Enhancing Empathy in Text-Based Teletherapy with Emotional State Inference

William R. Kearns

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2023

Reading Committee:
Trevor Cohen, Chair
Weichao Yuwen
Alex Marin

Program Authorized to Offer Degree:
Biomedical & Health Informatics

University of Washington

Abstract

Enhancing Empathy in Text-Based Teletherapy with Emotional State Inference

William R. Kearns
Chair of Supervisory Committee:
Professor Trevor Cohen
Biomedical Informatics and Medical Education

Over half of the U.S. population lives in an area without adequate access to mental health care and the unmet demand for mental health services has shifted to care providers who have not been trained to provide mental health support. This work represents a step toward addressing this supply-demand imbalance by applying recent advances in conversational AI.

The central hypothesis of this work is that both the quality and efficiency of text-based telehealth can be improved through recent advances in conversational AI. This hypothesis was evaluated with three aims: (Aim 1) explored the ability of computational methods to infer high-fidelity representations of emotional states as a precursor to empathy, (Aim 2) evaluated these representations as features for a transformer-based empathic response predictor, (Aim 3) piloted this system as a component of a teletherapy platform for the delivery of problem-solving therapy by nurses and psychologists. The results of these aims validate this core hypothesis by successfully collecting emotional health information through an automated SMS-based intervention and by significantly improving empathic accuracy and reducing response times of human care providers using an AI-augmented chat interface.

Together the components of this dissertation provide a unified solution that can help to increase access to mental health care by automating the remote monitoring of emotional health, expanding the number of

individuals who can provide protocolized care, and enhancing the efficiency and empathy of the care provided. During the course of this work, I developed a novel evaluation paradigm to better measure how emotion recognition systems can help to track emotional health through automated journaling exercises, applied these measures to predict empathic responses, and evaluated a support tool to assist care providers in delivering problem-solving therapy.

Acknowledgments

My journey through my doctoral program would not have been possible without the support of so many people. A special thank you to my parents for encouraging me to follow my passions and to be inquisitive about everything. To my spouse, Elizabeth, thank you for being by my side every step of the way and sharing the simplest to the most significant moments in my life. I thank my other family and friends across the world for the good laughs and deep discussions that kept me motivated during the tough times.

Thank you to Trevor Cohen for supervising my dissertation during a period of rapid change and for the thought-provoking conversations. Thank you to my committee member and COCO co-founder Weichao Yuwen for sharing clinical research knowledge and creating a compassionate and empathetic culture for the COCO team. Thank you to my committee member Alex Marin for sharing expert knowledge in cutting-edge methods for dialogue system development and industry applications. Thank you to Gina Levow for introducing me to dialogue systems and sitting in as my GSR. Thank you to the Department of Biomedical Informatics and Medical Education, especially to Peter Tarczy-Hornoch, John Gennari, Dave Masuda, Jim Pfaedtner, Marni Levy, and Annie Chen for providing the resources, knowledge, and guidance I needed to successfully complete my dissertation and beyond. Thank you to my BHI colleagues for the entertaining trivia nights, intellectual exchanges, and other moments of camaraderie that made for a more enriching university experience. I also thank the Department of Computational Linguistics, especially Fei Xia and Emily Bender, for teaching me how to develop machines that can read and write. Thank you to Pramod Gupta for teaching me Bayesian statistics from base principles which helped me calculate the statistical significance of my results.

I would like to thank the entire COCO team with a special thank you to Myra Divina, Liying Wang, Stanley Wang, Kelly Hou, Yinzhou Wang, and Alex Yuwen for investing their time and energy to develop the COCO platform and the entire care team for sharing their clinical expertise and helping to provide much needed care for caregivers during the pandemic.

Thank you to the Center for Innovation in Sleep Self-Management, especially George Demiris, Hilaire Thompson, and Margaret Heitkemper for funding the remainder of my doctoral studies and providing the mobile usability lab for the eye tracking experiments. Thank you to the National Library of Medicine for funding my initial years in the program and Comotion for providing funding for the development of COCO. I want to also thank the Research Computing Club and the Student Technology Fund which provided an allocation on the Hyak and Mox supercomputers and cloud computing credits that I used to conduct parts of my research.

I am deeply grateful for all these people, groups, and many more that supported me professionally and personally during the completion of this dissertation.

Table of Contents

List of Abbreviations

The following table contains the abbreviations used in this dissertation

AI	Artificial Intelligence
BART	Bidirectional Auto-Regressive Transformer
BERT	Bidirectional Encoder Representations from Transformers
CAI	Conversational Artificial Intelligence
COCO	Caring for Caregivers Online
COMET	Commonsense Reasoning from Transformers
COVID	Coronavirus Disease
DM	Dialog Management
EC	(Clinical) Empathy Criteria
ECA	Embodied Conversational Agent
EMA	Ecological Momentary Assessment
EMR	Electronic Medical Record
ERC	Emotion Recognition from Conversation
ERP	Empathic Response Prediction
ESI	Emotional State Inference
GAD	General Anxiety Disorder
GPT	Generative Pre-trained Transformer
IAA	Inter Annotator Agreement
KG	Knowledge Graph

MITI	Motivational Interviewing Treatment Integrity
NLG	Natural Language Generation
NLU	Natural Language Understanding
PHQ	Patient Health Questionnaire
PLM	Pre-trained Language Model
PST	Problem Solving Therapy
RCT	Randomized Controlled Trial
SMS	Short Messaging Service
USSD	Unstructured Supplementary Service Data
WHO	World Health Organization
WOZ	Wizard of Oz

Chapter 1: Introduction

Over half of the U.S. population lives in an area without adequate access to mental health care (HRSA 2019), with up to 90% of the population underserved in some countries (Freeman 2022). Both care providers and those who can afford their services are centralized in affluent, urban centers. As a result of this imbalance, the unmet demand for mental health services has shifted to care providers in primary care, nursing, coaching, or peers who have not been trained to provide mental health support (Cherry 2018). This work presents a step toward addressing the supply-demand imbalance between those who need mental health services and those who can provide treatment by applying conversational artificial intelligence (AI) to reduce response times during asynchronous communication for more immediate engagement and lower the educational barrier to entry, thereby increasing the number of individuals who can provide care.

Empathy is a key quality measure in healthcare delivery because it is linked to improved patient satisfaction as well as outcomes (Hojat 2016). It has been defined within the clinical setting (Mercer & Reynolds 2002) as satisfying the following three empathy criteria (EC):

(EC1) understanding the patient's situation, perspective, and feelings

(EC2) communicating that understanding and checking its accuracy

(EC3) acting on that understanding with the patient in a therapeutic way.

People differ in their ability to be empathic, even trained therapists or peer counselors. Fortunately, empathy is a skill that can be developed, and studies have shown it can be learned through assistive technology (Menezes

2021). However, empathy can also be negatively impacted by a condition called compassion fatigue or secondary traumatic stress, a work-related stress condition that leads to reduced empathy in care providers. This, combined with other work-related stress from the COVID-19 pandemic, is leading to unprecedented clinician burnout and exiting the profession (Sinclair 2017) further exacerbating the supply-demand imbalance. Toward the goals of reducing empathy fatigue and increasing the availability of treatment, I developed and evaluated a support tool to assist care providers without mental health training to deliver a protocolized therapy session based on problem-solving therapy (PST) (D'Zurilla 1999) emphatically.

Specifically, I describe the development and pilot evaluation of a conversational AI-assisted behavioral health care delivery system (Figure 1.1) that can assist in monitoring the emotional health of individuals in the community in between mental health care visits to provide context to their mental health care providers and support other health workers in the community without mental health training to deliver problem-solving therapy sessions accurately and efficiently.

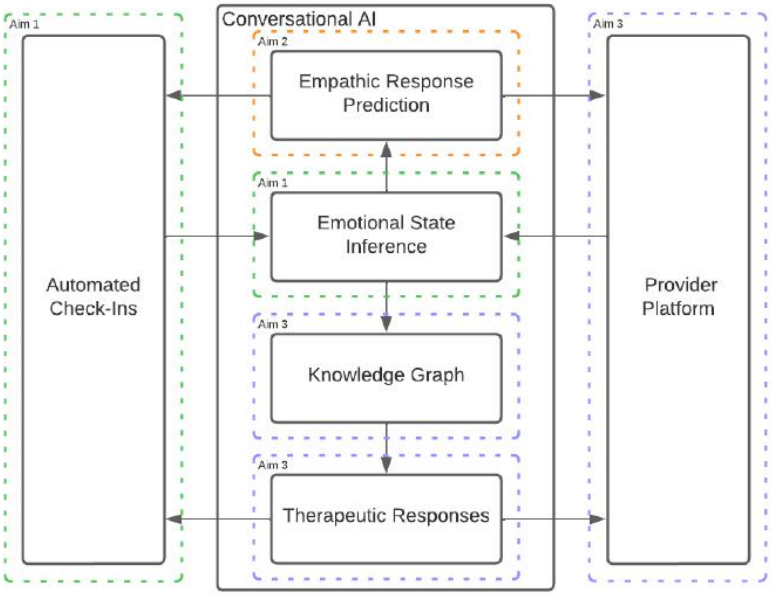


Figure 1.1: The AI-assisted care delivery system architecture proposed in this work with components numbered with the Aims in which they are evaluated.

The first component that will be described is the automated check-in delivered through a journaling exercise that provides insight into the emotional health of the population. The data received by that component can be used to monitor and continuously improve an emotional state inference component, a key feature for empathic response prediction (ERP). The knowledge graph is used to store information about the patient and expert-curated suggestions for goals, solutions, and resources. These components are all used to provide empathy-related and therapy-related response suggestions for the care provider through the provider platform.

1.1 Motivations

This work occurred within a wider context within health informatics that includes the proliferation of devices that track physical health to generate what are referred to as patient-generated health data. These data have been recognized as having the potential to improve medical decision-making by providing a more complete picture of health outside the point of care (Jim 2020). Due to the utility and vastness of the data generated by these devices, AI has emerged as a valuable component to incorporate these data into personalizing patient care while limiting additional demand on care providers that may lead to burnout (Ye 2021).

On the other hand, the effects of environmental influences on our emotional health are not well-tracked, yet many health behaviors and outcomes result from how we respond to our emotional states, such as *anger* or *grief*. Unfortunately, even when available, standard quantitative measures such as the Patient Health Questionnaire-9 (PHQ-9) and General Anxiety Disorder-7 (GAD-7) summarize this information into a single score to describe the past two weeks, making them susceptible to recency bias and devoid of insight into causal pathways (e.g. rising early and taking care of caregiving responsibilities before work leading to anxiety about performance at work which in turn led to sleep disturbances). It is well-known that people have difficulty recalling affective information (Thomas 1990, Safer 2002, Wenze 2012) and have a bias toward attributing more positive affect than was experienced at the time of the event (Colombo 2020). As a result, care providers spend

significant time attempting to collect information about the causes and effects of emotional health during the session and then make decisions based on the client's recollection at the point of care which is incomplete.

