Table 8.3: TrustBench – Benchmarking. We evaluate 14 models on the proposed dataset, , across four major categories: encoder-only, decoder-only, encoder-decoder, and closed-source. Evidently, smaller models outperform larger models across eight metrics, including Accuracy (Acc), F1, Precision (P), Recall (R), and Ordinal Cross-Entropy Loss (OL_{CE}).

8.5 TrustBench

Building on our definition of trust, we propose TrustBench, a benchmarking framework designed to evaluate how well existing methods can quantify trust in counseling conversations. We operationalize the problem statement as an ordinal classification task. This section outlines the methodologies, experiments, and evaluation metrics employed in TrustBench to benchmark state-of-the-art models.

8.5.1 Trust Modeling

The input to these models is meant to capture the dynamics of trust through conversational context rather than treating utterances in isolation. To capture the dynamics of trust as it evolves during a conversation, we consider contextual representations of patient utterances within their dialogue history. Given a dialogue D consisting of n utterances $\{u_1, u_2, \dots, u_n\}$, where each utterance u_i is either from the patient (P) or the therapist (T), the task is to predict the trust score $t_i \in \{1, 2, 3, 4, 5\}$ for each patient utterance u_i . The context window $\{u_{i-k}, \dots, u_{i-1}\}$, consisting of the k preceding utterances for context.

8.5.2 Competing Models

We discuss 14 competitive yet diverse baseline methods that we employ to benchmark.

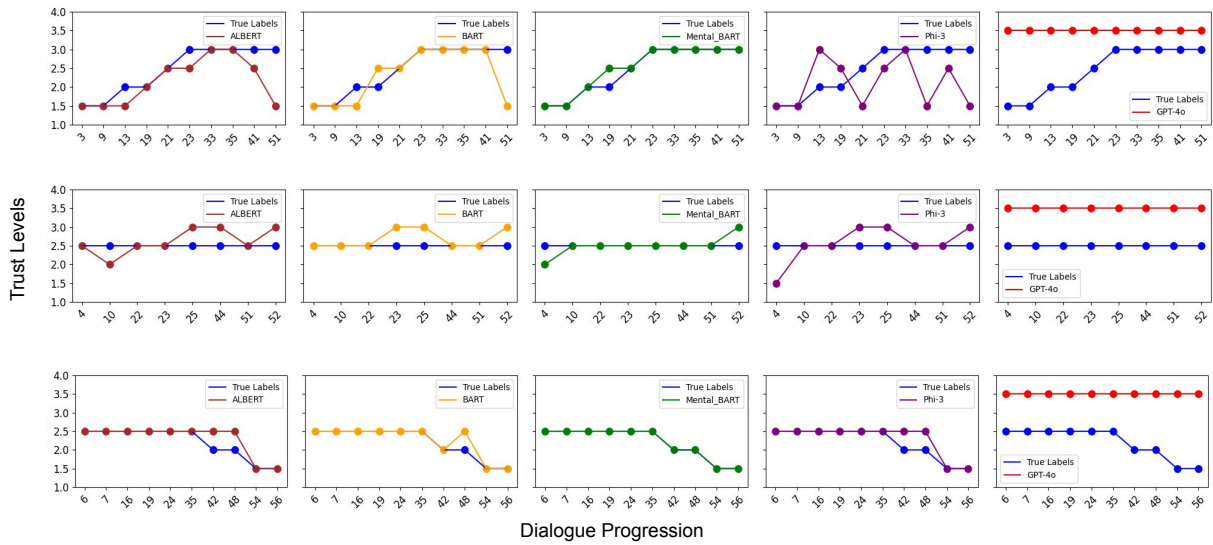


Figure 8.2: Examples of trust progression scenarios (*increasing, constant, decreasing*) across top-performing models. The ground-truth trust values are represented by the blue line. Decoder-only models, particularly domain-specific variants such as Mental-BART, achieve the closest alignment with true trust values, while closed-source models perform the worst due to their inherent rigidity.

Encoder-only Methods. ▶ **BERT** (312) is a bidirectional model that excels in understanding context through deep pretraining on large text corpora. ▶ **Mental-BERT** (313) is a variant fine-tuned on mental health data, making it well-suited for tasks with psycholinguistic nuances. ▶ **RoBERTa** (314) improves upon BERT by optimizing training strategies, making it highly effective for a wide range of NLP tasks. ▶ **ALBERT-base-v2** (315) is a lighter version of BERT that reduces model complexity while maintaining competitive performance. ▶ **DeBERTa-v3-base** (316) incorporates disentangled attention mechanisms, allowing it to better capture semantic relationships. ▶ **XLNet** (317) uses a permutation-based training approach, which enhances its capability to model bidirectional context.

Encoder-Decoder Methods. ▶ **BART** (318) is pre-trained as a denoising autoencoder. ▶ **Mental-BART** (319), fine-tuned on mental health conversations, makes it highly effective for generating therapeutic dialogue. ▶ **T5** (320) reframes NLP tasks into a text-to-text format, making it a versatile encoder-decoder model capable of diverse tasks.

Decoder-only Models. ▶ **LLaMA-3.1** (321) is designed for efficiency with smaller datasets, demonstrating strong performance. ▶ **Mistral-7B-v0.1** (213) is optimized for language modeling. ▶ **Phi-3.5** (322) is a small language model (SLM) particularly useful for easy compute power and inference while maintaining the abilities of LLMs.

Close-sourced Models. We also integrate closed-source, state-of-the-art models such as ▶ **GPT-4o** (323), known for its complex reasoning and strong performance in general NLP and instruction following tasks, and ▶ **Gemini** (324), which excels in multitasking scenarios, particularly in generative AI.

8.6 Results and Analysis

Here, we discuss the performances achieved by participating methods.

8.6.1 Performance Comparison

We benchmark the proposed dataset, MENTAL-TRUST, evaluating the performance of 14 competing baseline methods across nine evaluation metrics. The baseline methods are categorized into four major architectures: encoder-only, encoder-decoder, decoder-only, and closed-source language models. Table 8.3 highlights Mental-BART as the best performer.

Within the encoder-only category, DeBERTa achieves the best results, surpassing other methods with scores of 85.30 (accuracy), 85.23 ($F1_w$), 85.81 (P_w), and 3.46 (OL_{CE}). Notably, DeBERTa’s performance closely aligns with that of ALBERT, demonstrating competitive capabilities in this architecture. On the other hand, for the decoder-only methods, Phi-3.5, a small language model, emerges as the best performer, recording scores of 74.43 (accuracy), 73.10 ($F1_w$), 73.27 (P_w), and 5.27 (OL_{CE}). Meanwhile, in zero-shot comparisons with closed-source LLMs such as GPT-4o and Gemini-1.5, these models fall short of fine-tuned methods across all metrics. Surprisingly, despite their advanced capabilities, even these sophisticated models fail to outperform smaller, fine-tuned encoder-only models like DeBERTa.

Finally, we test full-fledged transformer architectures, including BART and its variant Mental-BART. These models consistently outperform all other baselines across all the segments, with Mental-BART achieving the best scores on seven out of nine metrics. For Mental-BART, we observe a performance of 89.03 (accuracy), 88.96 ($F1_w$), 89.15 (P_w), and 2.53 (OL_{CE}), further cementing its position as the top-performing model on the proposed dataset. Our findings underline the effectiveness of specialized and smaller models like Mental-BART in the trust modeling task and highlight the limitations of general-purpose LLMs in addressing the ordinal classification tasks. We discuss the effect of model sizes for trust modeling in Section 8.7.1.

It is worth noting that all decoder-only models in the TrustBench evaluation were assessed using identical few-shot, instruction-tuned prompts to ensure fairness in comparison. For smaller models, we additionally performed instruction-based fine-tuning, while larger proprietary models (e.g., GPT-4 and Gemini) were restricted to few-shot inference without fine-tuning due to proprietary constraints. The

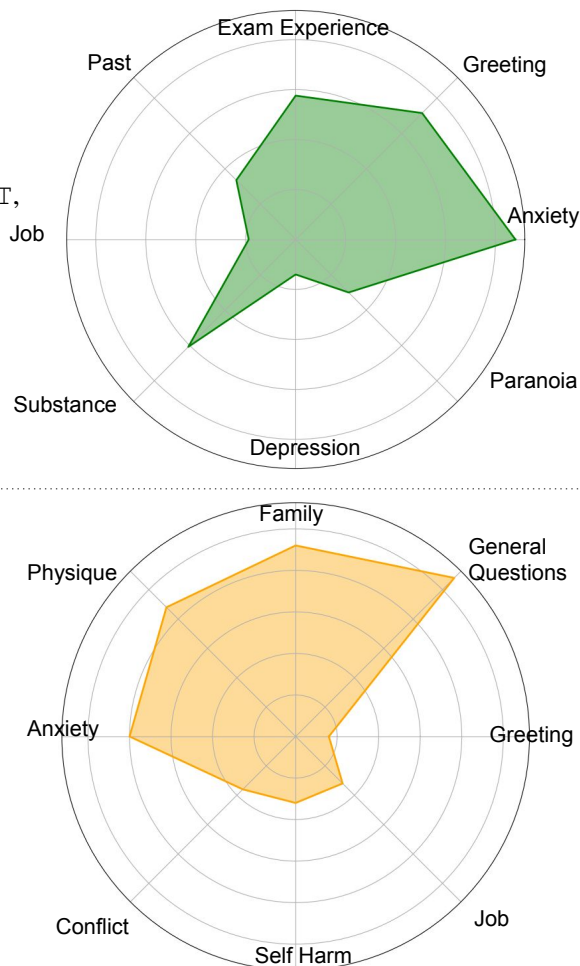


Figure 8.3: Topic of discussion in counseling sessions segmented into two major segments. Top: illustrating core topics in sessions with a positive topic alignment; Bottom: showing topics in sessions with negative topic alignment.

Splits	+ve Jumps	Max(+ Δ)	Min(+ Δ)	Avg(+ Δ)	-ve Jumps	Min(- Δ)	Max(- Δ)	Avg(- Δ)	Avg Streak
Train	228	1.0	0.5	0.504	110	-0.5	-2.0	-1.032	8.79
Test	37	0.5	0.5	0.5	16	-0.5	-2.0	-1.063	7.61
Val	77	1.0	0.5	0.506	36	-0.5	-2.0	-1.069	9.30

Table 8.4: Trust jumps and streak analysis. The table shows trust dynamics and includes the counts of trust level jump, their respective maximum, minimum, and average changes (+ Δ and - Δ), and the average streak length, highlighting the progression and fluctuations in trust trajectories.

observed performance gap between smaller and larger models is largely attributed to the differences in fine-tunability. Targeted fine-tuning allowed smaller models to align closely with domain-specific objectives. Closed-source inclusion is intended to assess whether current state-of-the-art LLMs could address the problem effectively in their default form. Our results indicate that, without domain-specific adaptation, these models fall short, necessitating the need for our proposed line of approach.

8.6.2 Analysis

We study the dynamics of trust modeling in counseling, examining two major fronts: (a) trust trajectory modeling and (b) topical analysis. We study these aspects for two therapeutic alignments: *positive topic alignment*, where concluding trust exceeds initial trust levels, and *negative topic alignment*, where concluding trust falls below initial trust levels. We lay out an evidential analysis and present our findings, highlighting both the strengths and limitations of models in capturing trust dynamics.

Trust Trajectory Analysis

The quantitative evaluation of trust modeling, as presented in Table 8.3, provides a comprehensive overview of each model’s performance when compared with expert-annotated trust levels as an ordinal classification task. While these scores reflect the overall ability of models to predict trust levels, we further study to understand how well models follow trust trajectories from the beginning of a session. To achieve this, we analyze model-predicted trajectories against gold-standard annotations for three distinct cases based on their overall alignments, including positive, negative, and neutral topic alignments (where trust levels remain relatively stable).

To balance thoroughness and efficiency, we focus on the best-performing model from each architecture paradigm, including LLMs, SLMs, encoder-only, decoder-only, and transformer-based. Evidently, as shown in Figure 8.2, Mental-BART demonstrates superior alignment with the gold standard, particularly excelling in scenarios involving abrupt fluctuations. Its performance closely rivals ALBERT, reflecting robustness in handling moderate trust variations. On the other hand, ALBERT, while competitive, shows limitations in scenarios involving sharp trust fluctuations, such as rapid dips or surges. Similarly, the SLM, Phi-3.5, struggles with rapid directional changes.

Topic vs Therapeutic Alignment

We perform a topical analysis of counseling sessions to investigate thematic trends across different therapeutic alignments. Figure 8.3 illustrates the frequency distribution of key topics discussed during these sessions, providing insights into the relationship between topic alignment and trust trajectories.

Sessions with positive topic alignment are characterized by ‘in-depth discussions’ on focused topics such as *anxiety*, *academic examinations*, and *substance* use or abuse. These conversations display ‘minimal digressions’, with patients maintaining alignment with the core topic of concern. Additionally, a recurring pattern emerges where sessions often begin or conclude with proper *greetings*, a seemingly routine element that contrasts sharply with negative topic alignment sessions, which frequently lack such exchanges. Furthermore, the topics discussed in positive alignment tend to be coherent and closely aligned with the therapeutic goal. Conversely, sessions with negative topic alignment exhibit scattered and tangential discussions, often touching on topics such as *general questions*, *casual chitchat*, *family dynamics*, *physical appearance*, and *anxiety*. A striking feature of these sessions is the prevalence of topic digressions, where patients shift focus away from the core concern to peripheral or unrelated issues, and an even bigger concern is ‘frequent digression’. These deviations are commonly linked to diminished trust, as the therapeutic bond is weakened when alignment on core emotional or therapeutic goals is not maintained.

8.7 Discussion

In this section, we discuss two key aspects of our work: (a) the patterns of trust evolution and (b) the counterintuitive relationship between model size and trust modeling.

8.7.1 Discussion on Effect of Model Sizes

Our findings reveal a counterintuitive relationship between model size and trust modeling performance. Despite the common assumption that LLMs lead to better performance, our analysis shows an inverse trend in the context of therapeutic trust modeling. Smaller encoder models like BERT (109M parameters) and ALBERT (11M parameters) achieve notably higher accuracy (85.04% and 84.48% respectively) compared to larger models such as Llama 3.1 (8B parameters) and Mistral (7B parameters), which achieve only 66.51% and 64.17% accuracy respectively. This inverse relationship becomes even more surprising with closed-source models like GPT-4 and Gemini 1.5, which, despite their large parameter counts, perform significantly worse with accuracies around 23%. **It is worth noting that all decoder-based models were evaluated using a few-shot instruction-tuned prompting strategy, with prompts kept consistent across all models for fairness. Smaller models underwent instruction-based fine-tuning, whereas large proprietary models (GPT-4, Gemini) were evaluated in few-shot inference mode without fine-tuning due to their closed-source nature.**

Our understanding of this pattern is that smaller models are more focused architectures suited to learning the specific patterns and subtle indicators of trust levels. For instance, smaller models are better at learning local, contextual patterns specific to trust dynamics rather than relying on broad world knowledge. Moreover, the specialized nature of trust modeling might benefit from models that can be more precisely fine-tuned to the task. **Also, the inclusion of closed-source models in our experiments serves primarily to assess whether state-of-the-art proprietary LLMs, in their standalone form, could already address the targeted problem. The results confirm they cannot fully solve it without domain-specific fine-tuning, underscoring the novelty and necessity of our proposed approach.**

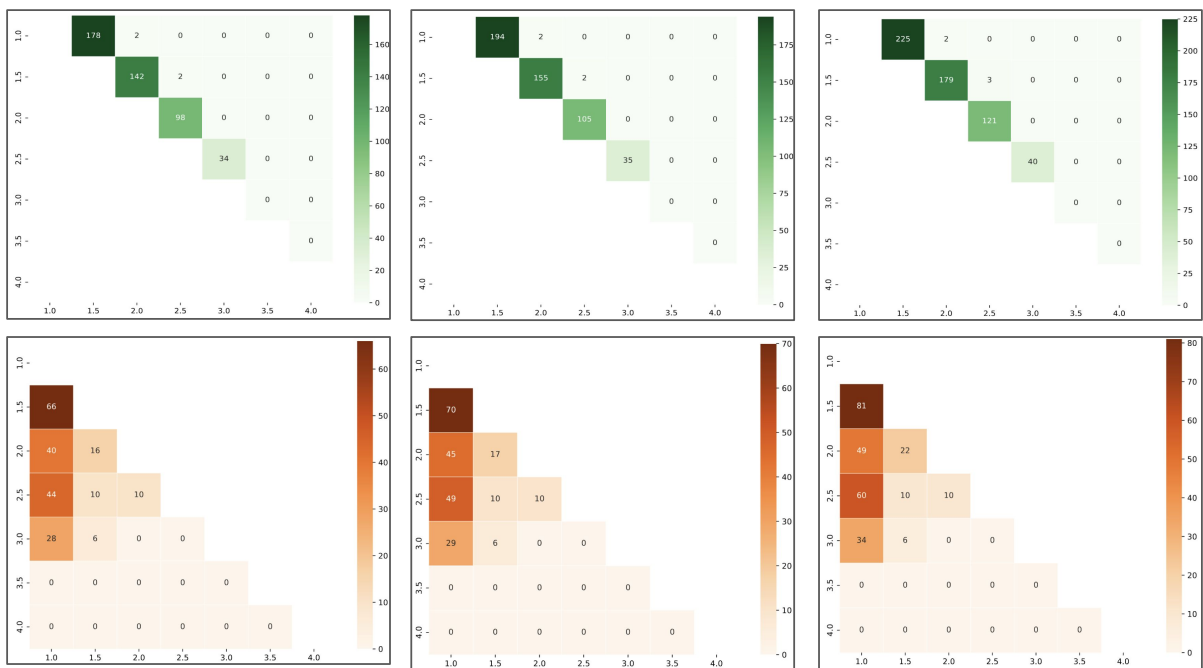


Figure 8.4: Illustration of the progression in trust scores, represented as either increasing or decreasing jumps. The y-axis denotes the starting trust score ("From"), while the x-axis indicates the resulting trust score ("To"). For instance, a progression from 2.0 to 2.5 corresponds to the (4, 5) cell in the upper triangular tables. Increasing trust score jumps are shown in green, while decreasing jumps are displayed in orange. Darker shades signify higher counts of the respective trust score changes.

8.7.2 Discussion on Trust Evolution Patterns

The trust level typically progresses through three stages: *refusal* (L1-L2), *digression* (L2-L3), and *opening-up* (L3-L4). Refusal involves reluctance or rejection of the therapist’s prompts. Digression occurs when peripheral topics are discussed instead of the core issue, and opening-up reflects direct engagement with the concern. For instance, a woman from initially dismisses discussing resentment toward her mother-in-law (refusal), later shares unrelated grievances (digression), and eventually addresses her hurt over personal comments (opening-up). This behavior is unique to this problem – trust modeling, and we further analyze this through the statistical point of view, as shown in Table 8.4.

Our analysis yields distinct patterns in how trust evolves during therapeutic conversations. Trust progression demonstrates an inherent positive inclination, with positive transitions occurring approximately 2× as negative ones, as building trust is a fundamental goal. Notably, *we observe that trust typically builds gradually*, with average positive changes of $\sim(+0.5)$ levels, while trust decreases tend to be more dramatic, with average $\sim(-1.0)$ levels. This asymmetry suggests that while trust is carefully cultivated in therapeutic relationships, it can be easily diminished by situational digression.

Another interesting finding is the presence of trust stability periods, averaging 8 – 9 consecutive utterances without level changes (c.f. *avg. Streak* in Table 8.4). These *trust plateaus* may represent stable therapeutic states where the patient-therapist bond has reached a temporary equilibrium. Figure 8.4 shows such plateaus in detail.