Digital health interventions, on the other hand, are capable of collecting a snapshot of emotional health at more regular intervals giving a clearer picture of emotional triggers and the effect of treatment (Huberty 2021, Schueller 2021). This has led to a rise in digital health interventions to track mood trajectories as an alternative measure of emotional health. These features fit within a broader class referred to as ecological momentary assessments (EMAs) which ask a person to provide in-the-moment information about their mood and behavior in relation to events. EMAs are widely used within clinical psychology research to address the aforementioned gaps in autobiographical memory that widen with time from the event (Shiffman 2008). While EMAs have proven valuable in research settings, they have yet to be fully adopted in clinical workflows due to concerns about the burden on patients and data quality (e.g., self-report biases and missing data) (Williams 2021, Stinson 2022). These factors have opened up the potential for new lines of inquiry in health informatics, namely the collection, analysis, and utilization of affective state information to improve behavioral health. As with other forms of patient-generated health data, care providers may benefit from AI-supported tools to incorporate this information into decision-making while limiting additional burden on care providers. Toward this goal, I evaluated methods to measure emotional health through automated journaling exercises, applied these inferred emotional states to improve empathic response prediction, and evaluated human-AI collaboration between care providers and these systems within an AI-augmented provider platform.

1.2. Gaps

Emotional state inference in teletherapy has primarily been studied from video or audio signals (Devault 2014, Stratou 2015). In the broader natural language processing literature, emotional state inference from text has been limited to lexical methods of sentiment detection (Alsaeedi 2019, Mehta 2020) and, more recently, emotion detection (Zahiri 2017, Li 2017, Poria 2019, Demszky 2020). Recent work has explored tuning language models to generate *mental state representations* that include emotional state inferences (Bosselut 2019) by pre-training on a crowdsourced commonsense knowledge graph (Sap 2019, Hwang 2020). This is significant because understanding the semantics of events potentially opens the door to more accurate mental state representations including emotional states than are possible from video or audio signals alone. These representations improved emotion detection on television dialogues datasets annotated for 6-7 emotion categories (Ghosal 2020, Zhu 2021). Within the mental health domain, contemporaneous work has shown that these representations improve depression and stress detection performance on social media forum data and stress factor classification from crowdsourced data (Yang 2022). However, the performance of these representations for *emotional state inference* on mental health intervention data has yet to be explored.

Unfortunately, all of the methods to date have been evaluated on datasets annotated by third-parties rather than by the speaker themselves, which has resulted in the evaluations measuring the ability for models to replicate normative expectations of how an individual would feel as a result of a particular event, which does not account for variance in the ground truth emotional state of the speaker. This in turn has resulted in both a focus on reductionist representations of emotion categories to improve inter annotator agreement (IAA) and a large percentage of *neutral* labels (as the annotators were unsure as to the correct emotional state) which limits their utility for downstream applications that rely on more granular emotional understanding. This presents a problem for applying these models within a healthcare setting. In the critical scenario of patient monitoring,

model performance is paramount. Knowing how an individual actually felt should take priority over mimicking the ability of another person to infer how they felt. To elucidate this point, it is possible for an individual to feel contrary to an expected emotion. For example, someone may feel guilt instead of pride at the accomplishment of a task, which would present an opportunity for a technique referred to as cognitive reframing. For this reason, I collected a dataset of self-reported *event-emotional state* pairs through a daily journaling exercise and proposed a novel paradigm to evaluate the emotion recognition methods on their ability to predict self-reported emotional states.

As research in empathic dialog is an emerging field, there are no publicly available tools to date that directly assist care professionals in responding emphatically to patient messages. Consequently, there is little known about which techniques would be best suited to achieve this goal and how this technology may be incorporated into their existing workflows.

1.3 Research Plan

The central hypothesis of this work is that both empathy and efficiency of text-based teletherapy (measured as the combination of quality and speed) can be improved through conversational AI. This hypothesis was evaluated with three main goals: (1) to explore the ability of computational methods to infer high-fidelity representations of emotional state as a precursor to empathy, (2) to evaluate emotional state representations as features for a transformer-based empathic response predictor, and (3) to evaluate an AI-assisted teletherapy interface for the delivery of problem-solving therapy by measuring the efficiency and accuracy of text-based problem-solving therapy sessions delivered by nurses and psychologists in two virtual patient scenarios. These goals are further specified as the following aims:

Aim 1: Compare methods to infer granular self-described emotional reactions to daily events.

Digital health interventions have thus far relied on self-reported values of affective states through button selection, limiting the options available to describe emotions. Similarly, text-based *emotional state inference* systems have used *sentiment detection*, including three labels (positive, neutral, or negative), or *emotion detection*, including distinguishing between 6 (Ekman 1971) and 28 (Cowen 2017) categories labeled by third-parties rather than self-reported. This reliance on human annotation rather than self-reports means the labels may not match an individual's experience and may result in models learning to pick up on only a subset of the possible signals in text (such as specific language or tone). This leads to model bias, as has been found to be the case for other annotation tasks (e.g. natural language inference where contradictions often contain negation simplifying the task and limiting generalizability) (Gururangan 2018). Further, it means that emotion detection models have not been validated to predict self-reported emotional states in the context of mood tracking. To address this gap, I evaluate the relative performance of *emotional state inference* methods to predict 3,465 *event-emotion state* pairs representing 217 unique self-reported emotional states from a text-based journaling intervention (**Example 1.1**).

Bot:	Tell me about an experience that could have been better for you today.	
User:	My son being really intense and stressed out	(Event)
Bot:	In one or two words, how did that make you feel?	
User:	Overwhelmed	(Emotional State)

Example 1.1: Emotional State Inference task is to predict the self-reported emotional state of the speaker resulting from the given the event

Aim 2: Evaluate mental state representation methods including explicit emotional state inference on the task of empathic response prediction.

Recognizing and responding to the emotional needs of clients during text-based teletherapy is challenging for humans and machines. In addition, clients often share events without explicitly expressing their emotional states or needs, suggesting a role for commonsense reasoning to infer implicit information from the text (**Example 1.2**). However, these tasks can be accomplished in several ways using contemporary neural architectures, with representational choices ranging from dense hidden state representations to predictions of discrete emotion categories. In this aim, I sought to determine the best representation methods to improve empathic response prediction within a contemporary dialog system infrastructure and the potential value of explicit emotional state inference.

User:	I wasn't as focused as I would like to be.	(Event)
Bot:	Staying focused can be a challenge, and it's completely normal to feel frustrated when we don't meet our own expectations. Is there anything you can do to help yourself stay more focused in the future?	(Response Text)
	[*Negative* + *Work* + *Unproductive*]	(Response Label)

Example 1.2: Empathic Response Prediction task is to predict the most appropriate response label given an event through classification

Aim 3: Evaluate novice and expert usage of a clinical decision support tool for empathic delivery of problem-solving therapy.

Responding to client messages is a significant portion of the daily work of telehealth care providers. This aim answers the question of to what extent a clinical decision support tool can enhance empathy, increase the efficiency of mental health care delivery, and reduce the gap between experts in psychotherapy and non-experts

with limited mental health training. This is achieved through the suggestion of both empathic responses, using the method developed in Aim 2, and therapeutic responses, using templates with slots filled by a knowledge graph based on symptoms inferred from client messages. The evaluation is conducted through care provider interactions with two virtual patient scenarios.

1.4 Relevance to Healthcare

Improving Access

56% of the population live in an area without enough therapists to support the mental health needs of the population, as defined by having more than 30,000 population per therapist (HRSA 2019). This was further exacerbated by the COVID-19 pandemic, during which rates of mental health-related diagnoses increased including anxiety diagnoses by 25.6% and depression diagnoses by 27.6% (COVID-19 Mental Disorders Collaborators 2021) globally with even higher rates for adolescents. The surge in demand for mental health services intersecting with the shortage of mental health providers led to burnout among care providers and required a shift toward alternative forms of care delivery by providers who were not necessarily trained in providing therapy, e.g., primary care providers, community health workers, or health and wellness coaches (Jetty 2021, Jordan 2021). I posit that technology akin to that developed in the present work can help address population health needs by assisting care providers in the delivery of protocolized therapies, reducing workforce fatigue and cost of care delivery, contributing toward the Institute of Healthcare Improvement (IHI) triple aim of reduced cost, improved population health, and improved healthcare quality.

Improving Measurement

Traditional therapy has been limited by the number of touchpoints between the therapist and their client. In addition, the electronic medical record (EMR) has limited data on patients' mental states between health care

visits and none for their caregivers, making risk management and personalizing therapy difficult. The gold standard clinical measurements of mental state, the PHQ-9 and GAD-7, are primarily taken at the point-of-care and ask patients to summarize their thoughts and emotions from the previous two weeks. The result is a score of average depression and anxiety levels over that period, which is not granular enough to help make actionable changes in relation to specific triggering events. This requires spending time during the visit or across multiple sessions to pinpoint these details as the patient themselves may not be aware of the underlying cause of their symptoms. Therefore, it is preferable to collect these mental states in the context of the events that triggered them and provide immediate feedback to patients. For this reason, this work included a daily journaling exercise to collect daily events in the lives of family caregivers and their emotional reactions to them.

1.5 Roadmap

The remainder of this work is divided into seven chapters. The second chapter covers the background information necessary to understand the methods and contribution of the work at hand. In the third chapter, I provide an historical account of shifts in mental health and telehealth that resulted from the COVID-19 pandemic as well as details of a system that I deployed during the early stages of the pandemic. In the fourth chapter, I introduce and describe the evaluation of approaches to emotional state inference from text to rank the fidelity of mental state representations. In the fifth chapter, I describe work in which I evaluated the utility of mental state representation methods on empathic response selection. In the sixth chapter, I describe work that incorporated this model into a teletherapy interface and evaluated its ability to improve the empathy and speed of teletherapy. In the seventh chapter, I conclude by summarizing the contribution of this work and provide directions for future work.

Chapter 2: Background

This chapter explains the necessary context for understanding the contributions of this work within the broader literature and provides the key technical background to understand the methods in subsequent chapters.

This work intersects with the field of *affective computing*, which aims to create emotionally intelligent systems (Picard 1997), and where recent advances have opened up new opportunities for empathic dialog systems that have yet to be explored in the health domain. A fundamental task for emotionally aware dialog systems involves recognizing the emotional state of a speaker. Work in this area has been informed by theories of human emotion and supported by publicly-available datasets for training and evaluation of models.

I begin by presenting a systematic review of health dialog systems to highlight the application areas within healthcare where the research presented in future chapters may be applied. I then introduce the fundamentals of natural language processing and recent advances that enable this work. With this context, I then outline emotional theories, existing datasets, and the top-performing models on these datasets. I conclude by summarizing the gaps that exist in the current literature and how my work contributes to advancing the field.

2.1 Health Dialog Systems

This section explores prior work in health dialog systems, where they have been applied, current limitations, and opportunity areas for further development. This section is reproduced from:

Kearns, W.R., Chi, N., Choi, Y.K., Lin, S., Thompson, H.J., & Demiris, G. (2019). A Systematic Review of Health Dialog Systems. *Methods of information in medicine, 58 6*, 179-193 .