8.8 Limitations and Future Work

While our work studies trust modeling in counseling conversations, several limitations highlight opportunities for future research. The dataset is limited to interactions in English-language only, restricting its applicability in multilingual contexts and overlooking cultural nuances critical to therapeutic dynamics. We plan to address such gaps in future works by expanding the dataset to include diverse languages and cultural contexts. Additionally, the reliance on textual data alone omits other modalities, such as vocal tone and facial expressions, which are vital for capturing the full spectrum of trust dynamics. Incorporating these modalities into future models is vital for real-world applicability. Furthermore, we benchmarked various models but did not propose a method for trust modeling, which we plan to focus on in the future.

8.9 Ethical Considerations

Our work focuses on trust modeling with the primary objective of assisting therapists in developing more effective and adaptive therapeutic strategies. Importantly, this study is not patient-facing, ensuring that our study remains risk-free for patients. The data we use in this study is derived from a publicly available counseling dataset, which has already undergone necessary procedures to ensure user anonymity and protection of personally identifiable information (PII). Furthermore, every step of data construction and annotation has been carried out in close consultation with domain experts to ensure reliability and alignment with ethical standards. Despite the ethical practices we followed, we acknowledge the possibility of cultural biases due to its limited size and diversity. We explicitly address this limitation in our work and view it as an area for improvement in future iterations. Our study adheres to established ethical standards in data handling, privacy protection, and annotation.

8.10 Summary

In this paper, we introduced the novel concept of using trust as a metric to evaluate therapeutic strategies in text-based counseling. Our approach to quantifying trust dynamics leverages psycholinguistics and psychological theories, providing an actionable tool for therapists to calibrate their methods in real-time. We proposed a robust, annotated dataset containing trust ratings and topic-switch patterns from real counseling dialogues, which we used to benchmark state-of-the-art AI models. Our experiments demonstrated the viability of using AI to detect and quantify trust in therapeutic contexts, showing promising results in predicting patient engagement. The insights gained from this work lay the foundation for future advancements in automated mental health counseling support.

Part V

Peer Interactions

9. Understanding Peer Behavior

Online Mental Health Communities (OMHCs), such as Reddit, have witnessed a surge in popularity as go-to platforms for seeking information and support in managing mental health needs. Platforms like Reddit offer immediate interactions with peers, granting users a vital space for seeking mental health assistance. However, the largely unregulated nature of these platforms introduces intricate challenges for both users and society at large. This study explores the factors that drive peer engagement within counseling threads, aiming to enhance our understanding of this critical phenomenon. We introduce BeCOPE, a novel behavior encoded Peer counseling dataset comprising over 10,118 posts and 58,279 comments sourced from 21 mental health-specific subreddits. The dataset is annotated using three major fine-grained behavior labels: (a) intent, (b) criticism, and (c) readability, along with the emotion labels. Our analysis indicates the prominence of “self-criticism” as the most prevalent form of criticism expressed by help-seekers, accounting for a significant 43% of interactions. Intriguingly, we observe that individuals who explicitly express their need for help are 18.01% more likely to receive assistance compared to those who present “surveys” or engage in “rants.” Furthermore, we highlight the pivotal role of well-articulated problem descriptions, showing that superior readability effectively doubles the likelihood of receiving the sought-after support. Our study emphasizes the essential role of OMHCs in offering personalized guidance and unveils behavior-driven engagement patterns.

9.1 Introduction

The prevalence of mental health distress has risen sharply in the last several years. A recent report reveals that one in six individuals suffers from mental health-related challenges¹. At the same time, there is a severe shortage of mental health providers to facilitate adequate support to those in need² (325; 326). As a result of these growing challenges, we specifically examined the patterns and factors that drive individuals to engage with peer-to-peer mental health threads, focusing on the impact of behavioral, emotional, textual, and topical signals during peer-to-peer interactions.

To this end, we develop the BeCOPE (BEhavior enCOded PEer Counseling) dataset, composed of peer-to-peer mental health conversational interactions across 10,118 posts and 58,279 comments from 21 mental health-specific subreddits. We inspect the level of engagement on Reddit for three different OMHC categories – (a) interactive, (b) non-interactive, and (c) isolated – based on the pattern of interaction between users and the original help-seeker (see Figure 9.1). Analyzing the critical factors in each engagement category, we comprehend factors and patterns that lead to constructive versus detrimental peer-to-peer mental health interactions. Understanding peer-to-peer interactions on OMHCs is key to the ethical and safe monitoring of these communities, including the moderation of safe interactions and sharing of accurate mental health information. We explore the following research questions:

¹<https://www.who.int/news/item/17-06-2022-who-highlights-urgent-need-to-transform-mental-health-and-mental-health-care>

²<https://www.newamericaneconomy.org/press-release/new-study-shows-60-percent-of-u-s-counties-without-a-single-psychiatrist/>

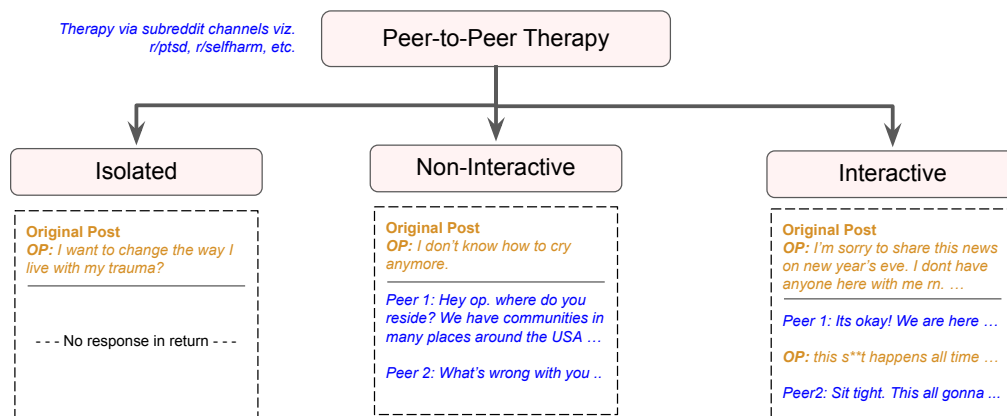


Figure 9.1: **Taxonomy of counseling methods along with examples.** Here, OP (original poster) is a common Internet terminology for the person who creates posts on peer-to-peer platforms. In peer-to-peer therapy, we inspect the level of engagement in three different categories based on the abundance of interaction with the help-seeker – **(a) interactive:** if there are back-and-forth conversations between the OP and peers, **(b) non-interactive:** if the post engages peers, but the OP does not reply to peers, and **(c) isolated:** if the post does not have any comment, but one-to-one therapy involves the continuous exchange of dialogues between therapist and client (help-seeker).

RQ1. When examining peer-to-peer OMHC interactions, how do intent (i.e., help-seeking), readability, and criticism impact peer willingness to engage with the original post (e.g., validation, advice-giving)?

RQ2. How does the expression of emotions in posts impact user engagement in the OMHC platforms?

Reddit is a popular OMHC platform that has steadily emerged as a platform for seeking help concerning a spectrum of mental challenges with specific posts devoted to disorders such as depression, attention-deficit/ hyperactivity disorder (ADHD, sometimes ADD), bipolar disorder and alcohol and substance use (327; 328; 329). Typically, users (i.e., support-seekers) create original posts to discuss their mental health issues, describing their symptoms and the contexts of their specific situations, like job loss or a recent divorce. The support-seekers, in turn, receive replies from peers (other users on the platform) with advice, recommendations for symptom management, and general support. This process allows support-seekers to share and ask for help for their mental health challenges in a cost-effective, convenient, and anonymous manner that typically results in immediate support. A recent study(330) analyzed patterns of posts on two popular OMHC platforms, Talklife and Reddit, by leveraging natural language processing for communication models in human-computer interaction and communication theory, operationalizing a set of four engagement indicators based on attention and interaction. The authors found that the back-and-forth peer platform communication effectively contributes to early support. A similar study(331) examined the change in sentiment to analyze peer-to-peer counseling settings to read whether a counseling thread or a post on the platform is correlated with a moment of cognitive change. It turned out that behavioral signals such as sentiment, affect, and topics associated with language are decisive toward effective counseling. On the same track, another study discussed the temporal engagement on social media correlating with patient disclosure(332). The authors developed an autoregressive time series computational model that assesses engagement patterns and subsequently forecasts alteration in the intimacy of disclosures. They found that attributes of audience engagement, like emotional support, personal behavior, and self-disclosure, strongly predict patterns in future counseling behavior.

Previous studies on the analysis of peer-to-peer mental health interactions identified threads that

Subreddits	Posts	Comments	Intent				Criticism				Readability	
			Help seeking	Rant	Survey	Chitchat	Self criticism	Other w/ Reason	Other w/o Reason	No criticism	Clear	Non-clear
r/Anxiety	469	1773	252	129	62	26	278	48	7	136	467	2
r/ptsd	494	1567	221	144	64	65	180	135	1	178	494	0
r/suicideWatch	403	2545	90	246	17	50	231	34	10	128	378	25
r/addiction	487	3581	217	148	43	79	246	67	6	168	466	21
r/ADHD	423	3856	169	104	78	72	139	31	9	247	418	5
r/alcoholicsanonymous	498	6021	181	107	47	163	155	58	5	280	490	8
r/Anger	464	2620	233	184	31	16	245	140	16	63	462	2
r/BPD	519	2744	180	185	113	41	234	99	4	182	518	1
r/depression	547	1951	83	363	26	75	243	91	18	195	546	1
r/domesticviolence	425	2847	254	94	25	52	34	277	1	113	421	4
r/eating_disorders	568	2021	256	209	51	52	346	43	1	178	567	1
r/getting_over_it	476	2551	230	163	35	48	258	72	2	144	473	3
r/mentalillness	484	1895	208	155	52	69	209	99	2	174	480	4
r/OpiatesRecovery	493	6112	215	116	62	100	185	28	3	277	493	0
r/rapecounseling	481	2390	288	142	26	25	125	269	1	86	481	0
r/sad	486	2258	44	287	27	128	115	71	8	292	485	1
r/selfharm	467	1928	136	232	52	47	243	39	0	185	465	2
r/selfhelp	419	2001	177	60	28	154	163	37	0	219	390	29
r/socialanxiety	461	2798	167	128	64	102	201	58	0	202	428	33
r/OCD	424	2528	159	117	63	85	209	29	3	183	424	0
r/helpmecope	473	2121	277	127	17	52	170	160	2	141	471	2
Total	9961	58108	4037	3440	983	1501	4209	1885	99	3771	9817	144
IAA (κ)	-	-	0.963				0.888				0.861	

Table 9.1: Statistics of the BeCOPE dataset. We collected a total of $\sim 10K$ posts and $\sim 50K$ comments. We annotated all the posts using three core labels – (i) intent, (ii) criticism, and (iii) readability (Clear: Excellent, Good and Average; Non-clear: Mediocre and Poor). IAA (κ) represents the inter-Annotator agreement using Cohen’s kappa score.

fall into affective (329), content-based (333; 334), and supportive (335) categories, thus demonstrating reliability for the functioning of peer-to-peer mental health platforms. However, little is known about how these categories of peer-to-peer mental health interactions are associated with constructive and/or detrimental outcomes. Understanding the characteristics of such OMHC users (336; 337; 338; 339) and given the widespread use of OMHC platforms, specific patterns and factors that drive engagement in peer-to-peer mental health interaction must be identified (340; 341; 342). In doing so, social media platforms should be better able to monitor and intervene for the benefit of their users in distress (343; 344; 345).

9.2 Dataset and Methods

9.2.1 Data Collection

To study latent signals in peer-to-peer mental health interactions, we develop BeCOPE by curating posts from 21 subreddits. Reddit is organized into spaces called subreddits, where each subreddit is specific to a certain discussion topic. To analyze behaviors on peer-to-peer mental health platforms, we scraped, processed, and annotated subreddit data to develop the dataset. We explored numerous subreddits and handpicked 21 most active mental health-related subreddits, as shown in Table 9.1. For each shown subreddit, we curated 500 posts and their comments from January 2020 to December 2020. Further, we performed a sanity check to ensure that conversations were acceptable (e.g., noise-free, written in English). We collected 10, 118 posts and 58, 279 comments along with their metadata, such as author information, score (upvotes), time of creation, and the number of comments.

Step 1: Categorization of interactions by the level of peer engagement. Depending on the comments on a post, we classified the collected conversations into one of the three engagement categories: (i)

interactive, (ii) non-interactive, or (iii) isolated. If an original post involved back-and-forth comments from the original user and peers, the conversation was deemed “interactive”. If an original post had zero comments, the conversation was deemed “isolated.” Finally, if an original post received more than one comment from peers, but the original user did not acknowledge or reply to peers’ comments, the conversation was deemed “non-interactive”.

Step 2: Annotation of posts by behavioral and emotional labels. The first step in the annotation process was the curation of Reddit posts on mental health topics by categorizing them based on (i) intent, (ii) criticism, (iii) readability, and (iv) emotion labels. We manually annotated ~5K posts and subsequently learned respective classifiers to obtain pseudo-labels for another ~5K posts. Next, a sanity check of the annotated dataset was performed to ensure the reliability of the annotations. Finally, we used the resultant dataset of ~10 posts for our analyses. Detailed statistics of the annotated BeCOPE dataset (including pseudo labels, discussed later) are presented in Table [9.1](#).

9.2.2 Data Annotation

Peer-to-peer counseling conversations are open-ended, where users express their diverse and different perspectives. We observe that users exhibit a variety of intents while discussing mental health issues *viz.* rating, seeking help survey, or doing general chit-chat. On the other hand, some alleged criticism for their issues. Moreover, another aspect of peer-to-peer counseling is to understand whether users convey their needs clearly and crisply. A well-written post may have attracted more and perhaps better responses than a poorly-written post. Therefore, the readability of the post is another factor that directly affects the chances of receiving help on a mental health post from peers. At the same time, studying the expressed emotions is highly imperative. We hypothesize that all the above-discussed factors – the knowledge of intent, the presence of criticism, the study of emotions, and the readability of the posts, are crucial in understanding the need of the help-seeker and accordingly providing appropriate assistance. Considering the literature and observations, we designed a set of guidelines to annotate the curated Reddit posts. A detailed discussion of the guidelines considering the four factors is presented below.

Intent

Intent defines the purpose of the original poster (OP) in the post. We divide the posts into four categories based on the user’s needs: help-seeking, rant, survey, and chitchat.

- **Help-Seeking:** Original posters explain mental health issues and expect peers to provide helpful suggestions to improve their condition.
- **Rant:** Original posters share their (strong) views on mental health issues without expecting help from peers.
- **Survey:** Original posters share mental health issues and ask peers to share their experiences. *Survey* differs from *help-seeking* as survey-labeled posts ask for a generic point of view on related mental health issues rather than individual-centric assistance.
- **Chitchat:** The Chitchat label is used for filler posts that are not directly related to mental health issues. Such posts include well wishes, general guidelines, occasional greetings, etc.

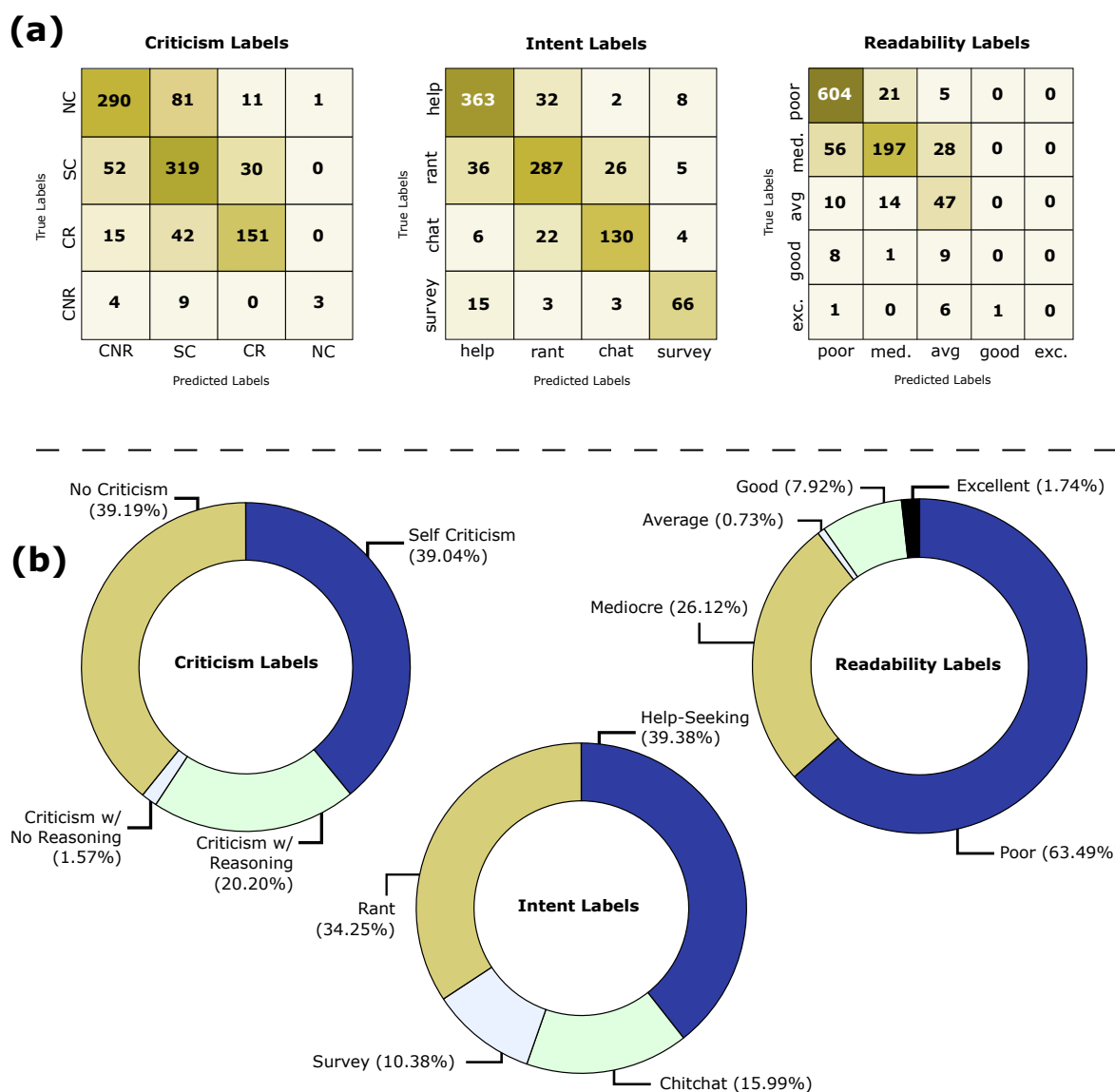


Figure 9.2: **(a)** Confusion matrix to represent the performance of pseudo labeling of criticism, intent, and readability labels. We exploit BERT to fine-tune on ~5K manually annotated posts to predict criticism, intent, and readability on the remaining posts. **(b)** Distribution of behavioral signals (criticism and intent) along with readability in the complete BeCOPE dataset.

Criticism

Original posters often criticize the situation caused due to their or others' mental health issues. Sometimes, criticism is on their own; other times, it is on others. Hence, it is important to study if showing criticism could be a cause to receive better help. In other words, do peers prefer helping others who use criticizing language in posting their mental state? To understand this, we define four criticism labels: *no-criticism*, *self-criticism*, *others' criticism with reason* and *without reason*.

- **Self-Criticism (SC).** We use this label for posts where original posters criticize themselves for their mental health issues.
- **Criticism on Others with Reason (CR).** We use this label for posts where original posters criticize

others for their mental health issues. Also, they provide some reasons (justification) to support their criticism.

- Criticism on Others with No Reason (CNR). This label differs from CR as the criticism is not backed by reasoning.
- No-Criticism (NC). We use this label for posts where there is no criticism.