Dialog systems are computer programs that simulate conversational intelligence often to carry out fixed tasks. Alexa, Cortana, Google Assistant, and Siri are commercial dialog systems that fulfill the role of digital assistants most frequently to help users in achieving simple tasks, e.g. to make a call, set a reminder, or access a weather report. While there is a large body of work on the development of dialog systems in the health domain, a systematic review found that conversational systems have seen limited integration into clinical practice (Laranjo 2018). This review seeks to enumerate opportunities to apply dialog systems toward the improvement of healthcare while identifying both gaps in the current literature that may impede their implementation and recommendations that may improve their success in medical practice.

A scoping review of embodied conversational agents (ECAs) in clinical psychology found that although this technology showed potential, limitations of the evaluation methods used to assess ECAs impede their implementation in clinical practice (Provoost 2017). This concern was reinforced by a systematic review of conversational agents in healthcare that found many systems were evaluated using ad hoc surveys rather than validated scales, suffered from a lack of reproducibility, and were often not evaluated with regard to patient safety (Laranjo 2018). Another review focused on dialog system methods and design (Montenegro 2019). This review answers a variety of research questions and provides a taxonomy of conversational agents in healthcare. However, a concern about this review is that the authors used the h-index of retrieved papers to automate part of the screening process, resulting in a bias toward well-established rather than emerging work. Our review updates and extends these prior works, using expanded search criteria that retrieved more health dialog system studies than had previously been reviewed. Further, this review focuses on two specific research questions:

(Q1) What are the application domains in which health dialog systems have been applied?

(Q2) To what extent have health dialog systems been evaluated?

Consequently, it provides more depth than prior reviews with regard to application domains and provides a perspective narrative through the lens of health services that was absent in prior reviews.

2.1.1 Background

Dialog system architectures have traditionally been built as a pipeline of components (Figure 2.1). Most typically, these consist of natural language understanding (NLU), dialog management (DM), and natural language generation (NLG). We briefly cover each of these components in the following subsections. For a methodological review of health dialog systems, we direct the reader to Bickmore and Giorgino, which covers methods for most of the dialog systems included in this study (Bickmore 2006). For a more recent review of neural approaches to conversational intelligence, we direct the interested reader to (Gao 2018).

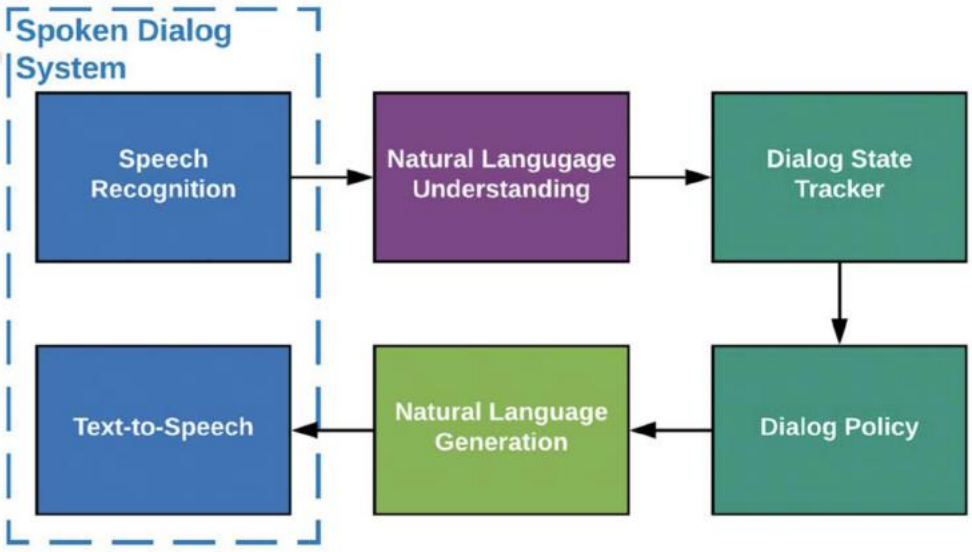


Figure 2.1 Common components of dialog systems

Natural Language Understanding

The NLU component extracts meaning from a user utterance at each turn of the dialog. Methods range from keyword and pattern matching to statistical models trained using distributional semantic representations. Having some form of NLU was a key criterion in our review and led to many systems being excluded for not having a conversational user interface, but rather relying on button or drop-down menu selections as input, a design choice that was largely made in the literature due to the added complexity of processing speech or free-text input. However, recent open-source frameworks (e.g., Rasa) and commercial services (e.g., Dialogflow, Alexa Skills Kit, Microsoft Bot Framework) for training HIPAA-aligned NLU models have reduced barriers to the implementation of these methods for digital health applications. Recent approaches to NLU, have trained models to recognize the entities, or slots, and the intent from a user utterance (Sarikaya 2016). Intent recognition has often been achieved through support vector machines as they can achieve high performance with only a few hundred examples (De Mori 2008). However, recent advances in semantic representation have improved the performance of neural approaches, especially with limited labeled training data. Neural approaches leveraging pre-trained language models (PLM) such as bidirectional encoder representations from transformers (BERT) have led to state-of-the-art performance across tasks relevant to NLU (Devlin 2019).

Dialog Management

State-based DM typically consists of two parts, that is, a state tracker that leverages an internal memory to maintain context over multiple turns and a dialog policy that governs the selection of system actions at each turn of the dialog based on this internal belief state (Young 2013). Finite-state machines (FSMs) are the simplest method of state-based DM that use an explicit graphical representation and deterministic transitions between dialog states. FSMs are common in interactive voice response (IVR) systems that have been used by banks and airlines for decades. Frame-based models (FBMs) improved the flexibility of FSMs by supporting user input

irrespective of order. These systems introduced tracking the dialog context using slots that fill a specific role within a given semantic frame. This facilitated mixed-initiative dialog where the user could preemptively provide the values for certain slots, while the system would prompt the user for the value of unfilled slots. Information state methods introduce additional complexity into the FBMs by tracking discourse elements, for example, the question under discussion or the agenda. Decision-making requires complex rules or machine learning models. Statistical dialog managers can use an explicit, latent, or hybrid representation of state. In each case, a mapping between an internal representation of the dialog state and dialog actions is learned from data often collected by Wizard-of-Oz (WOZ) methods and/or user simulation.

Natural Language Generation

Scripted responses are the simplest and most common form of NLG for health dialog systems included in this review. Template-based responses that confirm common ground by filling response slots using the state of the dialog manager are the second most common among accepted studies. This approach requires only a single template to be built for each dialog action rather than each potential response. These methods increase the scalability and task success of scripted solutions. Others have developed end-to-end statistical models for generating natural language from a continuous dialog state representation with no a priori template design (Serban 2015, Li 2016, Li 2017). However, production systems typically rely on retrieval-based selection of responses.

2.1.2 Methods

This review followed the Preferred Reporting Items for Systematic Review and Meta-Analyses guidelines (Moher 2009) (Figure 2.2). The review is registered with PROSPERO and available at: https://www.crd.york.ac.uk/PROSPERO/display_record.php?RecordID=104140.

Inclusion/Exclusion Criteria

Studies were included that demonstrated an application in health care or were health-related, interacted with users through natural language (text or speech), used state-based DM, and had either been user tested or deployed in a health care setting.

Studies were excluded that were not health care related, had no user testing or had not been implemented in practice, were operated by a human (WOZ), limited user interactions to buttons, or had stateless or no DM, for example, a simple voice command system.

Search Strategy

A search of PubMed and the ACM Digital Library was conducted on September 12, 2017 for articles related to healthcare and at least one of the following: dialog systems, conversational or relational agents, virtual or automated counselors, nurses, therapists, or patients. This strategy was arrived at through consultation with a health sciences research librarian experienced in conducting systematic reviews. For PubMed search, we included the term "User-Computer Interface" from the Medical Subject Headings thesaurus to increase the precision of our results while allowing for a broader keyword search. Queries were kept similar between both databases, and a complete table can be found in (Appendix A).

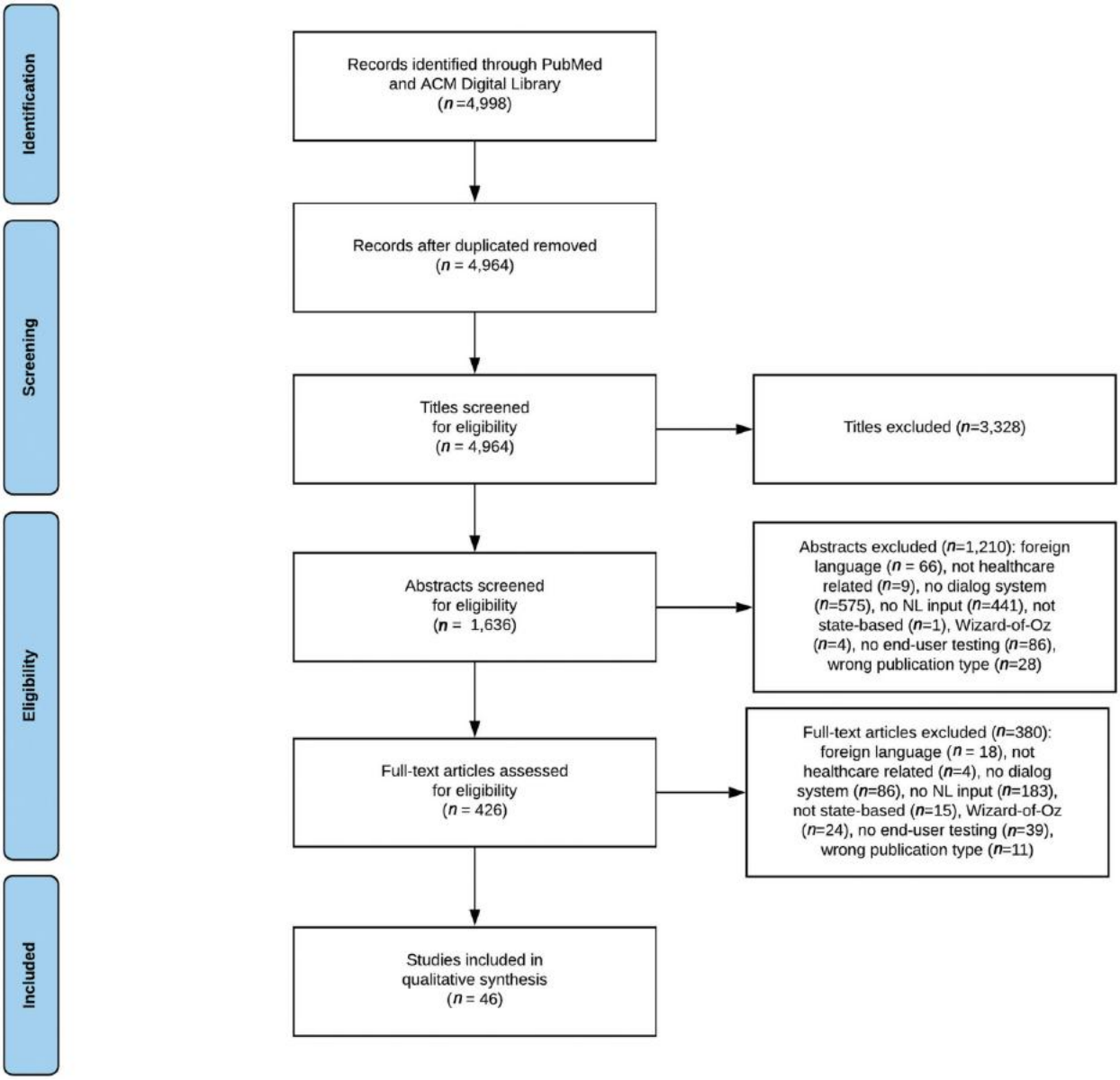


Figure 2.2 PRISMA diagram for the systematic review

Selection Process

Screening consisted of three parts: title screening, abstract screening, and full-text screening. The title screening served mostly to eliminate non-health-related articles and articles concerning other senses of the word agent, for example, agent-based modeling or pathogenic agent. Each abstract that passed the title screening was then

screened by two reviewers blinded to each other's results. Abstracts approved by at least one reviewer were accepted for full-text screening. The full-text of each accepted abstract was then read by two reviewers to determine whether it satisfied the inclusion/exclusion criteria. The final set of papers was reached by consensus between the reviewer pair. Data were then extracted from each full-text paper by the author (WK) using a table developed for this review.

Data Analysis

This review provides a narrative synthesis as the study design among retrieved papers was heterogeneous, which precluded the possibility of an aggregated statistical analysis. Instead, a set of fields was developed to capture details relevant to answering the aforementioned research questions along with additional meta-data. A convenience sample of the papers was selected to validate and improve upon these fields. The primary outcome of that process was to add a second level of classification for each domain. Once this process was complete, the full-text of each paper was reviewed, and the fields were extracted (Tables 2.1–2.5).

Table 2.1: Clinical Processing Systems

Title	Theme	Summary	Study Type	Outcomes
IVR and administrative operations in healthcare and hospitals. [126]	Clinical Processes	Overview of implemented IVR systems in a hospital	System Description	Not Reported
Ergonomic automated anesthesia recordkeeper using a mobile touch screen with voice navigation. [152]	Clinical Processes	IVR for clinical charting in anesthesiology	Quasi-Experimental	Adoption Rate
It's only a computer: the impact of human-agent interaction in clinical interviews [64]	Clinical Processes	A Virtual Human for conducting medical interviews	RCT	Fear of Negative Evaluation; Impression Management; System Usability; Self-deception
A history-taking system that uses continuous speech recognition. [85]	Clinical Processes	Dialog system that takes patient history prior to doctor visit	Quasi-Experimental	False positive finding rate; Utterance semantic accuracy; Word Recognition; Utterance Recognition
Telephone-linked care for physical activity: a qualitative evaluation of the use patterns of an information technology program for patients [51]	Clinical Processes	IVR to monitor and promote physical activity	Quasi-Experimental	Qualitative
Telephone follow-up in primary care: can interactive voice response calls work? [75]	Clinical Processes	An IVR system to follow-up with patients after ambulatory visits	RCT	Satisfaction

Table 2.1 (continued): Clinical Processing Systems

Title	Theme	Summary	Study Type	Outcomes
Closing the feedback loop: an interactive voice response system to provide follow-up and feedback in primary care settings [116]	Clinical Processes	An IVR to follow up with patients between primary care visits	Quasi-Experimental	Not Reported
Appraisal of a conversational artefact and its utility in remote patient monitoring [14]	Clinical Processes	An IVR system for reporting health metrics for patients with diabetes and co-existing hypertension	Quasi-Experimental	Calls completed, calls made, disconnections, modality
Using interactive voice response to improve disease management and compliance with acute coronary syndrome best practice guidelines: a randomized controlled trial. [160]	Clinical Processes	An IVR system to follow up with patients at regular intervals and ask questions related to medicine adherence and management.	RCT	Increased compliance and reduced adverse events
It's only a computer: virtual humans increase willingness to disclose [111]	Clinical Processes	A virtual therapist that conducts a clinical interview with patients	RCT	Willingness to disclose information
Interactive voice response telephony to promote smoking cessation in patients with heart disease: a pilot study. [145]	Clinical Processes	An IVR system to follow-up with patients who smoke that have recently been hospitalized with coronary heart disease.	Prospective Cohort Study	Abstinence rate at one year follow-up

Table 2.1 (continued): Clinical Processing Systems

Title	Theme	Summary	Study Type	Outcomes
Personal relationships with an intelligent interactive telephone health behavior advisor system: a multimethod study using surveys and ethnographic interviews. [90]	Clinical Processes	An IVR System to assist patients through information and advice to support patients with chronic diseases and their caregivers.	RCT	Patient Survey
Automated patient assessments after outpatient surgery using an interactive voice response system [58]	Clinical Processes	An IVR system for automated follow-up of women that underwent outpatient gynecologic surgery	Prospective Cohort Study	Adverse events, Follow-up Rate, Patient Survey
Automated conversation system before pediatric primary care visits: a randomized trial. [2]	Clinical Processes	An IVR system that screened and counseled caregivers prior to their child's pediatric visit.	RCT	Clinical effectiveness, Medication Adherence, Parent and Clinician satisfaction

Table 2.2: Medical Education Systems

Title	Theme	Summary	Study Type	Outcomes
Ellipsis and coreference resolution in a computerized virtual patient dialogue system. [106]	Medical Education	A VP for tutoring student to interact with Chinese patients	Dialog Analysis	Coreference and Ellipsis quantity in dialog. Precision, Recall, and F-measure for machine learning methods.
Virtual patients: assessment of synthesized versus recorded speech. [42]	Medical Education	A VP system was used to the impact of TTS fidelity	RCT	Intelligibility, Naturalness, Clarity
The use of virtual patients to teach medical students history taking and communication skills. [165]	Medical Education	A VP with bipolar diagnosis used to train medical students to address suicidal risk	Quasi-Experimental	Authenticity, stimulation to ask questions, acceptability, helpfulness of VI feedback
Medical student satisfaction using a virtual patient system to learn history-taking communication skills. [37]	Medical Education	A VP system for conducting medical interviews	Quasi-Experimental	Satisfaction
A pilot study to integrate an immersive virtual patient with a breast complaint and breast examination simulator into a surgery clerkship. [36]	Medical Education	A VP for practice history taking and breast exam	RCT	Confidence, anxiety

Table 2.2 (continued): Medical Education Systems

Title	Theme	Summary	Study Type	Outcomes
Are virtual patients effective to train diagnostic skills?: a study with bulimia nervosa virtual patients [67]	Medical Education	A VP for training students in diagnosing bulimia nervosa. The system allowed students to conduct a medical interview.	RCT	Test scores
Optimal learning in a virtual patient simulation of cranial nerve palsies: the interaction between social learning context and student aptitude. [86]	Medical Education	A VP for training medical students in diagnosing cranial nerve palsy	RCT	Understanding of the case
Virtual patient simulations and optimal social learning context: a replication of an aptitude-treatment interaction effect. [87]	Medical Education	A VP for learning about interacting with a patient with neurological problems	Quasi-Experimental	Understanding of the case
Virtual human personality masks: a human computation approach to modeling verbal personalities in virtual humans [96]	Medical Education	A VP that has an adjustable verbosity to emulate markers of depressed patients	RCT	Not Reported
Natural language understanding performance and use considerations in virtual medical encounters [170]	Medical Education	A Virtual Standardized Patient to train medical students in clinical interviewing for ear pain	Quasi-Experimental	NLU Accuracy Rate, Appropriate Response Rate, NLU Errors, Improper Questions, Commands

Table 2.2 (continued): Medical Education Systems

Title	Theme	Summary	Study Type	Outcomes
Informed consent procedures: an experimental test using a virtual character in a dialog system training application [77]	Medical Education	A VP for practicing administering informed consent	RCT	Not Reported
A tool for training primary health care medical students: the virtual simulated patient [112]	Medical Education	An ontology based ECA that maintains an internal emotional state to practice conversation in a primary care setting	Quasi-Experimental	NLU accuracy; Emotional accuracy
Shader lamps virtual patients: the physical manifestation of virtual patients [147]	Medical Education	A face was projected on a Styrofoam head as a manifestation of a VP with cranial nerve problems.	Quasi-Experimental	Not Reported
Let me introduce you to your first virtual patient [61]	Medical Education	A Comprehensive VP Platform developed by Shadow Health	Quasi-Experimental	Participant testimonies
Evaluation of Justina: a virtual patient with PTSD [92]	Medical Education	A VP that simulates PTSD for psychotherapy training	Quasi-Experimental	Tellegen Absorption Scale; Immersive tendencies questionnaire; Large number of these
A comparative analysis between experts and novices interacting with a virtual patient with PTSD. [93]	Medical Education	A VP exhibiting signs of PTSD and capable of answering questions using NL	Case Control	Information elicited from VP and other measures

Table 2.2 (continued): Medical Education Systems

Title	Theme	Summary	Study Type	Outcomes
Virtual patients as novel teaching tools in psychiatry. [110]	Medical Education	A VP that simulates PTSD for psychotherapy training	Quasi-Experimental	Not Reported
Virtual reality skills training for health care professionals in alcohol screening and brief intervention. [57]	Medical Education	A VP that used pre-recorded video responses to dynamically respond to questions based on rapport with the user.	RCT	Changes in Clinical Skills
Objective structured clinical interview training using a virtual human patient. [130]	Medical Education	A VP that embodies a case of an adolescent with conduct disorder.	Quasi-Experimental	User Satisfaction
Development and evaluation of web-based animated pedagogical agents for facilitating critical thinking in nursing [124]	Medical Education	An ECA that fulfills the role of a tutor for undergraduate nursing students over three scenarios.	Quasi-Experimental	Critical Thinking Process Test (CTPT) from Educational Resources Inc.