Readability

Readability is essential in imparting most of the information via textual communication in all professional domains (346). Earlier works utilized the readability criteria to decide the impact of a mental health post using statistical properties of posts such as length of the post (330). However, we argue that a shorter post could also be interpreted as poorly readable. Therefore, in this work, we define the readability score based on the clarity of the text and the amount of effort one needs to put into comprehending the post. We observe that lengthier sentences pose a degree of uneasiness in readers besides the use of SMS slang and abbreviations. Based on our observations, we define five readability levels for a post – *excellent*, *good*, *average*, *mediocre*, and *poor*.

Emotion

Emotion labeling is the practice of cultivating empathetic knowledge in conversations (347). We employ a set of 28 emotion classes – *admiration*, *amusement*, *anger*, *annoyance*, *approval*, *caring*, *confusion*, *curiosity*, *desire*, *disappointment*, *disapproval*, *disgust*, *embarrassment*, *excitement*, *fear*, *gratitude*, *grief*, *joy*, *love*, *nervousness*, *optimism*, *pride*, *realization*, *relief*, *remorse*, *sadness*, *surprise*, and *neutral*– for our Reddit posts. Moreover, we observed that many posts conveyed multiple emotions in a single post; hence, we assigned two emotions for each post, i.e., the *primary emotion* and the *secondary emotion*.

Ethical Consideration

Considering the sensitivity of research in mental health, this paper does not include any personal, identifiable information of any OMHC user. Further, our models involve sophisticated deep-learning models, which are careful not to take any bias toward any gender, caste, race, diagnosis, or peers with specific symptoms. We collected data solely based on the most relevant mental health subreddits and did not include any bias in the choice of particular subreddit channels. Finally, we conducted all experiments without compromising the anonymity of online users in BeCOPE.

9.3 Analysis and Findings

RQ1: When examining peer-to-peer OMHC interactions, how do intent (i.e., help-seeking), readability, and criticism impact peer willingness to engage with the original post (e.g., validation, advice giving)?

Intent. We observe that help-seekers on OMHC platforms are 18.01% more likely to receive help when they explicitly convey their pressing needs through queries, as opposed to when they make statements

Post	Intent	Criticism	Emotion		Read
			Primary	Secondary	
Hello all. I unfortunately used again today, despite going to my first NA meeting last night. My loneliness, is a trigger. I spent the whole day by myself, and decided it wouldn't hurt to light up one more time. I have a job interview, next Monday. I don't want to lose my life. I am usually a proud person, and solve my problems on my own. But, I know I need to reach out to someone before I spiral out of control. I honestly see myself giving up everything for it. Please, any advice and tips would help me in this low moment. I have no one in my life currently. Thank you. [Interactive]	<i>Help-seeking</i>	SC	<i>Sadness</i>	<i>Gratitude</i>	<i>Excellent</i>
I recently found out my boyfriend has been crushing or opening then snorting most of his prescription medications. We had been arguing non-stop over his marijuana addiction, but this brings it to the next level. Not just his ADHD meds, but his depression and anxiety meds too. He claims it makes them more effective. I tried negotiating with him over letting me dispense his ADHD meds to him one week at a time, and he goes on tirades about loss of control/lack of trust/T'm not his mother/the government and his doctors can't control him etc. I already am exhausted from our constant fighting over marijuana abuse. He says it's not my problem because it's his body and his decision. He has ADHD, depression, and anxiety concurrent with substance abuse disorder (marijuana and alcohol). He has been hospitalized and gone to rehabilitation multiple times. Now he's adding RX abuse to his problems. I told him I was considering reporting him to his doctors, and he told me to keep my mouth shut and let him make his own decisions. I think his illness keeps him from making logical decisions about this, as he is constantly seeking ways to "not be bored", "kill the anxiety", or "not feel anything". TLDR: SO says his decision to abuse prescription medications is not my problem, despite his diagnosed substance abuse disorders. [Interactive]	<i>Rant</i>	CR	<i>Sadness</i>	<i>Disappointment</i>	<i>Average</i>
I live in Canada. I started the application process for medical assistance in dying. It was nice to finally receive an intelligible response to "I really want to die." as opposed to the usual parroted phrases and hollow cheerleading. Not sure that I will go through with it tbh, but I hope that I do. Getting two witness signatures seems like the only real obstacle here. [Isolated]	<i>Chitchat</i>	CNR	<i>Optimism</i>	<i>Joy</i>	<i>Excellent</i>
Who's an alcoholic stay at home wife/mom? Just super interesting to me. What are your days typically like? Do you hide your consumption or own it? [Non-Interactive]	<i>Survey</i>	NC	<i>Excitement</i>	<i>Neutral</i>	<i>Excellent</i>
Every day I read to you from Daily Reflections. This book is published by Alcoholics Anonymous, and is an important resource for members of this LIFE-saving association. Many recovering alcoholics use this literature to start their day. AA is a fellowship of men and women who share their strengths and hopes with each other for the common purpose of helping the alcoholic who still suffers. It is the faith and the love of the LIFE which allows the cure or rather ashes the progression of the disease and helps us to leave the hell of alcoholism one day at the time. Addiction is very powerful and sneaky. Although these writings are primarily intended for alcoholics, their families and friends, many people who feel they have no contact with alcoholism greatly appreciate the wisdom that emanates from them. We hope you will gain the freedom we know! MERCI! Thank you for your support by subscribing to our YouTube channel. <youtube channel> #recovery #alcoholism #alcoholics anonymous love #LIFE #January [Isolated]	<i>Chitchat</i>	NC	<i>Neutral</i>	<i>Neutral</i>	<i>Poor</i>

Table 9.2: Example of posts and their corresponding labels in BeCOPE. **Intent:** *Help-seeking*, *Rant*, *Chitchat*, and *Survey*; **Criticism:** *Self-criticism* (SC), *Criticism with reasoning* (CR), *Criticism with no reasoning* (CNR), and *No-criticism* (NC); **Readability:** *Excellent* (5), *Good* (4); *Average* (3), *Mediocre* (2), and *Poor* (1); **Emotion:** *Admiration*, *Amusement*, *Anger*, *Annoyance*, *Approval*, *Caring*, *Confusion*, *Curiosity*, *Desire*, *Disappointment*, *Disapproval*, *Disgust*, *Embarrassment*, *Excitement*, *Fear*, *Gratitude*, *Grief*, *Joy*, *Love*, *Nervousness*, *Optimism*, *Pride*, *Realization*, *Relief*, *Remorse*, *Sadness*, *Surprise*, and *Neutral*; **Engagement:** *Interactive*, *Non-interactive*, and *Isolated*.

about their experiences. When an original post contains a help-seeking approach, it increases peer engagement. Specifically, 45.35% of interactive posts, 42.16% of non-interactive, and 27.34% of isolated posts are help-seeking in nature, indicating that peers who explicitly ask for help for their mental issues experience greater peer engagement. We also observe that when an original post is constructed as a “rant” (a long statement of the problem with no explicit ask for help/advice), it receives less peer engagement. The number of isolated posts labeled with the rant intent (38.11%) exceeds non-interactive posts (34.73%) and interactive posts (32.56%) by a margin of 3.38% and 5.55%, respectively. Further, posts with rant intent receive the least interaction compared to other intent labels across all engagement categories, showing that the survey posts do not elicit peers’ attention toward assistance. Our analysis sheds light on RQ1 by indicating the conveyance of explicit intentions through queries or the articulation of pressing needs on the OMHC platforms yields a more efficacious response. We present the distribution of intents across three engagement categories in Figure 9.3(a). The four annotated intent labels achieve statistically significant Cohen’s Kappa agreement scores with a confidence $\geq 95\%$ on the p -values of help-seeking (0.022), rant (0.046), chitchat (0.016), and survey (0.028).

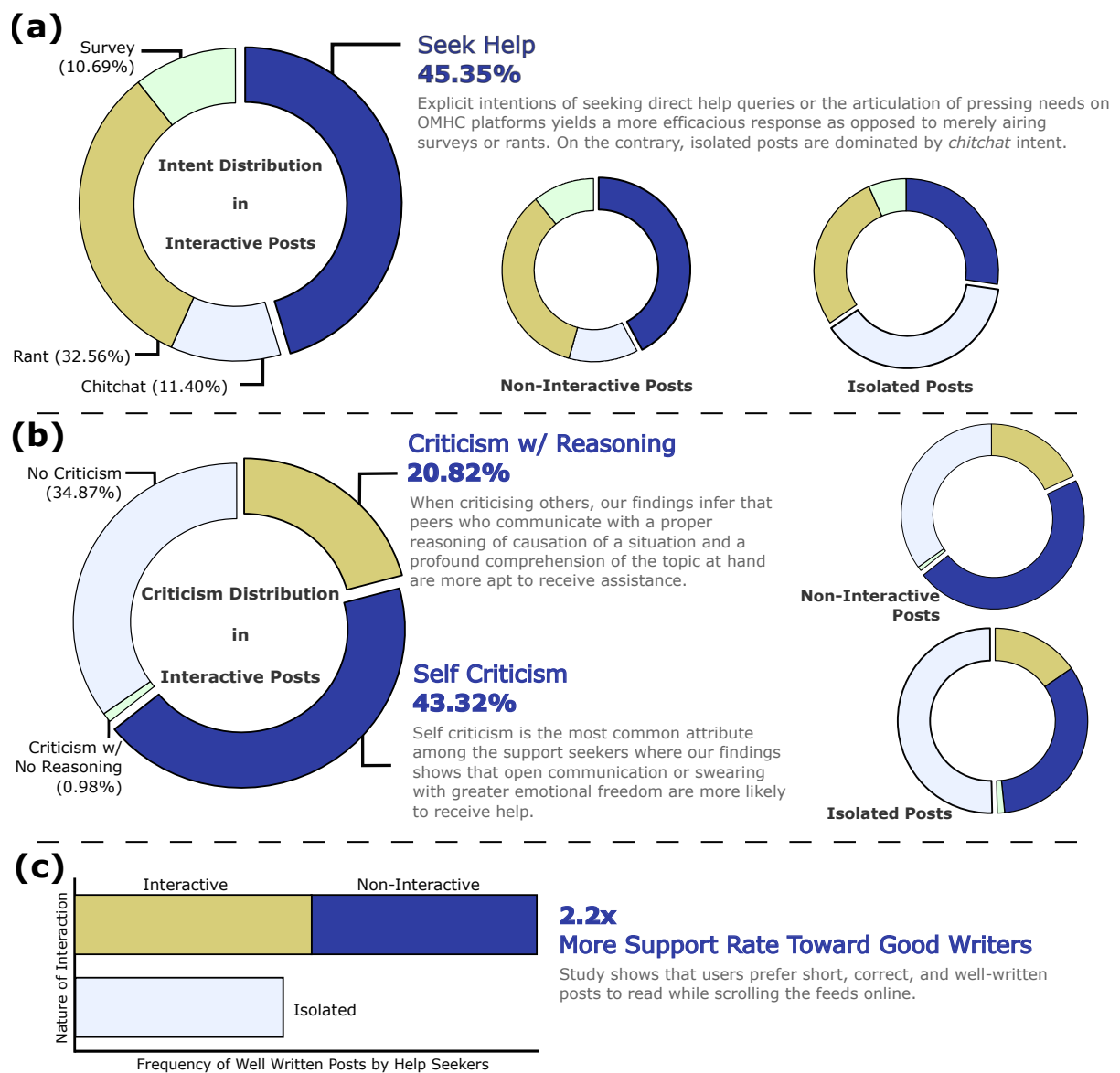


Figure 9.3: **Distribution of behavioral signals and readability in BeCOPE across all engagement categories.** (a) The intent distribution indicates that a majority (45.35%) of posts show explicit intentions (seek-help) through queries or the articulation of pressing needs on OMHC platforms, yielding a more productive response as opposed to merely airing surveys or rants. (b) The criticism distribution shows that help-seekers are more likely to engage in self-criticism (43.32%), and those who criticise openly on others with proper reasoning are more likely to receive assistance. (c) The readability statistics of posts in BeCOPE state that well-written posts receive 2.2 \times more support (responses) as compared to poorly written posts.

9.4 Experiments and Results

RQ1: When examining peer-to-peer OMHC interactions, how do intent (i.e., help-seeking), readability, and criticism impact peer willingness to engage with the original post (e.g., validation, advice giving)?

Intent. We observe that help-seekers on OMHC platforms are 18.01% more likely to receive help when they explicitly convey their pressing needs through queries, as opposed to when they make statements about their experiences. When an original post contains a help-seeking approach, it increases peer engagement. Specifically, 45.35% of interactive posts, 42.16% of non-interactive, and 27.34% of isolated posts are help-seeking in nature, indicating that peers who explicitly ask for help for their mental issues experience greater peer engagement. We also observe that when an original post is constructed as a “rant” (a long statement of the problem with no explicit ask for help/advice), it receives less peer engagement. The number of isolated posts labelled with the rant intent (38.11%) exceeds non-interactive posts (34.73%) and interactive posts (32.56%) by a margin of 3.38% and 5.55%, respectively. Further, posts with rant intent receive the least interaction compared to other intent labels across all engagement categories, showing that the survey posts do not elicit peers’ attention toward assistance. Our analysis sheds light on RQ1 by indicating the conveyance of explicit intentions through queries or the articulation of pressing needs on the OMHC platforms yields a more efficacious response. We present the distribution of intents across three engagement categories in Figure 9.3(a). The four annotated intent labels receive a significant agreement score with a confidence $\geq 95\%$ on the p -values of help-seeking (0.022), rant (0.046), chitchat (0.016), and survey (0.028).

Criticism. We observe that isolated posts have maximum *no-criticism* (NC) labels (50.34%) as compared to non-interactive (34.92%) and interactive (34.87%) posts. Figure 9.3(b) shows the distribution of *criticism* labels across all engagement categories. Conversely, individuals who can obtain support from their peers on OMHCs are frequently found to engage in criticising themselves and others. We bifurcate the criticism of others into two indicative categories – *criticism with reasoning* (CR) (i.e., a logical presentation of one’s experience), and *criticism with no-reasoning* (CNR). Out of all three engagement categories, interactive engagement carries the maximum CR label, 2.75% and 5.39% more than non-interactive and isolated engagement categories, respectively. This trend directly draws attention to the fact that proper reasoning in criticism is vital for receiving help. In contrast, CNR is most prevalent in isolated engagement, highlighting that criticism without proper reasoning only adds a noisy understanding to the reader’s mind. Similarly, *self-criticism* is considered the most prevalent type of criticism among those who receive help. This implies that people seeking support are more likely to engage in self-criticism, and those who express their emotions more openly are more likely to receive assistance. As a result, we infer that peers who criticise and have a profound comprehension of the topic at hand are more apt to receive assistance. The four annotated criticism labels receive adequate agreement score with confidence $\geq 95\%$ on the p -values of *criticism w/ reasoning* (0.043), *criticism w/ no reasoning* (0.010), *no criticism* (0.009), and *self-criticism* (0.035).

Readability. We hypothesize that well-written posts (i.e., easier to read) foster better understanding and subsequently attracted more peers to engage. Our initial observation supports the hypothesis; most of the posts in the BeCOPE dataset are hard to read, i.e., rated ≤ 2 on a scale of 1 to 5, with 1 being the least comprehensible. Our analyses reveal that posts scoring higher in readability result in $2.2\times$ greater support ratings from peers, as shown in Figure 9.3(c). We further employ experts in linguistics to understand what contributes more toward understanding posts. We observe that factors like the length of the post, the

division into paragraphs and listicles, grammar, spelling, clarity of the issue, and usage of short forms (SMS language) are critical that peers take into consideration when reading and deciding to engage with a post. The readability score receives significant confidence of $\geq 95\%$ with average p -values across all five labels to be 0.040.

RQ2: How does the expression of emotions in posts impact user engagement in the OMHC platforms?

Emotion labels. Emotions play a vital role in mental health support seeking. Empathetic understanding is an attempt by the observers/experts to regulate emotions that help-seekers express(348). Figure 9.4 shows a frequency-based radial distribution of the most frequent emotion labels in BeCOPE. Our analysis of emotion labels shows that 10% of the isolated posts carry *neutral* emotion labels. In contrast, only 3% posts carry *neutral* emotions for both interactive and non-interactive posts combined. Furthermore, 12.3% of the non-isolated posts exhibit *curiosity* as the secondary emotion compared to 7% isolated posts. Evidently, labels such as *sadness*, *curiosity*, *fear*, and *realization* are more prevalent in non-isolated posts. On the other hand, emotion labels such as *caring*, *confusion*, *approval*, *joy*, and *neutral* are more prevalent in isolated posts. Consequently, peers exhibiting explicit emotional expression in posts, such as curiosity, fear, and sadness, receive more significant support in 86% of the cases. For the remaining 14% of the posts, emotions are observed to be with tepid emotional labels, such as caring, confusion, or neutral, to which peers often ignored responding, leading to no interaction.

On analyzing a sample of 100 posts, we subjectively categorize extreme emotions expressed into various types, including fear, excitement, sadness, etc. In the category-wise emotion distribution (Figure 9.4), we observe that posts expressing such explicit extreme emotions have a higher chance of receiving a response, whereas posts with tepid emotional labels, such as caring, confusion, and neutral tend to be ignored.

9.4.1 Metadata and Content Analysis

We conduct an auxiliary analysis of the BeCOPE dataset with a prime focus on metadata and textual properties. These experiments aim to assess the impact of minor actions, such as subjectivity, interaction count, anonymity, etc., on help-seeking. We conclude that specific minor actions taken by help-seekers on OMHC platforms can increase the probability of receiving assistance. Our initial findings suggest that descriptive titles and body content attract more help than compact usage of words. Likewise, the active participation of the help-seeker in the conversation (through comments) increases the chances of receiving help two-fold. Such approaches might assist help-seekers in gaining early access to assistance. We observe that a few mental health subreddit channels like *r/OpiatesRecovery* are entirely dedicated to providing frequent assistance to help-seekers, including during late hours.