Table 2.3: Mental Health Systems

Title	Theme	Summary	Study Type	Outcomes
A demonstration of the perception system in SimSensei, a virtual human application for healthcare interviews [167]	Mental Health	A virtual therapist for PTSD	RCT	Not Reported
Evaluating an automated mental health care system: making meaning of human-computer interaction [50]	Mental Health	IVR to act as a human health professional for depression	Quasi-Experimental	Qualitative
On-demand virtual health counselor for delivering behavior-change health interventions [5]	Mental Health	An ECA that delivers MI for health behavior change for alcohol consumption	RCT	Attitude, Intention to Use, Perceived Enjoyment, Perceived Ease of Use, Perceived Sociability, Perceived Usefulness, Social Presence, Trust, Anxiety, Social Influence; Anthropomorphism
SimSensei kiosk: a virtual human interviewer for healthcare decision support [40]	Mental Health	A virtual therapist for PTSD	RCT	Not Reported
Toward robotic companions that enhance psychological wellbeing with smartphone technology [81]	Mental Health	A mobile app that engages the user in dialog about psychological wellbeing every day for three weeks	Quasi-Experimental	System accuracy, User Satisfaction

Table 2.4: Patient Education Systems

Title	Theme	Summary	Study Type	Outcomes
An assessment of the virtual conversations method for prostate cancer patient education. [68]	Patient Education	This system educates patients on Prostate Cancer. They interact with the system using a microphone, their questions are matched against a database of video clips of a recorded prostate cancer doctor. The system suggests follow-up questions given the system answer at the previous turn.	Quasi-Experimental	Knowledge Gain, User Satisfaction
Evaluation of a virtual dialogue method for breast cancer patient education. [69]	Patient Education	A spoken dialog system that responds to users using a library of video clips of clinicians	Case Control	Knowledge gain, Feasibility and Acceptance

Table 2.5: Personal Health Systems

Title	Theme	Summary	Study Type	Outcomes
vAssist: building the personal assistant for dependent people [153]	Personal Health	A SDS for interfacing with an assisted living hub	Quasi-Experimental	Not Reported
An adaptive computational model for personalized persuasion [89]	Personal Health	A virtual nurse that provides health advice and medication reminders for older adults	Quasi-Experimental	Rate of persuasion, social presence, frustration; Telepresence, Social Presence, Interactivity, Frustration
Multimodal and mobile conversational health and fitness companions [176]	Personal Health	Smart Speaker and Mobile agent that communicates with the user about various aspects of their day with an emphasis on healthy eating and exercise	Quasi-Experimental	WER and CER; Task Completion Rate
Multi-agent patient representation in primary care [144]	Personal Health	A CA schedules visits on behalf of the patient through SMS or email with a CA on behalf of the hospital.	Quasi-Experimental	Not Reported
Analysing user's reactions in advice-giving dialogues with a socially intelligent ECA. [128]	Personal Health	An ECA that attempts to persuade users to eat healthy through social dialog.	Quasi-Experimental	Percentage of Social Moves, Dialog Duration, Move length, Number of Questions, User Reactions to Persuasion Attempts

2.1.3 Results

The following two subsections provide our results that answer the research questions Q1 and Q2, respectively.

Application Domains

We identified five application domains described in detail below: clinical processes, medical education, mental health, patient education, and personal health.

Clinical Processes

Fourteen studies focused on automating clinical processes including eight systems for patient engagement and monitoring (Kaplan 2003, Black 2005, Farzanfar 2005, Reid 2007, Forster 2008, Willig 2013, Houser 2013, Sherrard 2015), four systems for history taking and screening (Johnson 1992, Gratch 2014, Lucas 2014, Adams 2014), one system for clinical documentation (Sanjo 1999), and one overview of implementations within a hospital (Mouza 2003). Patient engagement and monitoring have become increasingly important as hospital systems shift toward value-based reimbursement models (Fox 2017). Scalability is a strong motivating factor for the use of automated systems to fulfill this role. Each study in this category used IVR to follow-up with patients via phone calls at regular intervals in a structured, system-initiated dialog. This included three randomized control trials (RCTs), two studies had positive results finding a 60% increase in medication adherence (n = 1,608) (Sherrard 2015) and 90% patient satisfaction (n = 474) (Houser 2013) for the intervention groups. On the other hand, one manuscript described the limitations of their current IVR implementation and suggested improvements that they identified through surveys and user interviews (Kaplan 2003).

The earliest system for medical history taking was an IVR system with hand-crafted grammars for NLU and template-based NLG (Johnson 1992). The system was evaluated using system performance metrics related

to the Automatic Speech Recognition and NLU components including the word-level accuracy and semantic accuracy of the utterance parser. The authors reported that the false-positive rate was low enough to test with real patients; however, developing the knowledge base, text-generation templates, and grammars were prohibitively time-consuming for widespread adoption at the time.

SimSensei is an ECA that mimics a virtual therapist to conduct a clinical interview of patients with mental health concerns (Gratch 2014, Lucas 2014). An RCT of this system (n = 239) found that participants were more willing to disclose information when they believed the system was fully automated than when they believed it was operated by a human.

Voice-oriented-command-automated-anesthesia-record-keeper was developed as a clinical documentation system for anesthesiologists to enter data while maintaining visual contact with their patients (Sanjo 1999). The authors report adoption rate as their main evaluation criteria and report that 100% clinical usage was reached after 8 weeks. However, their sample size is unclear and additional factors that may have influenced this adoption rate are not reported.

Medical Education

We identified 20 studies of 12 unique virtual patient (VP) systems (Dickerson 2006, Hubal 2006, Stevens 2006, Louie 2007, Deladisma 2008, Kenny 2008, Lopez 2008, Parsons 2008, Fleming 2009, Deladisma 2009, Kenny 2009, Krishnan 2012, Rivera-Gutierrez 2012, Gutierrez-Maldonado 2013, Johnson 2013, Johnson 2014, Friedman 2014, Lin 2016, Talbot 2016) and 1 virtual tutor (Morey 2012) for inclusion in this review. VPs are dialog systems used in medical education that mimic the symptoms and demeanor of real patients enabling students to try out different approaches within a simulated learning environment. VPs included in this study cover several diagnoses including abdominal pain (Dickerson 2006, Stevens 2006, Deladisma 2008), breast cancer (Deladisma 2009), bulimia nervosa (Gutierrez-Maldonado 2013), depression (Krishnan 2012), alcohol

screening (Fleming 2009), conduct disorders (Parsons 2008), ear pain (Morey 2012), posttraumatic stress disorder (Louie 2007, Kenny 2008, Kenny 2009), cranial nerve palsy and other ophthalmic conditions (Rivera-Gutierrez 2013, Johnson 2013, Johnson 2014). While these systems are primarily concerned with training students in clinical interviewing skills, one system focused on training students in informed consent procedures (Hubal 2006). These systems were primarily evaluated through qualitative content analysis and by their ability to improve examination scores across student cohorts.

Mental Health

We identified five systems within the domain of mental health (Fanzafar 2007, Amini 2013, Devault 2014, Stratou 2015, Jeong 2017). Three of these studies were RCTs covering two ECA systems. The SimSensei system discussed in a prior section was used to deliver PTSD therapy in two studies (Devault 2014, Stratou 2015). Their study design consisted of three groups (n = 351) who interacted with either a human counselor, a system controlled by a wizard, or a fully autonomous system. They reported that participants reported higher rapport with the WOZ or ECA than face-to-face interaction, but the fully autonomous system did not reach the same level of rapport as the human-controlled avatar. The other system was an ECA that used a technique called motivational interviewing for health behavior change of alcohol consumption (Amini 2013). The investigators assigned a group of university student participants (n = 51) to interact with the ECA in either an empathic or non-empathic condition and found that participants were more likely to interact with the empathic system which also scored higher on their Likert-scale measures for attitude, intention to use, perceived enjoyment, perceived ease of use, perceived sociability, perceived usefulness, social presence, trust, social influence, and anthropomorphism.

Patient Education

Two studies focused on patient education using the same underlying DM system (Harless 2007, Harless 2009). This system used a unique approach to NLG that consisted of prerecorded video clips of doctors answering questions. The system matched spoken utterances against this database and suggested follow-up questions that appeared beneath each video clip. Both studies were similarly structured with a pretest, a dialog session, and a post-test with survey. The tests were meant to evaluate the knowledge gain of prostate cancer (n = 33) and breast cancer (n = 70) and found an improvement of 31 and 28 percentage points to their test scores, respectively.

Personal Health

Five systems focused on personal health management. One system was implemented in clinical practice to operate as a scheduling system (Reed 2005). Two systems were designed to communicate information about healthy eating with patients (Turunen 2011, Novielli 2012) and were evaluated using standard metrics for dialog systems. The evaluation of the former consisted of component and system performance metrics including word error rate and task completion rate, while the evaluation of the latter was evaluated based on percentage of social moves, dialog duration, move length, number of questions, and user reactions to persuasion attempts.

Two systems focused on assisted living solutions for older adults (Sansen 2014, Kang 2015). vAssist was a digital assistant designed to control and monitor Internet of things devices, functioning as a smart home hub (Sansen 2014). This technology had the potential to increase the time older adults are able to live independently. Florence was a system that simulated a virtual nurse encouraging study participants to exercise and eat healthily (Kang 2015). The system made use of the elaboration likelihood model of persuasion and presented a Model of Adaptive Persuasion (MAP) as a method for encouraging behavior change. They found in a study of 26 subjects that use of MAP persuaded 65% of participants whereas the baseline system using a single strategy only persuaded 35%.

Evaluation Methods

Forty-six studies met the eligibility criteria including 24 quasi-experimental studies, 16 RCTs, 2 case-control studies, 2 prospective cohort studies, 1 system description, and 1 human-computer conversation analysis.

Reproducibility

We applied a relaxed definition of reproducibility requiring that the system be accessible through any of the following methods: commercially available, open-source code, executable binary, or hosted on the Internet. By this definition, three studies were reproducible representing two systems, the NERVE VP (Johnson 2013, Johnson 2014) and ShadowHealth VP (Friedman 2014). A demonstration of the NERVE VP is available as a Web application and the ShadowHealth VP is commercially available as a standalone application.

2.1.4 Discussion

This review covered 46 studies spanning 5 application domains extending prior reviews that focused on conversational agents (n = 17)(Laranjo 2018) and ECAs (n = 49)(Montenegro 2019). The latter covers additional systems that did not fit our requirement of an NLU component; however, the scope of these systems was confined to the mental health domain. These earlier reviews reported no standardized evaluation methods among studies. However, our review adds additional nuance to this observation, that is, while no clinically validated standards were applied across studies, many used standard evaluations for dialog system evaluation, for example, task completion, user satisfaction, and NLU accuracy. Prior work has demonstrated that traditional thresholds and interpretations of these metrics may not be enough to ensure patient safety (Bickmore 2018, Fitzpatrick 2017) and additional research is necessary to develop ethical frameworks for conversational agents

that communicate sensitive health information and provide services to vulnerable populations particularly those with cognitive impairment.

Of the studies covered in this review, 34 spoken dialog systems were identified with many reporting word error rates around 25 to 30%. Unsurprisingly, many authors reported speech-to-text accuracy alongside text-to-speech fidelity as primary detractors from user satisfaction and barriers to practical implementation (Farzanfar 2005, Deladisma 2008, Turunen 2011). RCTs made up 34.8% of all studies covered in this review and the distribution was not uniform across all application domains. Dialog systems in the mental health domain had the highest percentage of RCTs at 60%, compared with 43% of studies of clinical processes, 35% of medical education studies, and 0% for both the personal health and patient education domains.

Of all studies, only the NERVE studies were reproducible with freely accessible software as an interactive demonstration was available over the Web at the time this manuscript was written (Johnson 2014). This reproducibility crisis obstructs the ability to independently assess and compare dialog systems in health care, thus delaying their implementation into clinical practice. This study confirms the finding that studies of dialog systems in health care lack clinically validated metrics for evaluation further hampering the ability to compare results (Laranjo 2018). While some studies had links to demonstrations of the applications, these links were no longer accessible at the time of this review. As hosting a VP may be computationally intensive, we would advise against using a hosted demonstration as the primary means of reproducibility and instead suggest providing a containerized application and/or source code. Despite these limitations, a review of this literature provides insight into the potential applications and motivation for health dialog systems. General domain speech recognition is nearing human-level performance (Chiu 2017) and simulated voices are increasing to near human-level fidelity (Simonyan 2016). These factors combined with the availability of open-source and commercial NLU platforms have the potential to democratize the application of this technology.