9.4.2 Topical Analysis

We also perform a topical analysis of peer-to-peer interactions, aiming to understand what specific topics and keywords drive the conversation in three engagement categories (viz. interactive, non-interactive, and isolated). To this end, we apply Latent Dirichlet Allocation (LDA)(2) on the posts in each engagement category. The idea is to understand the topics on which peers respond and don't respond. Therefore, we segregate isolated and non-isolated posts to study the topics on which support is received and not received, respectively. We observe that the most frequently occurring topics in isolated posts include discussions

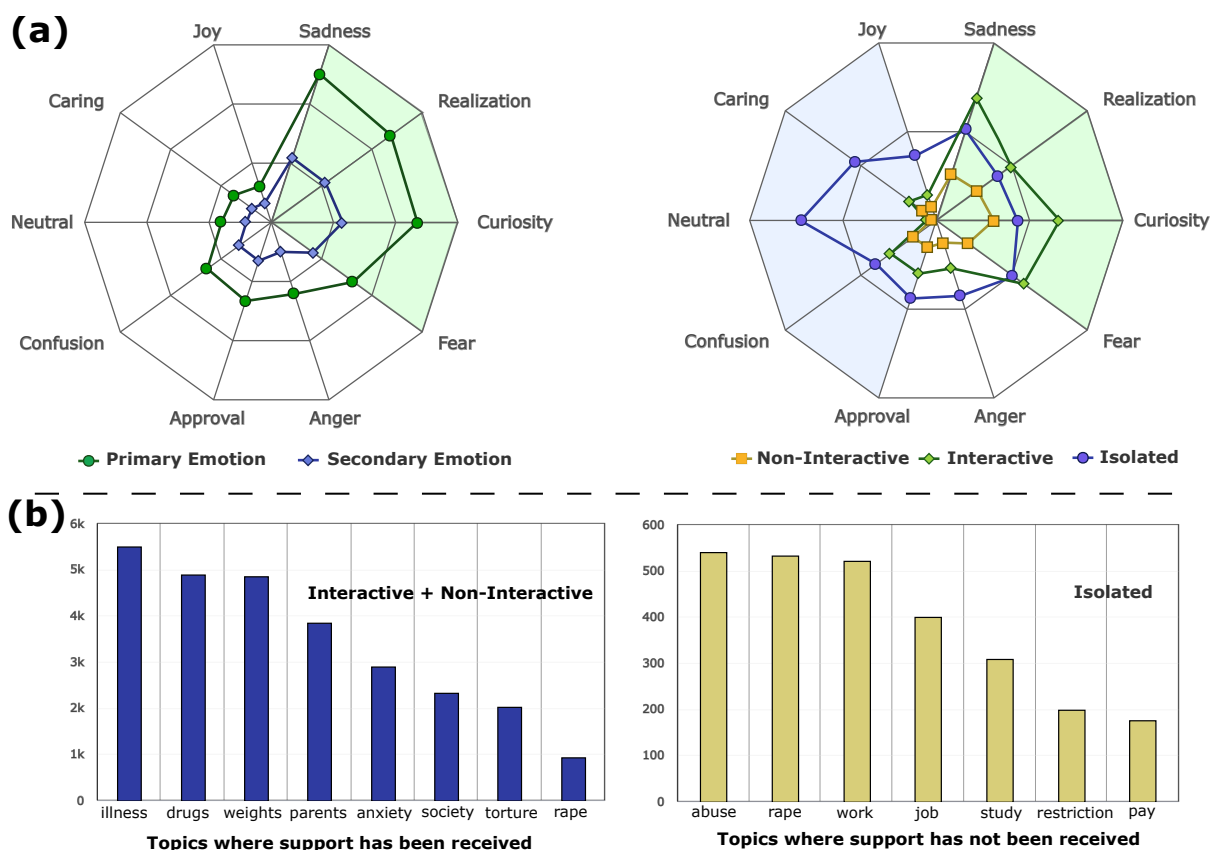


Figure 9.4: **(a) Distribution of emotion labels in the BeCOPE dataset.** For brevity, we show plots for the top 10 emotion labels only. Each post is tagged with primary and secondary emotion labels. We further analyze the emotion label distribution across three engagement categories. **(b) Topical analysis on the BeCOPE dataset.** We perform Latent Dirichlet Allocation (LDA) (2) to form 8 clusters of topics. To analyze the topics on which peers respond, we club interactive and non-interactive posts, where peers respond and compare them with topics from isolated posts.

about school-related issues, abuse, rape, pressure to meet society’s standards, salary, and freedom to express opinions and feelings. On the other hand, we observe that the frequently discussed topics from the non-isolated category are anxiety, drugs, common symptoms/illness and diagnosis, parenting behaviors, body image issues, food and weight, anxiety, and relapsing on drugs. Figure 9.4 shows a cluster of topics for posts from each category to obtain the most common topics in conversations. Evidently, the common topics of discussion in isolated posts elucidate that people shared experiences about many sensitive and stigmatized issues; subsequently, they remain unexplored, as indicated by the number of isolated posts. As a result, topics that resonate with peers and enjoy widespread prevalence tend to attract more interactions and are more likely to receive active engagement from peers on OMHCs.

9.4.3 Additional Auxiliary Analysis

This section analyses the auxiliary properties of the BeCOPE dataset. We run fundamental experiments to analyze BeCOPE’s focus on metadata and textual properties.

Title Length in BeCOPE

We run a short analysis on the length of the titles of posts from each category in BeCOPE. The aim is to comprehend if more descriptive titles are helpful while receiving help or make it cumbersome for helpers to skip over the post. We find that the mean length of the title for interactive posts is greater than that of non-interactive and isolated posts by a margin of 3.4 and 5.5 words, respectively. These statistics indicate that descriptive titles are more explanatory and likely to receive help. We further conclude that the original poster, who wrote descriptive titles, is likelier to engage with the peers trying to help. This explains the fact that peers primarily look for self-explanatory titles to decide to help.

Body Length in BeCOPE

Once a descriptive title catches the helper's attention, we hypothesize that the body of the text should also be descriptive enough to impart proper information to receive help. To verify this, we run a simple analysis of the body lengths of the post. Furthermore, the statistics show that the mean length of the body for interactive posts (189.95) is greater than that for non-interactive and isolated posts by a margin of 25.64 and 35.24, respectively. This supports our hypothesis, indicating that help-seekers who write descriptive posts are more likely to receive help. We also analyze that the mean body length for non-interactive posts is greater than for isolated posts by a margin of 9.6. Once again, this indicates that posts with longer body lengths are more likely to attract helpers by imparting information better.

The Number of Peer Comments

Here, we analyze how many peers interact with the posts. On average, an interactive post receives 10.12 comments, whereas a non-interactive post receives merely 3.41 comments. Evidently, the number of peer comments on interactive posts is tripled compared to the peer comments on non-interactive posts by a margin of 6.71. This analysis significantly differentiates the interactivity among the interactive and non-interactive posts on mental health subreddits. Moreover, the average number of unique peers interacting on the interactive posts is more than two times the number of non-interactive posts. The same trend holds for other statistical measures. For interactive and non-interactive posts, we observed a standard deviation of 12.47 and 5.94, median of 4 and 2, and 95th percentile of 22 and 10, respectively.

The Number of Unique Peers on a Thread

The engagement category influences the number of peers participating in a conversational thread. Our findings state that an average interactive post attracts 6.07 unique participants on the thread, and a non-interactive post receives merely 3.06. This shows that the average number of unique peers on interactive posts is almost double that of peers interacting on non-interactive posts. The help-seeker interacts with the peers by writing replies to their initial replies. The result indicates that the peers are more interactive, and the number of new peers pitching in to help also increased. Consequently, the help-seeker's interactivity also increases the post's overall engagement, resulting in an increase in the chance of receiving assistance.

Subreddit Division

We assess which mental health issues are more prevalent in peer-to-peer counseling. Also, it is an interesting comparative study to analyze which mental health issues usually attract interaction. Evidently,

*r/helpmeco*pe contributes to the minimum number of isolated posts out of all subreddits and the maximum number of non-interactive posts. The subreddit *r/OpiatesRecovery* contributes the maximum number of interactive posts to the dataset. The number of isolated posts is greater than that of non-interactive posts, which in turn was greater than the number of interactive posts for 13 out of 21 mental health subreddits in the BeCOPE dataset.

Anonymity

Reddit allows users to create *throwaway* accounts, which are temporary accounts for creating just an anonymous post. Thus, any content posted via an anonymous account does not reveal the user's personal account information. To explore the effect of anonymity on peer-to-peer counseling conversations, we extract all the posts with the throwaway user handles from BeCOPE. Next, we analyze these posts for differentiating characteristics. We find 245 posts from throwaway accounts in BeCOPE. Furthermore, 28 out of 245 posts are isolated, 101 are non-interactive, and 116 are interactive. Throwaway accounts contribute to a meagre 2.4% out of total posts. This indicates that help-seekers barely use the added anonymity offered by the OMHC platforms for posting mental health issues. Since it is clear that the posts are less likely to receive a reply if posted from throwaway accounts, we explore the alternate possibility of checking if the help varies with anonymity (Case 1). We also verify if peers' behavior while replying changes when they move from a regular to an anonymous account (Case 2). To this, we test the following two cases:

Case 1: Posts made from throwaway accounts are less likely to receive help as compared to the posts made from any regular account.

Case 2: Help-seekers posting from throwaway accounts are more likely to reply on receiving replies than those posting from regular accounts.

We observe that 88.5% posts by throwaway accounts received a reply compared to 81.3% posts that were not from throwaway accounts. Thus, the finding does not support our hypothesis. Moreover, the original poster replied in 53.5% of the posts from throwaway accounts, which received at least a reply. This is similar to the 53.8% posts not from throwaway accounts. Consequently, we do not observe any correlation between the behavior of the help seeker and anonymity; thus, the second case also contradicts the findings.

Analysis Based on Similarity

Earlier work showed that better responses from counselors uphold higher linguistic similarity with the clients' responses. Thus, we test the hypothesis to observe if the interactive engagements carry a higher similarity between the post and the comments than other engagements. Hence, to shed light on this and further clarify, we exploit sentence-bert (349) to compute the cosine-similarity between the context-rich BERT embeddings of help-seeker's post and peer comments. As a result, we observe that the earlier work's hypothesis contradicts the findings of our analysis. The average text-similarity between posts and comments for non-interactive posts is very similar to that of interactive posts with a slight difference of 0.0202 only. Thus unlike earlier work, the high textual similarity between the post and the comments is not a standard case.

9.5 Discussion

Understanding user behavior and online engagement is consistently challenging, particularly in comprehending the complexities of individuals in distress. OMHC platforms have emerged as crucial spaces for peer-based mental health discussions, enabling individuals to discuss their intrinsic thoughts and mental health issues openly. Beyond the OMHC's function, only a handful of these users interact, with even fewer users receiving the anticipated assistance. The most effective way of assessing peer engagement is to understand the factors on which peer interaction depends. Platforms like Reddit, containing dedicated mental health subreddits, offer rich repositories of discussions on relevant topics. Our formulated hypothesis posits that the comprehension of peer behavioral attributes such as intent, criticism, and readability significantly contributes to a holistic understanding. In addition, the expressivity of emotions on OMHCs can further concentrate on the causal underpinnings of these behavioral dynamics. However, this research area has remained under-resourced and insufficiently explored. Our newly introduced BeCOPE dataset holds significant implications beyond the insights drawn in this study. It can serve as a valuable resource across various research domains with dimensions ranging from empathetic to behavioral conduct of peers on OMHCs and further epitomizing explanations and causality of such implicit underlying causes.

Our research examines the behavioral, emotional, and topical dynamics associated with varying levels of engagement among peers within OMHCs. We perceive engagement as an indication of a peer's preparedness to provide support. Our findings underscore that simple behavioral characteristics such as explicitly seeking help and refraining from criticizing others can increase peer engagement, as observed in ~50% of the cases. This observation emphasizes that behaviors like ranting, criticising others, and generic chit-chatting do not elicit productive peer attention. At the same time, users express themselves in different styles, and the underlying concept of peers being able to understand others hinges on the clarity of the posts' readability. Earlier research shows that using short sentences is more engaging (350). In contrast, we show that peers with intricate thoughts aren't constrained to concise posts; instead, they often require more extensive elaboration (351). Our research demonstrates a twofold increase in support for individuals openly expressing their concerns on the OMHC platforms. Conversely, the illustration of emotion dynamics is an additional gauge to evaluate the user's context. In alignment with our formulated hypothesis, the intricate interplay of emotions articulated within OMHCs demonstrated a direct correlation with the level of peer interaction. Analogous to socio-cultural implications, instances where individuals convey heightened emotional intensity consistently involve more engagement, while expressions characterized by emotional neutrality tend to diminish in terms of peer involvement. This phenomenon potentially stems from underlying factors such as relatability, the emergence of a palpable sense of urgency, and a compelling inclination to provide empathetic validation and support. These emotionally charged interactions establish a conspicuously relatable presence, effectively motivating peers to participate in discussions and disseminate adaptive coping techniques actively. Consequently, the assessment of peer engagement within OMHCs stands as pertinent societal research that aims to assess the intricate dynamics underpinning an effective peer support framework. Such OMHCs serve as forums where peers engage in a wide spectrum of discussions, yet only a few receive the required assistance. We are convinced that a crucial void in this landscape lies in fostering societal awareness regarding the nature of these challenges and their appropriate navigation. For instance, individuals often discuss sensitive and stigmatized matters, which, although prevalent in volume, remain relatively unexplored, as substantiated by the prevalence of isolated posts. As a result, topics of a more general nature are observed to attract increased interaction. These perceptive insights inherently underscore the significance of understanding the factors of the support ecosystem before its effective utilization for constructive engagement.

9.6 Summary

OMHC platforms have become a popular way to seek help for people struggling with mental health issues (352; 353; 354; 355). Our work analyzed the granular user posting behaviors that foster peer engagement with the mental health content on OMHC platforms, specifically subreddits. The primary aim of this work was to better understand the behaviors of support seekers and the factors that drive peer engagement with the original post. We found that the intent of a post (seeking support versus ranting about one's experience), the readability, and the criticism elements of a post were associated with peer engagement. Our proposed dataset and empirical study call for more research to understand peer engagement on mental health platforms, including elements that lead to constructive versus detrimental engagement(354; 356; 357). These data are critical in understanding how OMHC can best support users experiencing distress in addition to preventing the proliferation of harmful and inaccurate mental health advice and information(358; 359; 360).

Understanding user behavior and online activity is challenging, and even harder to understand individuals in distress. The current study primarily focused on peer-to-peer engagement concerning mental health content. We understand that the findings can vary across other platforms like Twitter, Talklife, 7Cups, Facebook, Instagram, and even other subreddit channels. The future direction of this work will be to better understand user behavior on OMHCs, including how to monitor and moderate peer engagement so that it is not harmful to individuals in distress. Although our findings shed light on the connecting patterns of peer-to-peer online engagement, more research is needed to develop computational methods to gauge user satisfaction and behavior by exploiting the annotations we have done in BeCOPE.

10. Conclusion and Future Work

With this chapter, we now summarize the contributions made by this thesis and a discussion of possible directions in which future research could be extended.

The increasing demand for mental health support, exacerbated by the COVID-19 pandemic, has exposed the severe shortage of mental health professionals worldwide. As a response, Virtual Mental Health Assistants (VMHAs) have emerged as a promising solution to bridge this gap. However, despite their potential, existing dialogue systems remain rigid, contextually limited, and incapable of replicating the nuanced adaptability of human therapists. This thesis aimed to deconstruct the counseling interaction pipeline into its fundamental components, such as understanding, summarization, response generation, trust evaluation, and peer interactions, to develop modular novel methods that enhance virtual counseling. Through exhaustive experimentation, expert collaboration, and user evaluations, this thesis presents computational models that address the research gaps within digital mental health support, ensuring that VMHAs are not merely automated conversational agents but effective, structured, and clinically informed assistants capable of assisting mental health professionals in a typical counseling procedure.

A key contribution of this thesis is dialogue understanding, where we addressed the fundamental challenge of interpreting directive understanding in counseling dialogues. To this end, we developed the HOPE dataset, an extensively annotated dataset for the dialogue-act classification task in dyadic counseling conversations. We proposed SPARTA, a novel transformer-based model incorporating speaker-aware and time-aware joint contextual learning mechanisms. The model effectively leverages global and local context, improving the performance by 58.46%. This work lays the foundation for developing more contextually aware VMHAs that can better engage with clients and respond appropriately to their psychological needs. Beyond understanding conversations, another critical aspect of therapy is counseling note-making. We attempted to digitize the process as a dialogue summarization problem. In this work, we proposed MEMO, a novel summarization dataset for dyadic counseling, and proposed ConSum, a transformer-based domain-enriched summarization model that classifies key counseling components and integrates *PHQ-9* driven mental health knowledge. Our ablation studies demonstrated significant improvements over existing models. In another similar line of work, we proposed *PIECE*, a novel planning engine for LLMs that enhances summarization quality by orchestrating structured planning before language modeling. Extensive evaluations, including automatic, human, and expert assessments, revealed that summaries generated by *PIECE* often surpassed even human gold standards in terms of structure and relevance. This research highlights the importance of strategic content planning, opening avenues for further integration with multi-session therapy tracking and patient progress assessment.

One of the eventual goals of this research is dialogue generation, where we sought to hold more control over the LLM's generations with multiple controlling parameters. Existing VMHAs often produce generic, scripted responses that fail to align with the therapeutic goals of counseling interactions. To address this, we proposed *READER*, a response-act guided dialogue generation model that incorporates transformer-reinforcement learning with Proximal Policy Optimization (PPO), leveraging linguistic reward functions for *controlled* response generation. In another similar line of work, we proposed *EmpRes*, a commonsense-knowledge guided response generation model that utilizes speaker-context graphs and sentiment-guided commonsense relation graphs to structure responses based on conversational context. A large-scale human evaluation study revealed that 91% of users found *EmpRes*'s responses to be effective, with 80% reporting increased satisfaction with conversation quality compared to previous VMHAs. A critical aspect of any counseling interaction is the therapeutic bond, as it is the foundation of relationships. In this thesis, we proposed a novel approach to quantifying trust dynamics in counseling sessions using

psycholinguistic and psychological theories, *aka* trust modeling. We developed a robust annotated dataset containing trust levels and topic-switching patterns from real counseling dialogues, which we used to benchmark AI models for trust modeling. Our experiments demonstrated that models could effectively detect and predict clients' engagement levels, providing a quantifiable measure of trust-building in conversations. These findings offer a new direction for real-time AI-assisted therapist calibration, allowing either real therapists or bots to adapt responses based on trust levels.

Beyond one-on-one counseling, we also investigated peer interactions in Online Mental Health Communities (OMHCs), an increasingly popular avenue for individuals seeking digital peer support. Our work analyzed granular user behaviors on mental health-focused subreddits, identifying key factors that drive peer engagement. We found that post intent, readability, and critical elements in discussions significantly influenced the level of peer response and support. This research calls for further exploration into constructive vs. detrimental engagement patterns to prevent the support from going undelivered. Additionally, we emphasize the need to expand our findings to other OMHC platforms, such as Twitter, TalkLife, and 7Cups, to develop a more generalized understanding of support delivery, ultimately helping monitor and moderate peer interactions.

Deployment Pathways and Adaptation Protocols. While the primary contributions of this thesis are methodological, it is essential to outline the considerations for transitioning these research modules into deployment. The models developed here are designed with efficiency-conscious architectures, reducing inference latency and minimizing hardware demands. However, these efficiency gains alone are insufficient for deployment in sensitive domains such as mental health. Hence, the proposed thesis modules are explicitly intended as an assistive decision-support for mental health professionals, ensuring that unintended oversight is maintained at every stage. Model outputs should be treated as augmentations to professional judgment, not as replacements, with safeguards in place to detect uncertain responses. Additionally, before large-scale use, the methods must undergo domain adaptation to the target organization's specific context. This would require fine-tuning with anonymized organization-specific datasets, aligning the models with local clinical protocols, cultural norms, and communication styles. Continuous learning mechanisms should be incorporated, enabling iterative improvement based on expert feedback. Deployment itself should follow a staged approach: beginning with offline evaluations integrated into existing workflows, progressing to small-scale, monitored pilots, and only then considering broader adoption once reliability has been demonstrated in pilot settings. In parallel, compliance with local equivalent regulatory frameworks must be ensured.

Through rigorous experimentation, user studies, and expert collaboration, this thesis demonstrates that a carefully designed, context-aware, and psychologically informed approach to digital counseling can significantly enhance virtual mental health support. By breaking down the counseling process into structured computational modules, we present a framework that not only advances dialogue system research but also offers assistive methods that complement human therapists rather than replace them. Unlike existing VMHAs that function as static chatbots, this research highlights the importance of adaptive, personalized, and ethically grounded digital interventions in mental healthcare. Despite these advancements, several challenges remain. While this work significantly improves understanding, summarization, response generation, and trust evaluation, achieving initial intelligence, future research should explore multimodality incorporating audio, facial expressions, and physiological signals to enhance complete understanding. Additionally, ethical concerns surrounding bias mitigation, data privacy, and regulatory frameworks remain critical areas for exploration. On the other hand, while this thesis focuses primarily on therapy-specific scenarios, the modularity of the proposed frameworks holds clear potential for adaptation to primary care and related settings. Realizing this potential will require future research once sufficiently representative datasets from these alternative contexts become available, making this an important but

[data-dependent extension of the current work](#). In conclusion, this thesis sets a strong foundation for the next generation of VMHAs, paving the way for scalable, ethical, and contextually aware computational therapeutic models. With state-of-the-art computational dialogue modeling and real-world therapeutic insights, this research bridges the gap between automation and human-centered mental healthcare. As AI evolves, its ability to support, augment, and collaborate with mental health professionals will be crucial in addressing the global mental health crisis. This work encourages future researchers to build upon these findings, exploring novel methods to make mental health support more accessible, empathetic, and expert-supportive.