2.1.5 Limitations

This review excluded dialog systems that restricted user input to buttons or candidate responses as the aim was applicability toward current methods that support speech or free-text input. Regardless, we believe that the application domains identified in this study are representative of the broader literature as the same themes were present during the abstract and full-text review processes during which these studies were excluded.

2.1.6 Conclusion

Hospital systems can improve outcomes and efficiency by using dialog systems for screening, clinical documentation, and patient education. Medical schools can use VPs to increase student access to training prior to interacting with real patients. Conversational agents may reduce barriers to accessing mental health services and affect behavior change through continuous follow-up. However, progress is hampered by a lack of reproducibility and standard for health dialog system evaluation that accounts for clinical metrics.

2.1.7 Follow-up

There has been research published relevant to the present work that was not captured due to the time interval in which the papers were queried. Wysa is a therapy chatbot that was developed to listen emphatically and users who interacted with the system regularly had a higher reduction in PHQ-9 scores than those with low usage (Inkster 2018). Recent work across both Wysa and another CBT chatbot, Woebot, have shown that users of this system achieve a therapeutic alliance comparable to human-delivered forms of CBT (Prochaska 2021, Beatty 2022). Woebot has also been shown to reduce depression and anxiety symptoms in college students (Fitzpatrick 2017) and has received the FDA breakthrough device designation to treat postpartum depression (Darcy 2022). Wysa has also received FDA breakthrough device designation for use with patients with comorbid diagnosis of musculoskeletal pain and depression or anxiety following a study showing its effectiveness for helping to manage

chronic pain and associated depression and anxiety (Leo 2022). As these are closed commercial systems, no information is available on how these systems work.

2.2 Natural Language Processing

Having covered the need for the contribution within the space of health dialog systems, I will describe in this section a background of the natural language processing techniques required to understand the methodological contributions in the following chapters. Training neural models on natural language requires that free-form text be converted into a vector/matrix representation understandable by computers, referred to as an embedding. There are a number of methods to embed text which will be covered in this section. Most contemporary methods achieve this by utilizing the hidden state of neural language models. Once in vector/matrix format, neural models apply matrix multiplication between embeddings and matrices of trainable weights, which are tuned through stochastic gradient descent (Bottou 2010), to approximate an objective function which could be to predict a label, generate a response, or any number of tasks.

The remainder of this section begins with a description of distributional semantics, a key theory to understanding modern computational linguistics, followed by an introduction to language modeling. These set the stage to provide an explanation of state-of-the-art approaches to language modeling that have led to a large increase in model performance over the past few years. Lastly, I provide an overview of the commonsense reasoning models that build on these techniques and are extended in this work.

2.2.1 Distributional Semantic Representations

Distributional semantics is based on the hypothesis that the meaning of words can be defined by the contexts in which they occur such that words that occur in similar contexts have similar meaning (Firth 1957). The distributed representation of natural language has a long history in computational linguistics, beginning with

maximum entropy methods (Berger 1996), latent semantic analysis (LSA) (Dumais 1988), probabilistic latent semantic indexing (LSI) (Hofmann 1999), and latent Dirichlet allocation (LDA) (Blei 2003). Continuous-space vector representations of natural language have been broadly applied as a method of improving supervised model performance by learning word-level features from large unlabeled datasets (Weston 2008, Turian 2010, Mikolov 2013, Pennington 2014). However, these word-level methods generate a single 'static' embedding for each word from a corpus of text. Models at the time used these 'global' embeddings that were not modified by the 'local' contextual information of the sentence and thus relied on the task-specific model to handle sub-tasks like word-sense disambiguation. To address this issue, the field has moved toward models that incorporate contextual embeddings that compute a semantic representation of each token within the sequence in which it occurs.

2.2.1 Neural Language Models

Language models compute the probability of the next token in a sequence, x_t, given the prior context of tokens:

$$p(x_1, \ldots, x_T) = \prod_{t=1}^{T} p(x_t | x_{1:t-1})$$

Equation 2.1 Language modeling objective function

Neural network approaches for language modeling have existed for many years (Bengio 2003). However, due to limitations of contemporaneous computational resources and the architecture of early neural models they were limited in the amount of data they could incorporate. As will be shown shortly, incorporating more data has led to advances in neural methods but also necessitated a shift in underlying architectures.

Long Short-Term Memory (LSTM) neural networks (Hochreiter 1997) were the primary neural model for natural language processing tasks until the introduction of the transformer model covered in the next section. Bi-directional LSTMs were used to approximate the forward and backward language modeling objectives, that is the reverse of (Equation 2.1). After training neural models, the language model decoder can be used to generate natural language given an input sequence or removed and the hidden state representation of the model can be used as the embedding of the sequence of text. This embedding can then be used downstream by other task-specific layers. Early exploration of the multi-layer models found that different information was stored in each layer, e.g. syntactic or semantic information, leading to a model that computed a weighted embedding based on the task (Peters 2015). This discovery of additional information, beyond that directly related to the language model objective, latently encoded in the language model weights continued to be observed with increasingly complex model behaviors relying on less task-specific training (Liu 2019). This is covered in the next subsection.

Attention, a key component of recent advances in NLP, was first introduced for neural networks in LSTMs (Graves 2014), and quickly improved performance on machine translation (Bahdanau 2014), speech recognition (Chorowski 2015), question answering (Hermann 2015), and dialog systems within memory neural networks (Weston 2014, Sukhbaatar 2015).

2.2.2 Transformer-Based Models

The introduction of transformer models that could be trained more efficiently on web-scale data through an attention-only mechanism (Vaswani 2017) led to the proliferation of large pre-trained language models (PLMs), sometimes referred to as foundation models, that have significantly increased performance across a variety of natural language processing tasks in recent years (Duan 2020). By training on large web-scale data sources, these models have displayed emergent capabilities beyond the language modeling tasks for which they have been

trained. This allows them to be applied across a variety of tasks with limited in-domain data. In the remainder of this section, I will describe the foundation models used in this work.

Bidirectional Encoder Representations from Transformers (BERT) was the first method to introduce transformer models to natural language processing (Devlin 2018). This method substitutes two new training objectives in place of classical language models, masked language modeling and next sentence prediction, which accommodate the simultaneous attention mechanism of transformers. BERT replaces the forward and backward LSTM layers with a single transformer that simultaneously computes attention in the forward and backward direction. Since these are computed simultaneously, the authors introduce the masked language modeling objective that replaces certain words with a [MASK] label. The model is then tasked with predicting the masked words.

Rather than applying attention in both directions simultaneously, Generative Pre-trained Transformer (GPT) applies transformers as an auto-regressive unidirectional model to the task of language modeling in the forward direction using the standard objective in Equation 2.1 (Radford 2018). When trained on web-scale text, these language models were found to be capable of learning multiple tasks with state-of-the-art performance without supervision (Radford 2019). Since these models were trained without supervision, capabilities that are learned inherently in the language modeling tasks have been referred to as emergent behaviors. Recent work has focused on determining what behaviors are captured in these models through a technique called prompting, which uses natural language instructions to the model, which can be few-shot, where a few examples are provided, or zero-shot, where only instructions and no examples are provided (Liu 2022, Wei 2022).

Bidirectional Auto-Regressive Transformer (BART) (Lewis 2019) generalizes the BERT encoder model by adding noising functions, e.g., token masking or deletion, and adds a GPT decoder. The BERT encoder takes as input the sequence of tokens and converts it into an embedding that is passed to the GPT decoder, which converts this embedding back into the original sequence.

Adapting language models to new domains has been primarily achieved through fine-tuning the language models on corpora within the domain of interest by continuing training using the original training objective (Gururangan 2020). MentalBERT is one such model in the mental health domain and has been trained on a corpus of mental health forum posts and various publicly available mental health datasets (Ji 2022). By pre-training the model on these data, the model showed increased performance on detection of suicidal ideation, stress, and depression.

GPT-3 is the largest iteration of the GPT-line of models (175B parameters). Researchers have found the output generated by large PLMs can be toxic (Bender 2021). As a step toward generating less toxic and more accurate content, the GPT-3 model was improved through a novel human-in-the-loop reinforcement learning method that trained a scoring model to match human-judgment of the relative ranking of candidate responses based on natural language instructions as input, which was used to fine-tune the PLM (Ouyang 2022). This method also helped to distill the total number of parameters down to 1.3B parameters, which can be processed efficiently by contemporary hardware. This model is referred to as Instruct-GPT and will be evaluated in this work on the task of emotion recognition alongside the COMET model described in the next section.

The TED Policy is a dialogue management model that uses a transformer architecture to predict the next action a conversational agent should take given the input text and state of the dialog (Vlasov 2020). The model uses the StarSpace method of embedding the utterances and responses into a shared embedding space and then ranking the responses based on a similarity function (Wu 2018). In the case of the TED policy, it embeds the action label, for example “greet” for a response “Hi, how are you?”, and learns to match this to the embedding “Hi” with intent “greet” by embedding the context of “Hi” and “greet” using a transformer layer, projecting that embedding into the shared embedding space and comparing that to the embeddings for all available action labels, e.g. “greet” or “ask_name”. This model is used in the current work due to the use of the

Rasa open-source framework[1], which uses the TED Policy, for other aspects of the conversational AI system development.

2.2.3 Commonsense Reasoning from Transformers

ATOMIC is a knowledge graph of commonsense reasoning that consists of 877k tuples contributed by crowd workers. Each tuple describes inferential knowledge from events including predictions of agent mental states (Sap 2019). Given an event with an agent (subject) and grammatical patient (target), ATOMIC represents 4 main types of inference: *causes* for the agent, *attributes* of the agent, *effects* on the agent, and *effects* on the patient with subtypes relative to before or after the event took place.

My work extends COMET, a language modeling approach that generates commonsense triples that are not explicitly modeled in ATOMIC from events at the sentence-level (Bosselut 2019). While fine-tuned on data in a bespoke format ("*PersonX* fights with *PersonY*"), this method is able to support input of arbitrary utterances without needing to use a special format (beyond the addition of prompt tokens to control what inferences are made, e.g. "My heart is racing [xReact] [GEN]" (Figure 2.3). This model is selected due to its ability to predict a wider range of emotions through generation and its ability to infer emotional state information beyond surface-level lexical markers due to its recognition of event semantics. This work explores a few questions related to this model: 1) How well can COMET predict self-reported emotional states? 2) To what extent do large PLMs capture this information as an emergent behavior without fine-tuning? 3) What effect if any does the addition of this information have on empathic response prediction?