REFERENCES

- [1] G. Malhotra, A. Waheed, A. Srivastava, M. S. Akhtar, and T. Chakraborty, “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*, ser. WSDM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 735–745. [Online]. Available: <https://doi.org/10.1145/3488560.3498509>
- [2] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.
- [3] A. Pramanick, Y. Hou, S. M. Mohammad, and I. Gurevych, “The nature of nlp: Analyzing contributions in nlp papers,” 2024. [Online]. Available: <https://arxiv.org/abs/2409.19505>
- [4] M. Malgaroli, T. D. Hull, J. M. Zech, and T. Althoff, “Natural language processing for mental health interventions: A systematic review and research framework,” *Nature*, 2023. [Online]. Available: <https://www.nature.com/articles/s41398-023-02592-2>
- [5] M. Strand, L. S. Eng, and D. Gammon, “Combining online and offline peer support groups in community mental health care settings: a qualitative study of service users’ experiences,” *International journal of mental health systems*, vol. 14, no. 1, pp. 1–12, 2020.
- [6] L. S. Greenberg, “Emotion-focused therapy: A clinical synthesis,” *Focus*, vol. 8, no. 1, pp. 32–42, 2010.
- [7] R. B. Ardito and D. Rabellino, “Therapeutic alliance and outcome of psychotherapy: Historical excursus, measurements, and prospects for research,” *Frontiers in Psychology*, vol. 2, p. 270, 2011. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpsyg.2011.00270/full>
- [8] T. Alqahtani, R. Orji, S. Alhijji, A. Alhaji, A. Alhaji, A. Alhaji, A. Alhaji, A. Alhaji, A. Alhaji, and A. Alhaji, “Ai-powered mental health virtual assistants’ acceptance,” *Frontiers in Psychology*, vol. 14, p. 10753156, 2023. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpsyg.2023.10753156/full>
- [9] M. R. Bennion, G. E. Hardy, and R. K. Moore, “The use of artificial intelligence in psychotherapy,” *BMC Psychology*, vol. 11, no. 1, pp. 1–12, 2023. [Online]. Available: <https://bmcp psychology.biomedcentral.com/articles/10.1186/s40359-025-02491-9>
- [10] H. Chen, X. Liu, D. Yin, and J. Tang, “A survey on dialogue systems: Recent advances and new frontiers,” *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 2, pp. 25–35, 2017. [Online]. Available: <https://dl.acm.org/doi/10.1145/3166054.3166058>
- [11] B. J. Grosz, *The representation and use of focus in dialogue understanding*. University of California, Berkeley, 1977.
- [12] T. D. Eells, “Case conceptualization in clinical practice and training,” *Frontiers in Psychology*, vol. 14, p. 1130393, 2023. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1130393/full>

- [13] W. Zhang, X. Chen, M. Li, and J. Wang, "AI-enhanced cognitive behavioral therapy: Deep learning assessment of cognitive distortions," *arXiv preprint arXiv:2304.11449*, 2023. [Online]. Available: <https://arxiv.org/abs/2304.11449>
- [14] S. Tahir, S. A. Shah, and J. Abu-Khalaf, "Artificial empathy classification: A survey of deep learning techniques, datasets, and evaluation scales," *arXiv preprint arXiv:2310.00010*, 2023.
- [15] Z. Li, "The dark side of chatgpt: Legal and ethical challenges from stochastic parrots and hallucination," *arXiv preprint arXiv:2304.14347*, 2023. [Online]. Available: <https://arxiv.org/abs/2304.14347>
- [16] J. L. Borelli, L. Sohn, B. A. Wang, K. Hong, C. DeCoste, and N. E. Suchman, "Therapist–client language matching: Initial promise as a measure of therapist–client relationship quality." *Psychoanalytic Psychology*, vol. 36, no. 1, p. 9, 2019.
- [17] S. M. Schueller, J. F. Hunter, C. Figueroa, and A. Aguilera, "Use of digital mental health for marginalized and underserved populations," *Current Treatment Options in Psychiatry*, vol. 5, no. 3, pp. 243–255, 2018.
- [18] T. Althoff, P. Jindal, and J. Leskovec, "Online actions with offline impact: How online social networks influence online and offline user behavior," in *Proceedings of the Tenth International Conference on Web and Social Media (ICWSM)*. AAAI, 2016, pp. 1–10.
- [19] D. Vigo, G. Thornicroft, and R. Atun, "Estimating the true global burden of mental illness," *The Lancet Psychiatry*, vol. 3, no. 2, pp. 171–178, 2016. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2215036615005052>
- [20] T. Butryn, L. Bryant, C. Marchionni, and F. Sholevar, "The shortage of psychiatrists and other mental health providers: causes, current state, and potential solutions," *International Journal of Academic Medicine*, vol. 3, no. 1, pp. 5–9, 2017.
- [21] S. Rathod, N. Pinninti, M. Irfan, P. Gorczynski, P. Rathod, L. Gega, and F. Naeem, "Mental health service provision in low-and middle-income countries," *Health services insights*, vol. 10, p. 1178632917694350, 2017.
- [22] E. G. Lattie, C. Stiles-Shields, and A. K. Graham, "An overview of and recommendations for more accessible digital mental health services," *Nature Reviews Psychology*, vol. 1, no. 2, pp. 87–100, 2022.
- [23] T. J. Philippe, N. Sikder, A. Jackson, M. E. Koblanski, E. Liow, A. Pilarinos, and K. Vasarhelyi, "Digital health interventions for delivery of mental health care: systematic and comprehensive meta-review," *JMIR mental health*, vol. 9, no. 5, p. e35159, 2022.
- [24] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial," *JMIR mental health*, vol. 4, no. 2, p. e7785, 2017.
- [25] M. De Choudhury and S. De, "Mental health discourse on reddit: Self-disclosure, social support, and anonymity," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, no. 1, 2014, pp. 71–80.
- [26] J. A. Naslund, K. A. Aschbrenner, L. A. Marsch, and S. J. Bartels, "The future of mental health care: peer-to-peer support and social media," *Epidemiology and psychiatric sciences*, vol. 25, no. 2, pp. 113–122, 2016.

- [27] U. Lokala, J. Park, and M. Conway, “Computational analysis of self-disclosure in online mental health communities,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 371–382.
- [28] J. Park, M. Conway, A. T. Chen, and H. Park, “Examining the effects of self-disclosure and social support on the mental health of individuals with depression,” *International Journal of Medical Informatics*, vol. 119, pp. 127–135, 2018.
- [29] A. Barak and O. Gluck-Ofri, “The benefits of participants’ voluntary self-disclosure in online support groups,” *Computers in Human Behavior*, vol. 24, no. 5, pp. 1867–1883, 2008.
- [30] J. Huh, D. W. McDonald, A. Hartzler, and W. Pratt, “Patient moderator interaction in online health communities,” in *AMIA Annual Symposium Proceedings*, vol. 2013. American Medical Informatics Association, 2013, p. 627.
- [31] D. B. Olawade, N. Aderinto, G. Olatunji, E. Kokori, A. C. David-Olawade, and M. Hadi, “Advancements and applications of artificial intelligence in cardiology: Current trends and future prospects,” *Journal of Medicine, Surgery, and Public Health*, p. 100109, 2024.
- [32] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, “Building end-to-end dialogue systems using generative hierarchical neural network models,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/10489>
- [33] J. Torous, S. Bucci, I. H. Bell, L. V. Kessing, M. Faurholt-Jepsen, P. Whelan, A. F. Carvalho, M. Keshavan, J. Linardon, and J. Firth, “The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality,” *World Psychiatry*, vol. 20, no. 3, pp. 318–335, 2021.
- [34] A. Arrigoni, G. Rossetini, A. Palese, M. Thacker, and J. E. Esteves, “Exploring the role of therapeutic alliance and biobehavioural synchrony in musculoskeletal care: Insights from a qualitative study,” *Musculoskeletal Science and Practice*, vol. 73, p. 103164, 2024.
- [35] P. Crits-Christoph, A. Rieger, A. Gaines, and M. B. C. Gibbons, “Trust and respect in the patient-clinician relationship: preliminary development of a new scale,” *BMC psychology*, vol. 7, pp. 1–8, 2019.
- [36] J. A. Carter, “Therapeutic trust,” *Philosophical Psychology*, vol. 37, no. 1, pp. 38–61, 2024.
- [37] C. Primeau, M. Chau, M. R. Turner, and C. Paterson, “Patient experiences of patient–clinician communication among cancer multidisciplinary healthcare professionals during “breaking bad news”: A qualitative systematic review,” in *Seminars in Oncology Nursing*. Elsevier, 2024, p. 151680.
- [38] R. Gupta, D. Bone, S. Lee, and S. Narayanan, “Analysis of engagement behavior in children during dyadic interactions using prosodic cues,” *Computer speech & language*, vol. 37, pp. 47–66, 2016.
- [39] S. Shen, V. Pérez-Rosas, C. Welch, S. Poria, and R. Mihalcea, “Knowledge enhanced reflection generation for counseling dialogues,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 3096–3107.
- [40] S. Yang, R. Zhang, and S. Erfani, “Graphdialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems,” *arXiv preprint arXiv:2010.01447*, 2020.

- [41] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, “Comet: Commonsense transformers for automatic knowledge graph construction,” *arXiv preprint arXiv:1906.05317*, 2019.
- [42] Y. Shen, L. Hosseini, Y. Wang, J. Huang, and J. Gao, “Misc: A mixed initiative semantic control framework for grounded dialogue generation,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.
- [43] L. S. Nowell, J. M. Norris, D. E. White, and N. J. Moules, “Thematic analysis: Striving to meet the trustworthiness criteria,” *International journal of qualitative methods*, vol. 16, no. 1, p. 1609406917733847, 2017.
- [44] M. Kolagar, V. Parida, and D. Sjödin, “Ecosystem transformation for digital servitization: A systematic review, integrative framework, and future research agenda,” *Journal of Business Research*, vol. 146, pp. 176–200, 2022.
- [45] F. K. Qureshi, “The evolution of ai algorithms: From rule-based systems to deep learning,” *Frontiers in Artificial Intelligence Research*, vol. 1, no. 02, pp. 250–288, 2024.
- [46] H. Wang, L. Wang, Y. Du, L. Chen, J. Zhou, Y. Wang, and K.-F. Wong, “A survey of the evolution of language model-based dialogue systems,” *arXiv preprint arXiv:2311.16789*, 2023.
- [47] V. Balaraman, S. Sheikhalishahi, and B. Magnini, “Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey,” in *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue*, 2021, pp. 239–251.
- [48] J. Weizenbaum, “Eliza—a computer program for the study of natural language communication between man and machine,” in *Communications of the ACM*, vol. 9, no. 1. ACM, 1966, pp. 36–45.
- [49] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI blog*, 2019.
- [50] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, “Dialogpt: Large-scale generative pre-training for conversational response generation,” *arXiv preprint arXiv:1911.00536*, 2019.
- [51] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019.
- [52] W. H. Kwok, Y. Zhang, and G. Wang, “Artificial intelligence in perinatal mental health research: A scoping review,” *Computers in Biology and Medicine*, p. 108685, 2024.
- [53] A. Thakkar, A. Gupta, and A. De Sousa, “Artificial intelligence in positive mental health: a narrative review,” *Frontiers in digital health*, vol. 6, p. 1280235, 2024.
- [54] S. Schöbel, A. Schmitt, D. Benner, M. Saqr, A. Janson, and J. M. Leimeister, “Charting the evolution and future of conversational agents: A research agenda along five waves and new frontiers,” *Information Systems Frontiers*, vol. 26, no. 2, pp. 729–754, 2024.
- [55] Y. Zhang and N. Jaitly, “Sage: Steering and refining dialog generation with state-action augmentation,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.03040>
- [56] L. P. Kaelbling, M. L. Littman, and A. W. Moore, “Reinforcement learning: A survey,” *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

- [57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [58] R. AlMakinah, A. Norcini-Pala, L. Disney, and M. A. Canbaz, “Enhancing mental health support through human-ai collaboration: Toward secure and empathetic ai-enabled chatbots,” 2024. [Online]. Available: <https://arxiv.org/abs/2410.02783>
- [59] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, “Atomic: An atlas of machine commonsense for if-then reasoning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3027–3035. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/4191>
- [60] R. Speer, J. Chin, and C. Havasi, “Conceptnet 5.5: An open multilingual graph of general knowledge,” in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 4444–4451. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/11061>
- [61] T. Mizumoto, T. Yamazaki, K. Yoshikawa, M. Ohagi, T. Kawamoto, and T. Sato, “Dialogue systems can generate appropriate responses without the use of question marks?– a study of the effects of “?” for spoken dialogue systems –,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 4858–4864. [Online]. Available: <https://aclanthology.org/2024.lrec-main.435/>
- [62] F. Gouidis, K. Papantoniou, K. Papoutsakis, T. Patkos, A. Argyros, and D. Plexousakis, “Fusing domain-specific content from large language models into knowledge graphs for enhanced zero shot object state classification,” *Proceedings of the AAAI Symposium Series*, vol. 3, no. 1, p. 115–124, May 2024. [Online]. Available: <http://dx.doi.org/10.1609/aaais.v3i1.31190>
- [63] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, and S. Ananiadou, “Mentallama: Interpretable mental health analysis on social media with large language models,” in *Proceedings of the ACM Web Conference 2024*, ser. WWW ’24. ACM, May 2024, p. 4489–4500. [Online]. Available: <http://dx.doi.org/10.1145/3589334.3648137>
- [64] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, “Mentalbert: Publicly available pretrained language models for mental healthcare,” 2021. [Online]. Available: <https://arxiv.org/abs/2110.15621>
- [65] H. Cai, X. Shen, Q. Xu, W. Shen, X. Wang, W. Ge, X. Zheng, and X. Xue, “Improving empathetic dialogue generation by dynamically infusing commonsense knowledge,” *arXiv preprint arXiv:2306.04657*, 2023.
- [66] H. Chae, S. Moon, T. Wang, M. Bansal, C. Cardie, and R. Subba, “Dialogue chain-of-thought distillation for commonsense-aware conversational agents,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 4956–4974.
- [67] B. Y. Lin, X. Chen, J. Chen, and X. Ren, “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)*. Association for Computational Linguistics, 2019, pp. 2829–2839. [Online]. Available: <https://aclanthology.org/D19-1283/>

- [68] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu, “Ernie: Enhanced language representation with informative entities,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 1441–1451. [Online]. Available: <https://aclanthology.org/2020.acl-main.130/>
- [69] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, “K-bert: Enabling language representation with knowledge graph,” in *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020, pp. 2901–2908. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/5681>
- [70] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, “Grounded dialog generation with pre-trained language models,” 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)*, 2019, pp. 3762–3774. [Online]. Available: <https://aclanthology.org/D19-1381.pdf>
- [71] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, “Wizard of wikipedia: Knowledge-powered conversational agents,” *arXiv preprint arXiv:1811.01241*, 2019. [Online]. Available: <https://arxiv.org/pdf/1811.01241.pdf>
- [72] K. M. Laska, A. S. Gurman, and B. E. Wampold, “Common factors in psychotherapy: What we know and how we can use this knowledge,” *Cognitive and Behavioral Practice*, vol. 21, no. 3, pp. 293–305, 2014.
- [73] A. N. Niles, M. G. Craske, M. D. Lieberman, C.-H. Hur, and P. Vrticka, “Affect labeling enhances exposure effectiveness for public speaking anxiety,” *Behaviour Research and Therapy*, vol. 68, pp. 27–36, 2015.
- [74] Z. A. Nazi and W. Peng, “Large language models in healthcare and medical domain: A review,” 2024. [Online]. Available: <https://arxiv.org/abs/2401.06775>
- [75] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019. [Online]. Available: <https://arxiv.org/abs/1810.04805>
- [76] S. C. Moreno, C. Arango, C. Moreno, T. Wykes, S. Galderisi, M. Nordentoft, N. Crossley, N. Jones, M. Cannon, C. U. Correll, L. Byrne, S. Carr, E. Y. H. Chen, P. Gorwood, S. Johnson, H. Kärkkäinen, J. H. Krystal, J. Lee, J. Lieberman, C. López-Jaramillo, M. Männikkö, M. R. Phillips, H. Uchida, E. Vieta, A. Vita, and C. Arango, “Position Paper How mental health care should change as a consequence of the COVID-19 pandemic,” *The Lancet Psychiatry*, vol. 7, pp. 813–824, 2020.
- [77] X. Wang, S. Hegde, C. Son, B. Keller, A. Smith, and F. Sasangohar, “Investigating mental health of US college students during the COVID-19 pandemic: Cross-sectional survey study,” *JMIR*, vol. 22, no. 9, p. e22817, Sep.
- [78] A. Ahmadvand, J. I. Choi, and E. Agichtein, “Contextual dialogue act classification for open-domain conversational agents,” in *SIGIR*, 2019, pp. 1273–1276.
- [79] V. Raheja and J. Tetreault, “Dialogue Act Classification with Context-Aware Self-Attention,” in *NAACL*, Jun. 2019, pp. 3727–3733.
- [80] J. J. Godfrey, E. C. Holliman, and J. McDaniel, “Switchboard: telephone speech corpus for research and development,” in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1992, pp. 517–520 vol.1.

- [81] G. Shang, A. Tixier, M. Vazirgiannis, and J.-P. Lorré, “Speaker-change aware CRF for dialogue act classification,” in *COLING*, 2020, pp. 450–464.
- [82] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, “Dailydialog: A manually labelled multi-turn dialogue dataset,” in *IJCNLP*, 2017.
- [83] S. Kumar, A. Shrimal, M. S. Akhtar, and T. Chakraborty, “Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer,” 2021.
- [84] C. Cerisara, S. Jafaritazehjani, A. Oluokun, and H. T. Le, “Multi-task dialog act and sentiment recognition on mastodon,” in *COLING*, 2018, pp. 745–754.
- [85] M. Bedi, S. Kumar, M. S. Akhtar, and T. Chakraborty, “Multi-modal sarcasm detection and humor classification in code-mixed conversations,” *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.
- [86] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder–decoder for statistical machine translation,” in *EMNLP*, Oct. 2014, pp. 1724–1734.
- [87] T. A. Van Dijk, *Discourse studies: A multidisciplinary introduction*. Sage, 2011.
- [88] R. Mishra, P. P. Sinha, R. Sawhney, D. Mahata, P. Mathur, and R. R. Shah, “Snap-batnet: Cascading author profiling and social network graphs for suicide ideation detection on social media,” in *NAACL: Student Research Workshop*, 2019, pp. 147–156.
- [89] T. Naous, C. Hokayem, and H. Hajj, “Empathy-driven arabic conversational chatbot,” in *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, 2020, pp. 58–68.
- [90] M. Levis, C. Leonard Westgate, J. Gui, B. V. Watts, and B. Shiner, “Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models,” *Psychological Medicine*, p. 1–10, 2020.
- [91] A. Fine, P. Crutchley, J. Blase, J. Carroll, and G. Coppersmith, “Assessing population-level symptoms of anxiety, depression, and suicide risk in real time using nlp applied to social media data,” in *Natural Language Processing and Computational Social Science*, 2020, pp. 50–54.
- [92] J. Pennebaker, M. Mehl, and K. Niederhoffer, “Psychological aspects of natural language. use: our words, our selves.” *Annual review of psychology*, vol. 54, pp. 547–77, 2003.
- [93] W. Wei, J. Liu, X. Mao, G. Guo, F. Zhu, P. Zhou, and Y. Hu, “Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction,” in *CIKM*, 2019, p. 1401–1410.
- [94] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, “The distress analysis interview corpus of human and computer interviews.” in *LREC*, 2014, pp. 3123–3128.
- [95] F. Morbini, D. DeVault, K. Georgila, R. Artstein, D. Traum, and L.-P. Morency, “A Demonstration of Dialogue Processing in SimSensei Kiosk,” in *SIGDIAL*, Jun. 2014, p. 254.
- [96] K. Kretzschmar, H. Tyroll, G. Pavarini, A. Manzini, and I. Singh, “Can Your Phone Be Your Therapist? Young People’s Ethical Perspectives on the Use of Fully Automated Conversational Agents (Chatbots) in Mental Health Support,” *Biomed Inform Insights*, vol. 11, p. 1178222619829083, 2019.