[1] https://github.com/RasaHQ/rasa

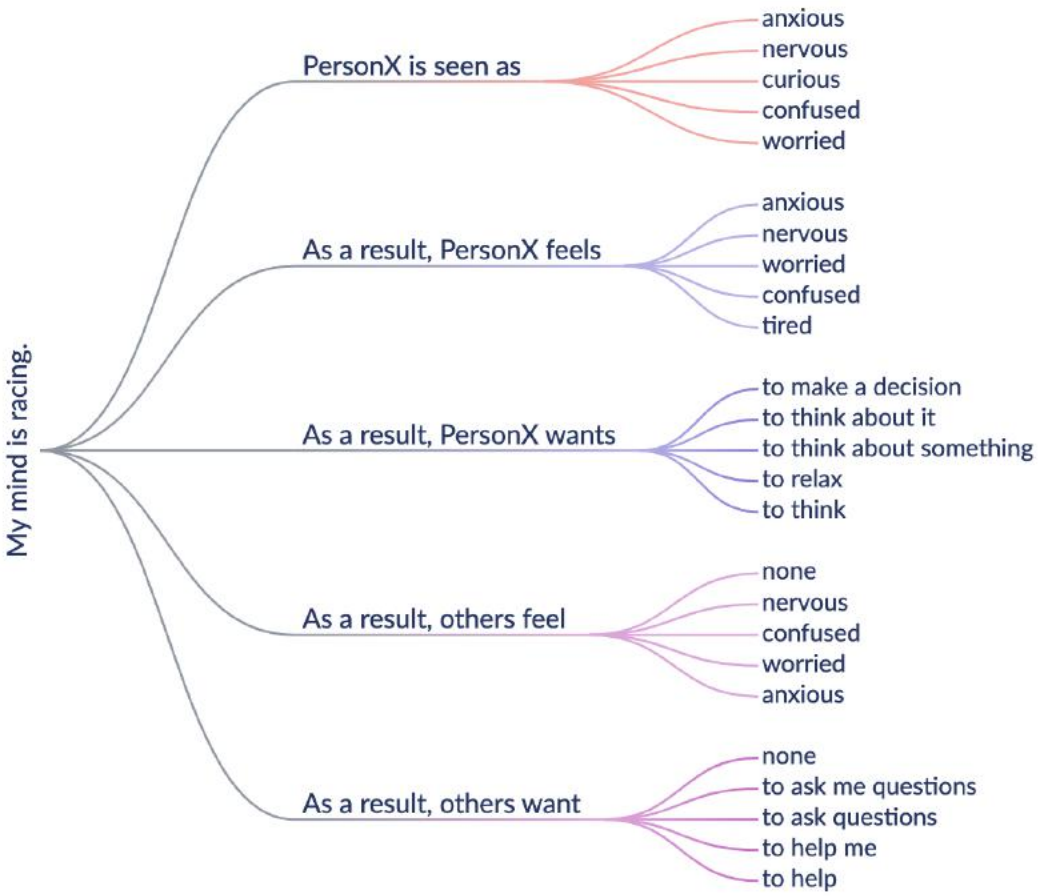


Figure 2.3: COMET mental state predictions for the patient utterance "My mind is racing."[2] As can be seen, the system is able to handle this idiomatic expression by recognizing it as a stressful situation.

2.3 Computational Empathy

In popular culture, machines are often represented as emotionless automata that while possessing super-human intelligence are incapable of recognizing human emotion. However, in reality, research into emotion recognition from facial expressions, speech, and most recently text has shown promising results (Poria 2017). This section begins with an overview of emotion theories, currently available open-source datasets, and methods that have been applied to the task of recognizing emotion from text and conversation.

[2] Generated from the publicly available deployment at https://mosaickg.apps.allenai.org/

2.3.1 Emotion Theory

There are several competing models of human emotions including Basic Emotions (Ekman 1971), the Plutchik wheel (Plutchik 1980), the circumplex model that focuses on two-dimensions valence and arousal (Posner 2005), and most recently the Cowen and Keltner model (Cowen & Keltner 2017). This section will cover only the emotion theories required to understand the work in future chapters and/or the relevant literature in this chapter. I surface concerns about these emotion categorizations that while they have continued to increase in granularity, they are still far coarser in granularity than the range of emotions that people describe through language. I also raise concerns over contemporary emotion recognition datasets that are not representative of mental health dialog and are annotated by third-parties using annotation schema derived from these emotion categorizations.

Ekman Emotion Categories

By far the most commonly-used, the Ekman emotion categories are derived from experiments on facial recognition. The Ekman emotion categories provide a proposed universal set of core emotions (*anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*) that were validated by experiments that presented participants with the set of six emotion categories and asked them to assign each category to an exaggerated facial expression from an actor selected by Ekman to be representative of that emotion category (Ekman 1971). Since participants in replicated experiments spanning different cultures, Ekman claimed to have developed a universal theory of basic emotion categories. However, others have argued that there were methodological issues with the Ekman experiments due to the requirement to select from the set of six categories, the exaggerated nature of the facial expressions chosen by Ekman, and evidence (including recreations of the experiment modifying the parameters

concerned above) indicating that facial expressions of emotions are dependent on cultural and situational contexts and vary even within a single individual (Barret 2019).

Cowen and Keltner emotion categories

Using newer methods, Cowen and Keltner derived 27 emotion categories from a study of participants' emotional responses to 2,185 emotionally evocative short videos (Cowen & Keltner 2017). The authors indicated in a response to a critique (Barret 2017) of this study that their results indicate that emotions can be represented in high dimensional space with gradients between what have been traditionally categorized as discrete emotions (Cowen 2018). This admission is at the core of issues that I alluded to at the start of this section and raise in the next subsection, concerning how the researchers and practitioners who developed all existing emotion recognition datasets have applied a limited theory of discrete emotion categories by assigning single labels of basic emotion categories. Instead, in this work participants are invited to provide their emotional state in one or two words, increasing granularity and supporting compositionality of emotion categories.

Wilcox Feeling Wheel

The feeling wheel (Figure 2.4) was created as a way for people to learn to recognize and express their emotions with more granularity (Wilcox 1982). The process of using the feeling wheel is to start with the inner emotions the client is feeling and then expand outward to the adjacent secondary ring of emotions and then to the third row. For example, a client may feel "scared", "insecure", and "foolish". Wilcox states that the wheel is necessary due to the proclivity of clients to describe their emotional states as "ok", "good", "bad", "better", or "worse". To date, no emotion recognition datasets have used the emotion categories within the feeling wheel beyond the inner circle (which roughly correspond to the Ekman basic emotions). An attempt was made, but was unsuccessful due to low inter-annotator agreement; instead, only the inner six emotion categories were used

(Zahiri 2017). A variation of this wheel is used in my study as a way to diversify emotional expression to obtain more granular ground truth labels for emotional states.

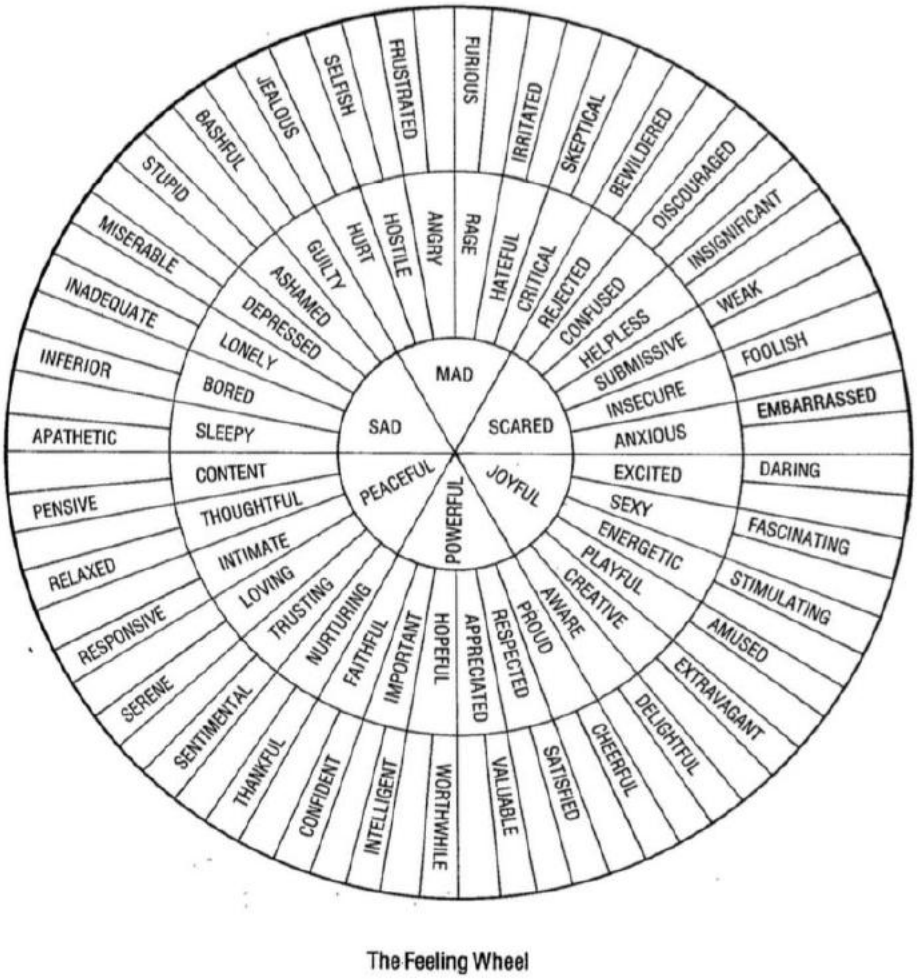


Figure 2.4 Feeling wheel of emotions (Wilcox 1982)

2.3.2 Emotion Recognition from Text

Research on text-based Emotion Recognition from Conversation (ERC) has primarily focused on the EmoryNLP, MELD, DailyDialog, and IEMOCAP datasets. Similar to the critique of the Ekman emotion studies (that actors portray stereotypical representations of emotions that are not representative of how people

express their emotional experiences in reality), research on emotion detection from text has relied almost entirely on actors and/or data labeled by third parties, both of which may be biased toward emotion category stereotypes and not capture the full-range of emotional expression in the lives of family caregivers. This indicates a gap in the current literature that my work attempts to address by exploring the ability of emotion detection systems to predict *self-reported* emotional states in relation to events described through a journaling exercise.

Open Source Datasets

The DailyDialog dataset (Li 2017) is a dataset derived from posts from English learning websites using web crawling. DailyDialog was annotated with the Ekman emotion categories, i.e. *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise*, and a default emotion of "other". The dataset was annotated by three experts and inter-annotator reliability on the emotion categories was not reported. The dataset was skewed with a majority label "other" (83.10%); with this label removed the majority label becomes "happiness" (74.02%).

The EmoryNLP dataset (Zahiri 2017) is based on transcripts of *Friends*, a television show, and uses the inner emotion categories of the feeling wheel, i.e. *joyful*, *mad*, *peaceful*, *powerful*, *sad*, *scared*, and a default of *neutral* (Wilcox 1982). The annotation was completed through Mechanical Turk, and the inter-annotator agreement (IAA) was very low (Cohen's Kappa, κ=0.14). Additionally, as may be expected from a situational comedy the dataset is heavily skewed toward the *neutral* (29.95%) and *joyful* (21.85%) categories. The authors indicate that they attempted to have the television show transcripts annotated with the secondary emotions in the feeling wheel, however that resulted in an even lower IAA with a κ-value of 0.08.