- [97] J. Weizenbaum, “ELIZA – A Computer Program For the Study of Natural Language Communication Between Man and Machine,” *CACM*, vol. 9, pp. 36–45, January 1966.
- [98] P. Colombo, E. Chapuis, M. Manica, E. Vignon, G. Varni, and C. Clavel, “Guiding attention in sequence-to-sequence models for dialogue act prediction,” in *AAAI*, 2020.
- [99] N. Reithinger and M. Klesen, “Dialogue act classification using language models,” in *Fifth European Conference on Speech Communication and Technology*, 1997.
- [100] S. Grau, E. Sanchis, M. J. Castro, and D. Vilar, “Dialogue act classification using a bayesian approach,” in *9th Conference Speech and Computer*, 2004.
- [101] D. Ortega, C.-Y. Li, G. Vallejo, P. Denisov, and N. T. Vu, “Context-aware neural-based dialog act classification on automatically generated transcriptions,” in *ICASSP*, 2019, pp. 7265–7269.
- [102] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural Comput.*, p. 541–551, Dec. 1989.
- [103] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *ICML*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, p. 282–289.
- [104] J. Y. Lee and F. Démoncourt, “Sequential short-text classification with recurrent and convolutional neural networks,” in *NAACL*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 515–520.
- [105] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Internal Representations by Error Propagation*. Cambridge, MA, USA: MIT Press, 1986, p. 318–362.
- [106] V. Raheja and J. Tetreault, “Dialogue act classification with context-aware self-attention,” in *NAACL*, 2019, pp. 3727–3733.
- [107] K. Hua, Z. Feng, C. Tao, R. Yan, and L. Zhang, “Learning to detect relevant contexts and knowledge for response selection in retrieval-based dialogue systems,” in *CIKM*, 2020, p. 525–534.
- [108] Z. Chen, R. Yang, Z. Zhao, D. Cai, and X. He, “Dialogue act recognition via crf-attentive structured network,” in *SIGIR*, 2018, pp. 225–234.
- [109] Y. Yu, S. Peng, and G. H. Yang, “Modeling long-range context for concurrent dialogue acts recognition,” in *CIKM*, 2019, p. 2277–2280.
- [110] L. Qin, W. Che, Y. Li, M. Ni, and T. Liu, “Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification,” in *AAAI*, 2020, pp. 8665–8672.
- [111] T. Saha, D. Gupta, S. Saha, and P. Bhattacharyya, “Emotion aided dialogue act classification for task-independent conversations in a multi-modal framework,” *Cognitive Computation*, pp. 1–13, 2020.
- [112] G. Shang, A. J.-P. Tixier, M. Vazirgiannis, and J. Lorré, “Speaker-change aware crf for dialogue act classification,” in *COLING*, 2020.
- [113] L. A. J. *How to Do Things with Words: The William James Lectures Delivered at Harvard University in 1955*, M. Sbisà and O. U. J. Eds. Oxford, England: Oxford University Press, 1962. [Online]. Available: <https://philpapers.org/rec/LAUHTD>

- [114] J. R. Searle, *Speech Acts: An Essay in the Philosophy of Language*. Cambridge, England: Cambridge University Press, 1969. [Online]. Available: <https://www.bibsonomy.org/bibtex/2e627c70f7a8b320bfba123f85adf0650/porta>
- [115] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [116] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized BERT pretraining approach,” *CoRR*, vol. abs/1907.11692, 2019.
- [117] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, “Patient subtyping via time-aware lstm networks,” in *SIGKDD*, 2017, p. 65–74.
- [118] Y. Zhang, P. Tiwari, D. Song, X. Mao, P. Wang, X. Li, and H. M. Pandey, “Learning interaction dynamics with an interactive lstm for conversational sentiment analysis,” *Neural Networks*, vol. 133, pp. 40 – 56, 2021.
- [119] B. Wang, “Disconnected recurrent neural networks for text categorization,” in *ACL*, 2018, pp. 2311–2320.
- [120] Z. Kozareva and S. Ravi, “ProSeqo: Projection sequence networks for on-device text classification,” in *EMNLP-IJCNLP*, 2019, pp. 3894–3903.
- [121] S. Ravi, “Projectionnet: Learning efficient on-device deep networks using neural projections,” *CoRR*, vol. abs/1708.00630, 2017. [Online]. Available: <http://arxiv.org/abs/1708.00630>
- [122] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, “Very deep convolutional networks for text classification,” in *EACL*, Apr. 2017, pp. 1107–1116.
- [123] P. Liu, X. Qiu, and X. Huang, “Recurrent neural network for text classification with multi-task learning,” in *IJCAI*. AAAI Press, 2016, p. 2873–2879.
- [124] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the NAACL*. Minnesota: ACL, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423>
- [125] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *NeurIPS*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds., 2019, pp. 8024–8035.
- [126] A. S. Miner, A. Haque, J. A. Fries, S. Fleming, D. Wilfley, G. T. Wilson, A. Milstein, D. Jurafsky, B. Arnow, W. S. Agras, L. Fei-Fei, and N. Shah, “Assessing the accuracy of automatic speech recognition for psychotherapy,” *NPJ Digital Medicine*, vol. 3, 2020.
- [127] A. See, P. J. Liu, and C. D. Manning, “Get to the point: Summarization with pointer-generator networks,” in *Proceedings of the 55th ACL*, 2017.
- [128] R. Nallapati, B. Zhou, C. dos Santos, Gulçehre, and B. Xiang, “Abstractive text summarization using sequence-to-sequence RNNs and beyond,” in *The 20th CoNLL*, 2016.

- [129] M. Afzal, F. Alam, K. M. Malik, and G. M. Malik, “Clinical context-aware biomedical text summarization using deep neural network: Model development and validation,” vol. 22, no. 10, 2020.
- [130] Y. Song, Y. Tian, N. Wang, and F. Xia, “Summarizing medical conversations via identifying important utterances,” in *Proceedings of the 28th COLING*, 2020.
- [131] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, “The PHQ-9,” *Journal of General Internal Medicine*, vol. 16, no. 9, Sep. 2001.
- [132] P. Chen and R. Verma, “A Query-Based Medical Information Summarization System Using Ontology Knowledge,” in *19th IEEE Symposium on Computer-Based Medical Systems (CBMS’06)*, Jun. 2006, iSSN: 1063-7125.
- [133] S. Konovalov, M. Scotch, L. Post, and C. Brandt, “Biomedical informatics techniques for processing and analyzing web blogs of military service members,” *Journal of Medical Internet Research*, vol. 12, no. 4, Oct. 2010.
- [134] J. Strauss, A. M. Peguero, and G. Hirst, “Machine learning methods for clinical forms analysis in mental health,” *Studies in Health Technology and Informatics*, vol. 192, p. 1024, 2013.
- [135] S. H. Kennedy, R. W. Lam, R. S. McIntyre, S. V. Tourjman, V. Bhat, P. Blier, M. Hasnain, F. Jollant, A. J. Levitt, G. M. MacQueen, S. J. McInerney, D. McIntosh, R. V. Milev, D. J. Müller, S. V. Parikh, N. L. Pearson, A. V. Ravindran, and R. Uher, “Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder,” *Canadian Journal of Psychiatry*, vol. 61, no. 9, Sep. 2016.
- [136] T. Tran and R. Kavuluru, “Predicting mental conditions based on “history of present illness” in psychiatric notes with deep neural networks,” vol. 75, 2017.
- [137] Y.-P. Chen, Y.-Y. Chen, J.-J. Lin, C.-H. Huang, and F. Lai, “Modified Bidirectional Encoder Representations From Transformers Extractive Summarization Model for Hospital Information Systems Based on Character-Level Tokens (AlphaBERT): Development and Performance Evaluation,” *JMIR Medical Informatics*, vol. 8, no. 4, p. e17787, Apr. 2020.
- [138] J. Ive, N. Viani, J. Kam, L. Yin, S. Verma, S. Puntis, R. N. Cardinal, A. Roberts, R. Stewart, and S. Velupillai, “Generation and evaluation of artificial mental health records for natural language processing,” vol. 3, 2020.
- [139] M. Gaur, V. Aribandi, U. Kursuncu, A. Alambo, V. L. Shalin, K. Thirunarayan, J. Beich, M. Narasimhan, and A. Sheth, “Knowledge-infused abstractive summarization of clinical diagnostic interviews: Framework development study,” *JMIR*, vol. 8, no. 5, May 2021.
- [140] L. Zhang, R. Negrinho, A. Ghosh, V. Jagannathan, H. R. Hassanzadeh, T. Schaaf, and M. R. Gormley, “Leveraging pretrained models for automatic summarization of doctor-patient conversations,” in *Findings of the ACL: EMNLP 2021*. ACL, Nov. 2021.
- [141] H. Zafari and F. Zulkernine, “Chatsum: An intelligent medical chat summarization tool,” in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2021.
- [142] Y. K. Atri, S. Pramanick, V. Goyal, and T. Chakraborty, “See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization,” *Knowledge-Based Systems*, vol. 227, p. 107152, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0950705121004159>

- [143] A. Dey, T. Chowdhury, Y. Kumar, and T. Chakraborty, “Corpora evaluation and system bias detection in multi-document summarization,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 2830–2840. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.254>
- [144] Y.-N. Chen, Y. Huang, C.-f. Yeh, and L.-S. Lee, “Spoken lecture summarization by random walk over a graph constructed with automatically extracted key terms.” 2011, pp. 933–936.
- [145] H. Saggion and T. Poibeau, “Automatic text summarization: Past, present and future,” in *Multi-source, Multilingual Information Extraction and Summarization*, ser. Theory and Applications of Natural Language Processing, R. Y. T. Poibeau; H. Saggion. J. Piskorski, Ed. Springer, 2012.
- [146] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” *arXiv:1706.03762 [cs]*, Dec. 2017, arXiv: 1706.03762.
- [147] Y.-C. Chen and M. Bansal, “Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting,” *arXiv:1805.11080 [cs]*, May 2018.
- [148] S. Narayan, S. B. Cohen, and M. Lapata, “Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. ACL, 2018.
- [149] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” vol. 3, 2003.
- [150] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” *arXiv:1912.08777*, Jul. 2020.
- [151] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *JMLR*, jan 2020.
- [152] J. C. Quiroz, L. Laranjo, A. B. Kocaballi, A. Briatore, S. Berkovsky, D. Rezazadegan, and E. Coiera, “Identifying relevant information in medical conversations to summarize a clinician-patient encounter,” *Health Informatics Journal*, vol. 26, no. 4, 2020, PMID: 32865113.
- [153] K. Krishna, S. Khosla, J. Bigham, and Z. C. Lipton, “Generating SOAP notes from doctor-patient conversations using modular summarization techniques,” in *Proceedings of the 59th ACL and the 11th IJCNLP*, Aug. 2021.
- [154] M. Sekhon, M. Cartwright, and J. J. Francis, “Acceptability of healthcare interventions: an overview of reviews and development of a theoretical framework,” *BMC Health Services Research*, vol. 17, no. 1, Jan 2017.
- [155] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” *arXiv:1910.01108 [cs]*, Feb. 2020, arXiv: 1910.01108.
- [156] A. H. Yazdavar, H. S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, and A. Sheth, “Semi-supervised approach to monitoring clinical depressive symptoms in social media,” *IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, vol. 2017, 2017.
- [157] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with bert,” 2020.
- [158] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” 2004.

- [159] T. Scialom, P.-A. Dray, P. Gallinari, S. Lamprier, B. Piwowarski, J. Staiano, and A. Wang, “QuestEval: Summarization Asks for Fact-based Evaluation,” *arXiv:2103.12693 [cs]*, Apr. 2021.
- [160] T. Sellam, D. Das, and A. Parikh, “BLEURT: Learning Robust Metrics for Text Generation,” in *Proceedings of the 58th ACL*, Online, 2020.
- [161] X. Feng, X. Feng, L. Qin, B. Qin, and T. Liu, “Language model as an annotator: Exploring DialoGPT for dialogue summarization,” in *Proceedings of the 59th ACL and the 11th IJCNLP*, Online, Aug. 2021.
- [162] Y. Zou, J. Lin, L. Zhao, Y. Kang, Z. Jiang, C. Sun, Q. Zhang, X. Huang, and X. Liu, “Unsupervised summarization for chat logs with topic-oriented ranking and context-aware auto-encoders,” *Proceedings of the AAAI Conference*, vol. 35, no. 16, May 2021.
- [163] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” Jan. 2008.
- [164] C.-W. Goo and Y.-N. Chen, “Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts,” *SLT*, 2018.
- [165] A. Joshi, N. Katariya, X. Amatriain, and A. Kannan, “Dr. summarize: Global summarization of medical dialogue by exploiting local structures.” in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020.
- [166] S. Molenaar, L. Maas, V. Burriel, F. Dalpiaz, and S. Brinkkemper, *Medical Dialogue Summarization for Automated Reporting in Healthcare*, 05 2020.
- [167] A. Thirunavukarasu, D. Ting, and K. Elangovan, “Large language models in medicine,” *Nature Medicine*, 2023.
- [168] A. C. Heerden, J. R. Pozuelo, and B. A. Kohrt, “Global Mental Health Services and the Impact of Artificial Intelligence–Powered Large Language Models,” *JAMA Psychiatry*, vol. 80, no. 7, pp. 662–664, 07 2023. [Online]. Available: <https://doi.org/10.1001/jamapsychiatry.2023.1253>
- [169] M. Cascella, J. Montomoli, and V. e. a. Bellini, “Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios,” *J Med Syst*, 2023. [Online]. Available: <https://link.springer.com/article/10.1007/s10916-023-01925-4>
- [170] A. Srivastava, T. Suresh, S. P. Lord, M. S. Akhtar, and T. Chakraborty, “Counseling summarization using mental health knowledge guided utterance filtering,” in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3920–3930. [Online]. Available: <https://doi.org/10.1145/3534678.3539187>
- [171] Z. Feng, W. Ma, W. Yu, L. Huang, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. liu, “Trends in integration of knowledge and large language models: A survey and taxonomy of methods, benchmarks, and applications,” 2023.
- [172] W. Yu, C. Zhu, Z. Li, Z. Hu, Q. Wang, H. Ji, and M. Jiang, “A survey of knowledge-enhanced text generation,” *ACM Computing Survey (CSUR)*, 2022.
- [173] W. Yim, Y. Fu, A. Ben Abacha, N. Snider, T. Lin, and M. Yetisgen, “Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation,” *Nature Scientific Data*, 2023.

- [174] A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, and T. Althoff, “Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support,” *Nature Machine Intelligence*, 2023.
- [175] —, “Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach,” in *Proceedings of the Web Conference 2021*, ser. WWW ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 194–205. [Online]. Available: <https://doi.org/10.1145/3442381.3450097>
- [176] T. Saha, V. Gakhreja, A. S. Das, S. Chakraborty, and S. Saha, “Towards motivational and empathetic response generation in online mental health support,” in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’22. New York, NY, USA: ACM, 2022, p. 2650–2656. [Online]. Available: <https://doi.org/10.1145/3477495.3531912>
- [177] T. Saha, S. Reddy, A. Das, S. Saha, and P. Bhattacharyya, “A shoulder to cry on: Towards a motivational virtual assistant for assuaging mental agony,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 2436–2449. [Online]. Available: <https://aclanthology.org/2022.naacl-main.174>
- [178] P. Priya, K. Mishra, P. Totala, and A. Ekbal, “Partner: A persuasive mental health and legal counselling dialogue system for women and children crime victims,” in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, E. Elkind, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2023, pp. 6183–6191, aI for Good. [Online]. Available: <https://doi.org/10.24963/ijcai.2023/686>
- [179] A. Srivastava, I. Pandey, M. S. Akhtar, and T. Chakraborty, “Response-act guided reinforced dialogue generation for mental health counseling,” in *Proceedings of the ACM Web Conference 2023*, ser. WWW ’23. New York, NY, USA: Association for Computing Machinery, 2023, p. 1118–1129. [Online]. Available: <https://doi.org/10.1145/3543507.3583380>
- [180] K. Mishra, P. Priya, M. Burja, and A. Ekbal, “e-THERAPIST: I suggest you to cultivate a mindset of positivity and nurture uplifting thoughts,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13 952–13 967. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.861>
- [181] K. Mishra, P. Priya, and A. Ekbal, “Help me heal: A reinforced polite and empathetic mental health and legal counseling dialogue system for crime victims,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, pp. 14 408–14 416, Jun. 2023. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/26685>
- [182] —, “PAL to lend a helping hand: Towards building an emotion adaptive polite and empathetic counseling conversational agent,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 12 254–12 271. [Online]. Available: <https://aclanthology.org/2023.acl-long.685>
- [183] S. Liu, N. Deng, S. Sabour, Y. Jia, M. Huang, and R. Mihalcea, “Task-adaptive tokenization: Enhancing long-form text generation efficacy in mental health and beyond,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor,

- J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 15 264–15 281. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.944>
- [184] K. Yang, S. Ji, T. Zhang, Q. Xie, Z. Kuang, and S. Ananiadou, “Towards interpretable mental health analysis with large language models,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6056–6077. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.370>
- [185] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, “How can we know what language models know?” *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 423–438, 2020. [Online]. Available: <https://aclanthology.org/2020.tacl-1.28>
- [186] Z. Jiang, J. Araki, H. Ding, and G. Neubig, “How can we know when language models know? on the calibration of language models for question answering,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 962–977, 2021. [Online]. Available: <https://aclanthology.org/2021.tacl-1.57>
- [187] M. Choi, J. Pei, S. Kumar, C. Shu, and D. Jurgens, “Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 11 370–11 403. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.699>
- [188] K. Singhal, S. Azizi, and T. Tu, “Large language models encode clinical knowledge,” *Nature*, 2023.
- [189] Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, “Llama 2: Open foundation and fine-tuned chat models,” 2023.
- [190] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” 2019.
- [191] K. Yang, T. Zhang, Z. Kuang, Q. Xie, S. Ananiadou, and J. Huang, “Mentallama: Interpretable mental health analysis on social media with large language models,” 2023.
- [192] K. Valmeekam, M. Marquez, S. Sreedharan, and S. Kambhampati, “On the planning abilities of large language models - a critical investigation,” in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: <https://openreview.net/forum?id=X6dEqXIsEW>
- [193] X. Wang, L. Caccia, O. Ostapenko, X. Yuan, W. Y. Wang, and A. Sordoni, “Guiding language model math reasoning with planning tokens,” 2024.
- [194] H. Zhang, W. Xu, and H. Yu, “Generative planning for temporally coordinated exploration in reinforcement learning,” in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=YZHES8wIdE>

- [195] S. Zhang, Z. Chen, Y. Shen, M. Ding, J. B. Tenenbaum, and C. Gan, “Planning with large language models for code generation,” in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: <https://openreview.net/forum?id=Lr8cOOtYbfl>
- [196] A. Sharma, K. Rudra, and N. Ganguly, “Going beyond content richness: Verified information aware summarization of crisis-related microblogs,” in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM ’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 921–930. [Online]. Available: <https://doi.org/10.1145/3357384.3358020>
- [197] R. Jain, T. Saha, J. Lalwani, and S. Saha, “Can you summarize my learnings? towards perspective-based educational dialogue summarization,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 3158–3173. [Online]. Available: <https://aclanthology.org/2023.findings-emnlp.208>
- [198] S. Kumar, G. S. Kohli, T. Ghosal, and A. Ekbal, “Mup-scidocsum: Leveraging multi-perspective peer review summaries for scientific document summarization,” in *Leveraging Generative Intelligence in Digital Libraries: Towards Human-Machine Collaboration*, D. H. Goh, S.-J. Chen, and S. Tuarob, Eds. Singapore: Springer Nature Singapore, 2023, pp. 250–267.
- [199] Y. K. Atri, V. Goyal, and T. Chakraborty, “Fusing multimodal signals on hyper-complex space for extreme abstractive text summarization (tl;dr) of scientific contents,” in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD ’23. New York, NY, USA: Association for Computing Machinery, 2023, p. 3724–3736. [Online]. Available: <https://doi.org/10.1145/3580305.3599830>
- [200] A. Tiwari, A. Saha, S. Saha, P. Bhattacharyya, and M. Dhar, “Experience and evidence are the eyes of an excellent summarizer! towards knowledge infused multi-modal clinical conversation summarization,” in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, ser. CIKM ’23. New York, NY, USA: Association for Computing Machinery, 2023, p. 2452–2461. [Online]. Available: <https://doi.org/10.1145/3583780.3614870>
- [201] S. Mukherjee, A. Jatowt, R. Kumar, A. Jangra, and S. Saha, “Can multimodal pointer generator transformers produce topically relevant summaries?” in *2023 International Joint Conference on Neural Networks (IJCNN)*, 2023, pp. 1–8.
- [202] Y. K. Atri, V. Goyal, and T. Chakraborty, “Multi-document summarization using selective attention span and reinforcement learning,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3457–3467, 2023.
- [203] C. Bodnar, F. D. Giovanni, B. P. Chamberlain, P. Lio, and M. M. Bronstein, “Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in GNNs,” in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: <https://openreview.net/forum?id=vbPsD-BhOZ>
- [204] Y. Atri, A. Iyer, T. Chakraborty, and V. Goyal, “Promoting topic coherence and inter-document consorts in multi-document summarization via simplicial complex and sheaf graph,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2154–2166. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.133>

- [205] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, “The PHQ-9,” *Journal of General Internal Medicine*, vol. 16, no. 9, Sep. 2001.
- [206] A. Cohan, W. Ammar, M. van Zuylen, and F. Cady, “Structural scaffolds for citation intent classification in scientific publications,” 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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3586–3596. [Online]. Available: <https://aclanthology.org/N19-1361>
- [207] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: <https://aclanthology.org/D14-1162>
- [208] A. Shepard, “A cellular description of the derived category of a stratified space,” in *Brown University*, 1985. [Online]. Available: <http://justinmcurry.com/wp-content/uploads/2022/02/shepard.pdf>
- [209] J. Curry, “Sheaves, cosheaves and applications,” 2014.
- [210] Y. Zhang, A. Ni, Z. Mao, C. H. Wu, C. Zhu, B. Deb, A. Awadallah, D. Radev, and R. Zhang, “Summⁿ: A multi-stage summarization framework for long input dialogues and documents,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1592–1604. [Online]. Available: <https://aclanthology.org/2022.acl-long.112>
- [211] M. Zhong, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, “Dialoglm: Pre-trained model for long dialogue understanding and summarization,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 11 765–11 773, Jun. 2022. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/21432>
- [212] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, “Scaling instruction-finetuned language models,” 2022.
- [213] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, “Mistral 7b,” 2023. [Online]. Available: <https://arxiv.org/abs/2310.06825>
- [214] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourier, N. Habib, N. Sarrazin, O. Sansevero, A. M. Rush, and T. Wolf, “Zephyr: Direct distillation of lm alignment,” 2023.
- [215] A. A. Abd-Alrazaq, M. Alajlani, N. Ali, K. Denecke, B. M. Bewick, and M. Househ, “Perceptions and opinions of patients about mental health chatbots: Scoping review,” *J Med Internet Res*, vol. 23, no. 1, p. e17828, Jan 2021. [Online]. Available: <http://www.jmir.org/2021/1/e17828/>
- [216] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, “The design and implementation of xiaoice, an empathetic social chatbot,” *Computational Linguistics*, vol. 46, pp. 1–62, 01 2020.

- [217] M. Zhang and J. Li, “A commentary of gpt-3 in mit technology review 2021,” *Fundamental Research*, vol. 1, no. 6, pp. 831–833, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2667325821002193>
- [218] B. Noble and V. Maraev, “Large-scale text pre-training helps with dialogue act recognition, but not without fine-tuning,” in *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*. Groningen, The Netherlands (online): Association for Computational Linguistics, Jun. 2021, pp. 166–172. [Online]. Available: <https://aclanthology.org/2021.iwcs-1.16>
- [219] C. H. Chao, X. J. Hou, and Y. C. Chiu, “Improve chit-chat and QA sentence classification in user messages of dialogue system using dialogue act embedding,” in *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*. Taoyuan, Taiwan: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), Oct. 2021, pp. 138–143. [Online]. Available: <https://aclanthology.org/2021.rocling-1.19>
- [220] E. Hosseini-Asl, B. McCann, C.-S. Wu, S. Yavuz, and R. Socher, “A simple language model for task-oriented dialogue,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 20 179–20 191. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/e946209592563be0f01c844ab2170f0c-Paper.pdf>
- [221] P. Gupta, H. Jhamtani, and J. Bigham, “Target-guided dialogue response generation using commonsense and data augmentation,” in *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1301–1317. [Online]. Available: <https://aclanthology.org/2022.findings-naacl.97>
- [222] J. Tang, T. Zhao, C. Xiong, X. Liang, E. Xing, and Z. Hu, “Target-guided open-domain conversation,” in *Proceedings of the 57th Annual Meeting of the ACL*. Florence, Italy: ACL, Jul. 2019, pp. 5624–5634.
- [223] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, “Dialogue act modeling for automatic tagging and recognition of conversational speech,” *Computational Linguistics*, vol. 26, no. 3, pp. 339–374, 2000. [Online]. Available: <https://aclanthology.org/J00-3003>
- [224] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić, “MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 5016–5026. [Online]. Available: <https://aclanthology.org/D18-1547>
- [225] P.-H. Su, D. Vandyke, M. Gašić, D. Kim, N. Mrkšić, T. H. Wen, and S. Young, “Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems,” 09 2015.
- [226] X. Chen, J. Xu, and B. Xu, “A working memory model for task-oriented dialog response generation,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2687–2693. [Online]. Available: <https://aclanthology.org/P19-1258>
- [227] J. Pei, P. Ren, C. Monz, and M. de Rijke, “Retrospective and prospective mixture-of-generators for task-oriented dialogue response generation,” *CoRR*, vol. abs/1911.08151, 2019. [Online]. Available: <http://arxiv.org/abs/1911.08151>

- [228] S. Mehri, T. Srinivasan, and M. Eskenazi, “Structured fusion networks for dialog,” 01 2019, pp. 165–177.
- [229] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. F. Christiano, and G. Irving, “Fine-tuning language models from human preferences,” *CoRR*, vol. abs/1909.08593, 2019. [Online]. Available: <http://arxiv.org/abs/1909.08593>
- [230] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. [Online]. Available: <https://aclanthology.org/N18-1202>
- [231] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. Red Hook, NY, USA: Curran Associates Inc., 2019.
- [232] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, “Transfertransfo: A transfer learning approach for neural network based conversational agents,” *ArXiv*, vol. abs/1901.08149, 2019.
- [233] C. Xu, W. Wu, C. Tao, H. Hu, M. Schuerman, and Y. Wang, “Neural response generation with meta-words,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5416–5426. [Online]. Available: <https://aclanthology.org/P19-1538>
- [234] M. Khalifa, H. Elsahar, and M. Dymetman, “A distributional approach to controlled text generation,” 2021. [Online]. Available: <https://openreview.net/forum?id=jWkw45-9AbL>
- [235] P. Lippe, P. Ren, H. Haned, B. Voorn, and M. de Rijke, “Diversifying task-oriented dialogue response generation with prototype guided paraphrasing,” *CoRR*, vol. abs/2008.03391, 2020. [Online]. Available: <https://arxiv.org/abs/2008.03391>
- [236] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” 2017. [Online]. Available: <https://arxiv.org/abs/1707.06347>
- [237] S. Dingliwal, A. Shenoy, S. Bodapati, A. Gandhe, R. T. Gadde, and K. Kirchhoff, “Prompt tuning gpt-2 language model for parameter-efficient domain adaptation of asr systems,” 2021. [Online]. Available: <https://arxiv.org/abs/2112.08718>
- [238] S. Gretz, Y. Bilu, E. Cohen-Karlik, and N. Slonim, “The workweek is the best time to start a family – a study of gpt-2 based claim generation,” 2020. [Online]. Available: <https://arxiv.org/abs/2010.06185>
- [239] W. Lin, B.-H. Tseng, and B. Byrne, “Knowledge-aware graph-enhanced GPT-2 for dialogue state tracking,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7871–7881. [Online]. Available: <https://aclanthology.org/2021.emnlp-main.620>
- [240] Y. Yang, Y. Li, and X. Quan, “Ubar: Towards fully end-to-end task-oriented dialog systems with gpt-2,” 2020. [Online]. Available: <https://arxiv.org/abs/2012.03539>

- [241] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, “DIALOGPT : Large-scale generative pre-training for conversational response generation,” in *Proceedings of the 58th Annual Meeting of the ACL: System Demonstrations*. Online: ACL, Jul. 2020, pp. 270–278.
- [242] W. Chen, Y. Gong, S. Wang, B. Yao, W. Qi, Z. Wei, X. Hu, B. Zhou, Y. Mao, W. Chen, B. Cheng, and N. Duan, “DialogVED: A pre-trained latent variable encoder-decoder model for dialog response generation,” in *Proceedings of the 60th Annual Meeting of the ACL*. Dublin, Ireland: ACL, May 2022, pp. 4852–4864.
- [243] W. Qi, Y. Gong, Y. Yan, C. Xu, B. Yao, B. Zhou, B. Cheng, D. Jiang, J. Chen, R. Zhang, H. Li, and N. Duan, “ProphetNet-X: Large-scale pre-training models for English, Chinese, multi-lingual, dialog, and code generation,” in *Proceedings of the 59th Annual Meeting of the ACL and the 11th IJCNLP: System Demonstrations*. Online: ACL, Aug. 2021, pp. 232–239.
- [244] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, “Building end-to-end dialogue systems using generative hierarchical neural network models,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI’16. AAAI Press, 2016, p. 3776–3783.
- [245] T. Zhao and T. Kawahara, “Effective incorporation of speaker information in utterance encoding in dialog,” 2019. [Online]. Available: <https://arxiv.org/abs/1907.05599>
- [246] Y. Park, J. Cho, and G. Kim, “A hierarchical latent structure for variational conversation modeling,” in *Proceedings of the 2018 Conference of the NAACL*, Louisiana, Jun. 2018, pp. 1792–1801. [Online]. Available: <https://aclanthology.org/N18-1162>
- [247] C. van der Lee, A. Gatt, E. van Miltenburg, and E. Kraemer, “Human evaluation of automatically generated text: Current trends and best practice guidelines,” *Computer Speech & Language*, vol. 67, p. 101151, 2021.
- [248] T. DeAngelis, “Better relationships with patients lead to better outcomes.” in *Monitor on Psychology*. Monitor on Psychology, 2019. [Online]. Available: <https://www.apa.org/monitor/2019/11/ce-corner-relationships>
- [249] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, and M. Montes-y Gómez, “Detecting mental disorders in social media through emotional patterns - the case of anorexia and depression,” *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 211–222, 2023.
- [250] A. Srivastava, T. Gupta, A. Cerezo, S. P. Lord, M. S. Akhtar, and T. Chakraborty, “Critical behavioral traits foster peer engagement in online mental health communities,” *CoRR*, vol. abs/2309.01618, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2309.01618>
- [251] U. Lokala, A. Srivastava, T. G. Dastidar, T. Chakraborty, M. S. Akhtar, M. Panahiazar, and A. P. Sheth, “A computational approach to understand mental health from reddit: Knowledge-aware multitask learning framework,” in *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022*, C. Budak, M. Cha, and D. Quercia, Eds. AAAI Press, 2022, pp. 640–650. [Online]. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/19322>
- [252] J. Gibson, A. Katsamanis, F. Romero, B. Xiao, P. Georgiou, and S. Narayanan, “Multiple instance learning for behavioral coding,” *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 81–94, 2017.

- [253] I. Chatterjee, M. Gorsic, M. S. Hossain, J. D. Clapp, and V. D. Novak, “Automated classification of dyadic conversation scenarios using autonomic nervous system responses,” *IEEE Transactions on Affective Computing*, vol. 14, no. 04, pp. 3388–3395, oct 2023.
- [254] P. K. Adhikary, A. Srivastava, S. Kumar, S. M. Singh, P. Manuja, J. K. Gopinath, V. Krishnan, S. Kedia, K. S. Deb, and T. Chakraborty, “Exploring the efficacy of large language models in summarizing mental health counseling sessions: A benchmark study,” *CoRR*, vol. abs/2402.19052, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2402.19052>
- [255] A. Srivastava, T. Suresh, S. P. Lord, M. S. Akhtar, and T. Chakraborty, “Counseling summarization using mental health knowledge guided utterance filtering,” in *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, A. Zhang and H. Rangwala, Eds. ACM, 2022, pp. 3920–3930. [Online]. Available: <https://doi.org/10.1145/3534678.3539187>
- [256] O. Team, “Gpt-4 technical report,” 2024.
- [257] L. Brocki, G. C. Dyer, A. Gładka, and N. C. Chung, “Deep learning mental health dialogue system,” in *2023 IEEE International Conference on Big Data and Smart Computing (BigComp), 2023*, pp. 395–398.
- [258] J. Lu, A. Kannan, J. Yang, D. Parikh, and D. Batra, “Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model,” in *NeurIPS*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.
- [259] X. Shen, H. Su, Y. Li, W. Li, S. Niu, Y. Zhao, A. Aizawa, and G. Long, “A conditional variational framework for dialog generation,” in *Proceedings of the 55th Annual Meeting of the ACL*. Vancouver, Canada: ACL, Jul. 2017, pp. 504–509.
- [260] X. Wu, A. Martínez, and M. Klyen, “Dialog generation using multi-turn reasoning neural networks,” in *Proceedings of the 2018 Conference of the NAACL*. New Orleans, Louisiana: ACL, Jun. 2018, pp. 2049–2059.
- [261] A. Srivastava, I. Pandey, M. S. Akhtar, and T. Chakraborty, “Response-act guided reinforced dialogue generation for mental health counseling,” in *Proceedings of the ACM Web Conference 2023, ser. WWW '23*. New York, NY, USA: Association for Computing Machinery, 2023, p. 1118–1129. [Online]. Available: <https://doi.org/10.1145/3543507.3583380>
- [262] S. M. Schueller, A. Aguilera, and D. C. Mohr, “Ecological momentary interventions for depression and anxiety,” *Depression and anxiety*, vol. 34, no. 6, pp. 540–545, 2017.
- [263] M. J. Sullivan, “Strategies for working with difficult clients,” in *Parenting coordination in post-separation disputes: A comprehensive guide for practitioners*, S. A. Higuchi and S. J. Lally, Eds. American Psychological Association, 2014, pp. 107–122.
- [264] F. Ilievski, A. Oltramari, K. Ma, B. Zhang, D. L. McGuinness, and P. Szekely, “Dimensions of commonsense knowledge,” *Knowledge-Based Systems*, vol. 229, p. 107347, 2021.
- [265] N. Tandon, A. S. Varde, and G. de Melo, “Commonsense knowledge in machine intelligence,” *ACM SIGMOD Record*, vol. 46, no. 4, pp. 49–52, 2018.
- [266] S. Chaturvedi, H. Peng, and D. Roth, “Story comprehension for predicting what happens next,” in *Proceedings of the 2017 Conference on EMNLP*, 2017, pp. 1603–1614.

- [267] I. B. Schlicht, E. Sezerer, S. Tekir, O. Han, and Z. Boukhers, “Leveraging commonsense knowledge on classifying false news and determining checkworthiness of claims,” *arXiv preprint arXiv:2108.03731*, 2021.
- [268] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, “The curious case of neural text degeneration,” *arXiv preprint arXiv:1904.09751*, 2019.
- [269] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, “COMET: commonsense transformers for automatic knowledge graph construction,” *CoRR*, vol. abs/1906.05317, 2019. [Online]. Available: <http://arxiv.org/abs/1906.05317>
- [270] Q. Tu, Y. Li, J. Cui, B. Wang, J.-R. Wen, and R. Yan, “Misc: A mixed strategy-aware model integrating comet for emotional support conversation,” *arXiv preprint arXiv:2203.13560*, 2022.
- [271] D. Li, Y. Li, J. Zhang, K. Li, C. Wei, J. Cui, and B. Wang, “C3kg: A chinese commonsense conversation knowledge graph,” *arXiv preprint arXiv:2204.02549*, 2022.
- [272] M. Y. Chen, S. Li, and Y. Yang, “EmpHi: Generating empathetic responses with human-like intents,” in *Proceedings of the 2022 Conference of the NAACL: Human Language Technologies*. Seattle, United States: ACL, Jul. 2022, pp. 1063–1074.
- [273] J. Gao, Y. Liu, H. Deng, W. Wang, Y. Cao, J. Du, and R. Xu, “Improving empathetic response generation by recognizing emotion cause in conversations,” in *Findings of the EMNLP*, 2021, pp. 807–819.
- [274] M. Firdaus, U. Jain, A. Ekbal, and P. Bhattacharyya, “Seprg: sentiment aware emotion controlled personalized response generation,” in *Proceedings of the 14th International Conference on Natural Language Generation*, 2021, pp. 353–363.
- [275] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky, “Deep reinforcement learning for dialogue generation,” *arXiv preprint arXiv:1606.01541*, 2016.
- [276] T. Saha, S. Chopra, S. Saha, and P. Bhattacharyya, “Reinforcement learning based personalized neural dialogue generation,” in *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part IV 27*. Springer, 2020, pp. 709–716.
- [277] J. Wang, Y. Cheng, and W. Li, “Care: Causality reasoning for empathetic responses by conditional graph generation,” *arXiv preprint arXiv:2211.00255*, 2022.
- [278] J.-C. Gu, C.-H. Tan, C. Tao, Z.-H. Ling, H. Hu, X. Geng, and D. Jiang, “Hetermpc: A heterogeneous graph neural network for response generation in multi-party conversations,” *arXiv preprint arXiv:2203.08500*, 2022.
- [279] T. Zhao, K. Xie, and M. Eskenazi, “Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models,” in *Proceedings of the 2019 Conference of the NAACL*. Minnesota: ACL, Jun. 2019, pp. 1208–1218.
- [280] A. Amutio, C. Franco, M. de Carmen Pérez-Fuentes, J. J. Gázquez, and I. Mercader, “Mindfulness training for reducing anger, anxiety, and depression in fibromyalgia patients,” *Frontiers in Psychology*, vol. 5, p. 1572, 2015.
- [281] R. Karkar, N. Almuhanha, A. Althumali, A. Alkhalifah, R. Aljamaan, M. Alsuwaidan, and T. Alshammari, “Understanding people’s use of and perspectives on mood-tracking apps: Mixed methods study,” *JMIR mental health*, vol. 8, no. 8, p. e29368, 2021.

- [282] S. Wu, Y. Li, D. Zhang, Y. Zhou, and Z. Wu, “Diverse and informative dialogue generation with context-specific commonsense knowledge awareness,” in *Proceedings of the 58th Annual Meeting of the ACL*. Online: ACL, Jul. 2020, pp. 5811–5820.
- [283] C. Zheng, Y. Liu, W. Chen, Y. Leng, and M. Huang, “CoMAE: A multi-factor hierarchical framework for empathetic response generation,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 813–824. [Online]. Available: <https://aclanthology.org/2021.findings-acl.72>
- [284] S. Sabour, C. Zheng, and M. Huang, “Cem: Commonsense-aware empathetic response generation,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 11 229–11 237, Jun. 2022.
- [285] C. Beatty, T. Malik, S. Meheli, and C. Sinha, “Evaluating the therapeutic alliance with a free-text cbt conversational agent (wysa): A mixed-methods study,” *Frontiers in Digital Health*, vol. 4, 2022. [Online]. Available: <https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgh.2022.847991>
- [286] P. Crits-Christoph, A. Rieger, A. Gaines, and M. B. C. Gibbons, “Trust and respect in the patient-clinician relationship: preliminary development of a new scale,” *BMC Psychol*, 2019.
- [287] A. Li, L. Ma, Y. Mei, H. He, S. Zhang, H. Qiu, and Z. Lan, “Understanding client reactions in online mental health counseling,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 10 358–10 376. [Online]. Available: <https://aclanthology.org/2023.acl-long.577>
- [288] A. Darcy, J. Daniels, D. Salinger, P. Wicks, and A. Robinson, “Evidence of human-level bonds established with a digital conversational agent: Cross-sectional, retrospective observational study,” *JMIR Form Res*, vol. 5, no. 5, p. e27868, May 2021. [Online]. Available: <https://formative.jmir.org/2021/5/e27868>
- [289] E. L. van der Schyff, B. Ridout, K. L. Amon, R. Forsyth, and A. J. Campbell, “Providing self-led mental health support through an artificial intelligence–powered chat bot (leora) to meet the demand of mental health care,” *J Med Internet Res*, vol. 25, p. e46448, Jun 2023. [Online]. Available: <https://www.jmir.org/2023/1/e46448>
- [290] B. Spadaro, N. A. Martin-Key, E. Funnell, J. Benáček, and S. Bahn, “Opportunities for the implementation of a digital mental health assessment tool in the united kingdom: Exploratory survey study,” *JMIR Form Res*, vol. 7, p. e43271, Aug 2023. [Online]. Available: <https://formative.jmir.org/2023/1/e43271>
- [291] U. Smrke, I. Mlakar, S. Lin, B. Musil, and N. Plohl, “Language, speech, and facial expression features for artificial intelligence–based detection of cancer survivors’ depression: Scoping meta-review,” *JMIR Ment Health*, vol. 8, no. 12, p. e30439, Dec 2021. [Online]. Available: <https://mental.jmir.org/2021/12/e30439>
- [292] S. Cross, I. Bell, J. Nicholas, L. Valentine, S. Mangelsdorf, S. Baker, N. Titov, and M. Alvarez-Jimenez, “Use of ai in mental health care: Community and mental health professionals survey,” *JMIR Ment Health*, vol. 11, p. e60589, Oct 2024. [Online]. Available: <https://mental.jmir.org/2024/1/e60589>

- [293] K. Berry, A. Salter, R. Morris, S. James, and S. Bucci, “Assessing therapeutic alliance in the context of mhealth interventions for mental health problems: Development of the mobile agnew relationship measure (marm) questionnaire,” *J Med Internet Res*, vol. 20, no. 4, p. e90, Apr 2018. [Online]. Available: <http://www.jmir.org/2018/4/e90/>
- [294] G. Malhotra, A. Waheed, A. Srivastava, M. S. Akhtar, and T. Chakraborty, “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*, ser. WSDM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 735–745. [Online]. Available: <https://doi.org/10.1145/3488560.3498509>
- [295] A. Srivastava, T. Gupta, A. Cerezo, S. Peregrine, Lord, M. S. Akhtar, and T. Chakraborty, “Critical behavioral traits foster peer engagement in online mental health communities,” 2023. [Online]. Available: <https://arxiv.org/abs/2309.01618>
- [296] D. B. Olawade, O. Z. Wada, A. Odetayo, A. C. David-Olawade, F. Asaolu, and J. Eberhardt, “Enhancing mental health with artificial intelligence: Current trends and future prospects,” *Journal of Medicine, Surgery, and Public Health*, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:269222234>
- [297] A. Srivastava, S. Joshi, T. Chakraborty, and M. S. Akhtar, “Knowledge planning in large language models for domain-aligned counseling summarization,” 2024. [Online]. Available: <https://arxiv.org/abs/2409.14907>
- [298] M. Merino, J. F. Tornero-Aguilera, A. Rubio-Zarapuz, C. V. Villanueva-Tobaldo, A. Martin-Rodriguez, and V. J. Clemente-Suarez, “Body perceptions and psychological well-being: A review of the impact of social media and physical measurements on self-esteem and mental health with a focus on body image satisfaction and its relationship with cultural and gender factors,” in *Healthcare*, vol. 12, no. 14. MDPI, 2024, p. 1396.
- [299] G. Park, J. Chung, and S. Lee, “Human vs. machine-like representation in chatbot mental health counseling: the serial mediation of psychological distance and trust on compliance intention,” *Current Psychology*, vol. 43, no. 5, pp. 4352–4363, 2024.
- [300] P. K. Adhikary, A. Srivastava, S. Kumar, S. M. Singh, P. Manuja, J. K. Gopinath, V. Krishnan, S. Kedia, K. S. Deb, and T. Chakraborty, “Exploring the efficacy of large language models in summarizing mental health counseling sessions: A benchmark study,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.19052>
- [301] J. Bolsinger, M. Jaeger, P. Hoff, and A. Theodoridou, “Challenges and opportunities in building and maintaining a good therapeutic relationship in acute psychiatric settings: A narrative review,” *Frontiers in Psychiatry*, vol. 10, 2020. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pubmed/32009995>
- [302] S. Priebe and R. McCabe, “Therapeutic relationships in psychiatry: The basis of therapy or therapy in itself?” *International Review of Psychiatry*, vol. 20, pp. 521 – 526, 2008. [Online]. Available: <https://doi.org/10.1080/09540260802565257>
- [303] R. McCabe and S. Priebe, “The therapeutic relationship in the treatment of severe mental illness: A review of methods and findings,” *International Journal of Social Psychiatry*, vol. 50, pp. 115 – 128, 2004. [Online]. Available: <https://doi.org/10.1177/0020764004040959>

- [304] M. A. Kuhail, N. Alturki, J. Thomas, A. K. Alkhalifa, and A. Alshardan, “Human-human vs human-ai therapy: An empirical study,” *International Journal of Human–Computer Interaction*, pp. 1–12, 2024.
- [305] R. Plutchik, “A psychoevolutionary theory of emotions,” in *Social Science Information*, 1982, pp. 529–553.
- [306] J. Bowlby, *Attachment and loss*. Random House, 1969, no. 79.
- [307] E. H. Erikson, *Childhood and society*. Norton New York, 1963, vol. 2.
- [308] J. Birkhäuser, J. Gaab, J. Kossowsky, S. Hasler, P. Krummenacher, C. Werner, and H. Gerger, “Trust in the health care professional and health outcome: A meta-analysis,” *PloS one*, vol. 12, no. 2, p. e0170988, 2017.
- [309] D. Tao, H. Chui, S. Luk, and T. Lee, “Cuempathy: A counseling speech dataset for psychotherapy research,” in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2022, pp. 354–358.
- [310] Z. Zhu, C. Li, J. Pan, X. Li, Y. Xiao, Y. Chang, F. Zheng, and S. Wang, “Medic: A multimodal empathy dataset in counseling,” in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM ’23. New York, NY, USA: Association for Computing Machinery, 2023, p. 6054–6062. [Online]. Available: <https://doi.org/10.1145/3581783.3612346>
- [311] P.-C. Chen, M. Rohmatillah, Y.-T. Lin, and J.-T. Chien, “Convounsel: A conversational dataset for student counseling,” 2024. [Online]. Available: <https://arxiv.org/abs/2411.00604>
- [312] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019. [Online]. Available: <https://arxiv.org/abs/1810.04805>
- [313] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, “Mentalbert: Publicly available pretrained language models for mental healthcare,” 2021. [Online]. Available: <https://arxiv.org/abs/2110.15621>
- [314] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” 2019. [Online]. Available: <https://arxiv.org/abs/1907.11692>
- [315] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “Albert: A lite bert for self-supervised learning of language representations,” 2020. [Online]. Available: <https://arxiv.org/abs/1909.11942>
- [316] P. He, X. Liu, J. Gao, and W. Chen, “Deberta: Decoding-enhanced bert with disentangled attention,” 2021. [Online]. Available: <https://arxiv.org/abs/2006.03654>
- [317] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, “Xlnet: Generalized autoregressive pretraining for language understanding,” 2020. [Online]. Available: <https://arxiv.org/abs/1906.08237>
- [318] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” 2019. [Online]. Available: <https://arxiv.org/abs/1910.13461>

- [319] K. Yang, T. Zhang, Z. Kuang, Q. Xie, and S. Ananiadou, “Mentalllama: Interpretable mental health analysis on social media with large language models,” 2023.
- [320] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: <http://jmlr.org/papers/v21/20-074.html>
- [321] D. et al., “The llama 3 herd of models,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.21783>
- [322] M. A. et al., “Phi-3 technical report: A highly capable language model locally on your phone,” 2024. [Online]. Available: <https://arxiv.org/abs/2404.14219>
- [323] F. Petruzzellis, A. Testolin, and A. Sperduti, “Benchmarking gpt-4 on algorithmic problems: A systematic evaluation of prompting strategies,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.17396>
- [324] G. Team, “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” 2024. [Online]. Available: <https://arxiv.org/abs/2403.05530>
- [325] T. Butryn, L. Bryant, C. Marchionni, and F. Sholevar, “The shortage of psychiatrists and other mental health providers: Causes, current state, and potential solutions,” *International Journal of Academic Medicine*, vol. 3, no. 1, pp. 5–9, 2017. [Online]. Available: <https://www.ijam-web.org/article.asp?issn=2455-5568;year=2017;volume=3;issue=1;spage=5;epage=9;aulast=Butryn;t=6>
- [326] S. Rathod, N. Pinninti, M. Irfan, P. Gorczynski, P. Rathod, L. Gega, and F. Naeem, “Mental health service provision in low- and middle-income countries,” *Health Services Insights*, vol. 10, p. 1178632917694350, 2017, pMID: 28469456. [Online]. Available: <https://doi.org/10.1177/1178632917694350>
- [327] U. Lokala, A. Srivastava, T. G. Dastidar, T. Chakraborty, M. S. Akhtar, M. Panahiazar, and A. Sheth, “A computational approach to understand mental health from reddit: Knowledge-aware multitask learning framework,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, no. 1, pp. 640–650, May 2022. [Online]. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/19322>
- [328] A. Park, M. Conway, and A. T. Chen, “Examining thematic similarity, difference, and membership in three online mental health communities from reddit,” *Comput. Hum. Behav.*, vol. 78, no. C, p. 98–112, jan 2018. [Online]. Available: <https://doi.org/10.1016/j.chb.2017.09.001>
- [329] M. De Choudhury and S. De, “Mental health discourse on reddit: Self-disclosure, social support, and anonymity,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 71–80, May 2014. [Online]. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/14526>
- [330] A. Sharma, M. Choudhury, T. Althoff, and A. Sharma, “Engagement patterns of peer-to-peer interactions on mental health platforms,” in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 614–625.
- [331] Y. Pruksachatkun, S. R. Pendse, and A. Sharma, “Moments of change: Analyzing peer-based cognitive support in online mental health forums,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI ’19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–13. [Online]. Available: <https://doi.org/10.1145/3290605.3300294>

- [332] S. K. Ernala, T. Labetoulle, F. Bane, M. L. Birnbaum, A. F. Rizvi, J. M. Kane, and M. De Choudhury, “Characterizing audience engagement and assessing its impact on social media disclosures of mental illnesses,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, Jun. 2018. [Online]. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/15027>
- [333] M. K. Evans, L. Donelle, and L. Hume-Loveland, “Social support and online postpartum depression discussion groups: a content analysis.” *Patient education and counseling*, vol. 87 3, pp. 405–10, 2012.
- [334] S. Rains, E. Peterson, and K. Wright, “Communicating social support in computer-mediated contexts: A meta-analytic review of content analyses examining support messages shared online among individuals coping with illness,” *Communication Monographs*, vol. 82, pp. 1–28, 03 2015.
- [335] N. Andalibi, P. Ozturk, and A. Forte, “Sensitive self-disclosures, responses, and social support on instagram: The case of depression,” in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, ser. CSCW ’17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1485–1500. [Online]. Available: <https://doi.org/10.1145/2998181.2998243>
- [336] A. Donnelly, “A global mental health approach to depression in adolescence. nat. mental health,” *Nature Mental Health*, p. 527–528, 2023. [Online]. Available: <https://doi.org/10.1038/s44220-023-00107-y>
- [337] C. Seiferth, L. Vogel, and B. Aas, “How to e-mental health: a guideline for researchers and practitioners using digital technology in the context of mental health,” *Nature Mental Health*, p. 542–554, 2023. [Online]. Available: <https://doi.org/10.1038/s44220-023-00085-1>
- [338] M. Hoven, J. Luigjes, and D. Denys, “How do confidence and self-beliefs relate in psychopathology: a transdiagnostic approach,” *Nature Mental Health*, p. 337–345, 2023. [Online]. Available: <https://doi.org/10.1038/s44220-023-00062-8>
- [339] Y. Re’em, E. Stelson, and H. Davis, “Factors associated with psychiatric outcomes and coping in long covid,” *Nature Mental Health*, p. 361–372, 2023. [Online]. Available: <https://doi.org/10.1038/s44220-023-00064-6>
- [340] A. Sharma, M. Choudhury, T. Althoff, and A. Sharma, “Engagement patterns of peer-to-peer interactions on mental health platforms,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, no. 1, pp. 614–625, May 2020. [Online]. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/7328>
- [341] K. Ali, L. Farrer, A. Gulliver, and K. M. Griffiths, “Online peer-to-peer support for young people with mental health problems: A systematic review,” *JMIR Mental Health*, vol. 2, no. 2, p. e19, May 2015. [Online]. Available: <http://mental.jmir.org/2015/2/e19/>
- [342] B. Biagiante, S. H. Quraishi, and D. A. Schlosser, “Potential benefits of incorporating peer-to-peer interactions into digital interventions for psychotic disorders: A systematic review,” *Psychiatric Services*, vol. 69, no. 4, pp. 377–388, 2018, PMID: 29241435. [Online]. Available: <https://doi.org/10.1176/appi.ps.201700283>
- [343] J. A. Naslund, K. A. Aschbrenner, L. A. Marsch, and S. J. Bartels, “The future of mental health care: peer-to-peer support and social media,” *Epidemiology and Psychiatric Sciences*, vol. 25, no. 2, p. 113–122, 2016.

- [344] A. E. Kazdin, “Annual research review: Expanding mental health services through novel models of intervention delivery,” *Journal of Child Psychology and Psychiatry*, vol. 60, no. 4, pp. 455–472, 2019. [Online]. Available: <https://acamh.onlinelibrary.wiley.com/doi/abs/10.1111/jcpp.12937>
- [345] H. Yan, E. E. Fitzsimmons-Craft, M. Goodman, M. Krauss, S. Das, and P. Cavazos-Rehg, “Automatic detection of eating disorder-related social media posts that could benefit from a mental health intervention,” *International Journal of Eating Disorders*, vol. 52, no. 10, pp. 1150–1156, 2019. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/eat.23148>
- [346] S. B. Bonsall, A. J. Leone, B. P. Miller, and K. Rennekamp, “A plain english measure of financial reporting readability,” *Journal of Accounting and Economics*, vol. 63, no. 2, pp. 329–357, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0165410117300162>
- [347] A. S. Cowen and D. Keltner, “Self-report captures 27 distinct categories of emotion bridged by continuous gradients,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 38, pp. E7900–E7909, 2017.
- [348] J. S. Kim and C. Franklin, “Understanding emotional change in solution-focused brief therapy: Facilitating positive emotions,” *Best Practices in Mental Health*, vol. 11, no. 1, pp. 25–41, 2015.
- [349] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: <https://arxiv.org/abs/1908.10084>
- [350] T. Althoff, K. Clark, and J. Leskovec, “Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health,” *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 463–476, 08 2016. [Online]. Available: https://doi.org/10.1162/tacl_a_00111
- [351] M. Sit, S. A. Elliott, K. S. Wright, S. D. Scott, and L. Hartling, “Youth mental health help-seeking information needs and experiences: A thematic analysis of reddit posts,” *Youth & Society*, vol. 0, no. 0, p. 0044118X221129642, 0. [Online]. Available: <https://doi.org/10.1177/0044118X221129642>
- [352] K. M. Griffiths, A. J. Mackinnon, D. A. Crisp, H. Christensen, K. Bennett, and L. Farrer, “The effectiveness of an online support group for members of the community with depression: A randomised controlled trial,” *PLOS ONE*, vol. 7, no. 12, pp. 1–9, 12 2012. [Online]. Available: <https://doi.org/10.1371/journal.pone.0053244>
- [353] J. A. Naslund, S. W. Grande, K. A. Aschbrenner, and G. Elwyn, “Naturally occurring peer support through social media: The experiences of individuals with severe mental illness using youtube,” *PLOS ONE*, vol. 9, no. 10, pp. 1–9, 10 2014. [Online]. Available: <https://doi.org/10.1371/journal.pone.0110171>
- [354] C. L. Odgers and M. R. Jensen, “Annual research review: Adolescent mental health in the digital age: facts, fears, and future directions,” *Journal of Child Psychology and Psychiatry*, vol. 61, no. 3, pp. 336–348, 2020. [Online]. Available: <https://acamh.onlinelibrary.wiley.com/doi/abs/10.1111/jcpp.13190>
- [355] D. Elmquist and C. McLaughlin, “Social media use among adolescents coping with mental health,” *Contemporary School Psychology*, vol. 22, 11 2017.

- [356] D. Z. Q. Gan, L. McGillivray, M. E. Larsen, and M. Torok, "Promoting engagement with self-guided digital therapeutics for mental health: Insights from a cross-sectional survey of end-users," *Journal of Clinical Psychology*, vol. 79, no. 5, pp. 1386–1397, 2023. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jclp.23486>
- [357] J. Torous, J. Nicholas, M. E. Larsen, J. Firth, and H. Christensen, "Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements," *BMJ Ment Health*, vol. 21, no. 3, pp. 116–119, 2018. [Online]. Available: <https://mentalhealth.bmj.com/content/21/3/116>
- [358] K. Palmer and V. Burrows, "Ethical and safety concerns regarding the use of mental health-related apps in counseling: Considerations for counselors," *Journal of Technology in Behavioral Science*, vol. 6, 03 2021.
- [359] C. Burr, J. Morley, M. Taddeo, and L. Floridi, "Digital psychiatry: Risks and opportunities for public health and wellbeing," *IEEE Transactions on Technology and Society*, vol. 1, no. 1, pp. 21–33, 2020.
- [360] O. Oyeboade, F. Alqahtani, and R. Orji, "Using machine learning and thematic analysis methods to evaluate mental health apps based on user reviews," *IEEE Access*, vol. 8, pp. 111 141–111 158, 2020.