The MELD dataset (Poria 2019) is also based on transcripts of *Friends*. However, this set is labeled with the Ekman emotion categories, i.e. *joy*, *sadness*, *fear*, *anger*, *surprise*, *disgust*, and a default of *neutral*. The developers of MELD through the addition of visual information extended the EmotionLines dataset (Chen 2018). EmotionLines annotators labeled the Friends TV transcripts with Ekman emotion categories using

Mechanical Turk with inter-annotator reliability that was low (κ=0.34), albeit considerably better than that reported for the EmoryNLP set. Since the MELD dataset included multi-modal information (video and text) and the original EmotionLines annotation process only involved showing the text (transcribed audio) to the annotators, the MELD developers coordinated the re-annotation of the data by showing the Mechanical Turk workers the video clips as well as their transcripts. This resulted in higher inter-annotator reliability (κ=0.43). As with the previously described dataset, the MELD dataset is also unbalanced with label prevalence of *neutral* (47%), *joy* (17%), *anger* (12%), and *surprise* (12%). Despite the source of the dataset, which will likely lead to culturally biased algorithms (the Friends cast lacks diversity and the show was focused on the lives of young urbanites), the authors make the following claim: "this dataset is useful to train a conversational emotion recognition classifier which can be plugged into **any** dialogue system to generate empathetic responses" (emphasis added).

The IEMOCAP dataset (Busso 2008) is a multimodal dataset of video, audio, and text of actors reading scripts under direction. Data are labeled with Ekman emotion categories, i.e. *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*, as well as additional states of *excitement, frustration*,, and *neutral*. The distribution of labels in the dataset is not presented, however, they are likely more varied than the others above given they were controlled by the researchers through the direction of actors.

The dataset most relevant to the current work is the GoEmotions dataset (Demszky 2020). It consists of 58k Reddit comments annotated with 27 emotion categories from (Cowen 2017), plus a *neutral* category for comments that did not convey any emotion. The authors indicate a motivation of their work was to move beyond the Ekman emotion categories to more granular categories. The authors trained a BERT-large model with a classifier head that achieves an .46 macro-F1 on the emotion recognition task. They purport that their large dataset of granular emotion categories will be useful for downstream applications in conversational agents (a claim that I ultimately tested and found lacked support in the current context of developing empathic

conversational agents for mental health). The macro-average kappa value (averaging IAA equally across the 27 categories) was 0.293 indicating low agreement except for a few classes having a kappa value above 0.4. Those were *gratitude* (κ=0.749), *love*(κ=0.555), *amusement* (κ=0.474), and *admiration*(κ=0.468), which the authors identified as having more lexical correlates (e.g. "lol" for *amusement* and "thanks" for *gratitude*) than the categories with lower IAA. This underscores my critique of these tasks: that the annotators are picking up on surface level cues that bias the dataset and lead to unrealistic reporting of the performance of emotion recognition systems. This can lead to propagation of bias to downstream systems, such as conversational agents.

To summarize, there is a trade-off between increased granularity in emotion labels and IAA, evident in the drop in reliability score when moving from the Ekman categories (EmotionLines, MELD) through the inner to the outer rim of Wilcox's wheel (the EmoryNLP variants). This has led to datasets that rely on a dated and controversial theory of basic emotions since the limited size of the label set helps to increase IAA. A further concern is that third-party perception of emotion will vary from person to person and so may not accurately reflect the internal mental state of the author of a text message, especially when annotators and the author lack a shared cultural background. To address these concerns, my work takes a different approach by asking individuals to self-report their own emotional states. While this approach does not solve the variance between how individuals would assign a feeling to an event, it gives a ground truth emotional state label for that individual on that day in relation to a described event with the granularity of the emotional state limited only by the participant's range of emotional expression. In the present work, this gives a more accurate measure of the performance of text-based ERC methods and creates the opportunity in future work to personalize predictions to the individual by taking into account self-report data across time.

Additionally, the messages in existing ERC datasets are biased toward specific types of conversations that may not generalize to other scenarios, and so there exists a need to validate these models in critical scenarios such as text-based teletherapy. Specifically, they are from either a television show about urban-dwelling white

young adults in the 90s, basic conversations as part of learning English as a second language, or paid actors. Finally, the majority label in these datasets is "other" or "neutral" which as we will see in the next section, leads to models that largely predict these labels with limited utility in downstream applications.

State-of-the-Art Models

Contemporaneous work has shown that the inclusion of COMET output improves performance on emotion detection on the MELD, DailyDialog, EmoryNLP, and IEMOCAP datasets (Ghosal 2020, Zhu 2021). Table 2.6 presents the results reported in the COSMIC (Ghosal 2020) and TODKAT (Zhu 2021) papers. Ghosal et al. introduced the COSMIC model which assigns COMET inferences based on relation to the speaker (as the information is known from the dataset and all utterances are taken to have the speaker as the agent), e.g. *xReact* for the speaker or *oReact* for the listener. COSMIC improved performance over a RoBERTa based model on IEMOCAP (+Δ0.005 weighted-average), DailyDialog (Δ+0.014 MacroF1), MELD (Δ+0.016 weighted-average), and EmoryNLP (Δ+0.007 weighted-average). Zhu et al. introduced the TODKAT model that combined topical information (e.g. office or family) alongside COMET commonsense inferences and which improved performance over the COSMIC model on DailyDialog (Δ+0.015 Macro-F1), MELD (Δ+0.030 weighted-average), and EmoryNLP (Δ+0.050 weighted-average) but not IEMOCAP (Δ-0.040 weighted-average).

Table 2.6: Top performing models compared to a RoBERTa-based model without fine-tuning on ATOMIC as reported in the original papers

Models	DailyDialog (Macro-F1)	MELD (Weighted-F1)	IEMOCAP (Weighted-F1)	EmoryNLP (Weighted-F1)
RoBERTa Dialogue RNN (Ghosal 2020)	0.4965	0.6361	0.6476	0.3744
COSMIC (Ghosal 2020)	0.5105	0.6521	0.6528	0.3811
TODKAT (Zhu 2021)	0.5256	0.6823	0.6133	0.4312

These models generally provide an improvement over the RoBERTa baseline, however, the open-source ERC datasets tested do not address the health domain, and so the utility of commonsense inference remains untested for health-related ERC.

2.3.3 Empathic Response Prediction

Empathic response generation is a new field of research that began with the introduction of the EmpatheticDialogues dataset with the intention to improve the ability of conversational agents to respond emphatically in social dialogue (Rashkin 2018). The dataset is a collection of 25k peer-to-peer dialogs centered around a set of 32 emotion categories. The speaker is given an emotion label and asked to provide a situation in which they felt that emotion. They then share that situation with another peer in a two turn dialogue (Example 2.1).

Label: Afraid

Situation: Speaker felt this when...

"I've been hearing noises around the house at night"

Conversation:

Speaker: I've been hearing some strange noises around the house at night.

Listener: oh no! That's scary! What do you think it is?

Speaker: I don't know, that's what's making me anxious.

Listener: I'm sorry to hear that. I wish I could help you figure it out

Example 2.1: Example from EmpatheticDialogues training set (Rashkin 2018)

Modeling Approaches

Several systems have developed to generate empathic responses in social dialog (Lin 2019, Majumder 2020) and more recently peer-to-peer mental health support dialog (Sharma 2021). Recent work has reported positive improvements to empathy, as determined by empathic response retrieval metrics and human evaluation, by incorporating emotion recognition into dialog systems (Zhou 2018, Wang 2021).

Transformer-based models for empathic response generation have achieved state-of-the-art results on the EmpatheticDialogues dataset by incorporating emotional state information (Lin 2019, Majumder 2020). Mixture of empathic listeners (MoEL) is a proposed method for generating empathic responses that takes into account a distribution over a set of emotions (Lin 2019). Each emotion has its own "listener" that then generates an empathic response. These responses are weighted and sent to a "meta-listener" that then decides which response to return to the user. The MIME model builds on MoEL by adding a term that incentivizes the model to mimic the emotion of the speaker. This works well in social dialog scenarios between peers,. however, in therapy dialog the therapist may want to respond in a non-mirroring way. For example, mirroring may be inappropriate in cases where the emotional tone of the client does not match the expected emotional response to an event, e.g. if the client is laughing in response to describing a breakup. The mirroring concern along with concerns about the potential toxicity and unpredictability of open-ended text generation (in contrast with text retrieval from a constrained set of alternatives) were explored by recent research attempting to develop an empathic response prediction system for nursing using a system similar to MoEL. The investigators reported

generated responses that encouraged inappropriate behavior, insulted the patient, and bordered on harassment (Table 2.7) in supposedly empathic responses that attempted to mirror the feelings expressed by the patient (Shi 2019).

Table 2.7: Examples of inappropriate behavior of a generative model trained on empathic dialogues between nurses and patients (Shi 2019)

Emotion	Patient Message	Generated Nurse Response
Like	Be my girlfriend, I will make you happy.	I'll make you happy too.
Sad	I have a sore throat and swollen gums.	You're so pathetic.
Happy	The good news of their marriage made me want to get married, too.	Ha, you come here and we'll get married.

Due to the aforementioned challenges to meeting the demand for therapy, there has been a renewed interest in developing empathic models within the mental health domain. Sharma et al. developed models to evaluate empathy, as assessed by human raters along three dimensions using a custom rating scale, in peer-to-peer mental health support forums (Sharma 2020). This group subsequently developed a reinforcement learning approach to improve the empathy of peer-to-peer messages through textual insertions and deletions (Sharma 2021). Lubis et al. explored the elicitation of positive emotions through interactions with conversational agents (Lubis 2019). To my knowledge, the present work is the first to apply common-sense reasoning to the task of selecting empathic responses in the context of text-based teletherapy.

2.4 Gaps and Contributions

Empathy has been recognized as an essential component for conversational intelligence (Daher 2022) and for building a strong therapeutic alliance between a therapist and their client (Nienhuis 2018). Yet, emotions may

not be expressed explicitly, and most conversational agents have limited capacity to recognize and understand information that is not directly stated (Montenegro 2019). This work is motivated by the hypothesis that commonsense transformers (COMET) (Bosselut 2019), which can infer mental states in relation to events, may improve the ability of automated systems to identify empathic responses in these settings.

Methods to address this task have urgent applicability to assisting with the chronic shortage of mental health providers, on account of their ability to assist teletherapy providers with the selection of appropriate responses (as well as potentially preventing autonomous agents from selecting inappropriate ones). In this work, I explore the potential for such systems to enable care providers without extensive mental health training to provide protocolized therapy.

Such support may be of benefit even to experienced care providers, on account of a phenomenon known as "empathy fatigue" that leads to burnout and impaired ability to empathize with care recipients (Mottaghi 2020, Zhang 2021). Empathy fatigue is a work-stress condition caused by overexertion of empathic regions of the brain due to repeated secondary exposure to traumatic events (Cocker 2016, Yi 2019). Empathic response prediction systems may help care providers experiencing empathy fatigue to empathize with patients even when their ability to empathize is temporarily impaired.

Further, crafting empathic messages can be a bottleneck for text-based care efficiency, resulting in a trade-off between care quality and the number of clients that can be served per care provider (Lieu 2019). A human-in-the-loop system that assists care providers in responding emphatically would enable providers to provide quality care to more clients in a shorter time frame, increasing care efficiency. Continued integration of these systems with a provider-in-the-loop is one path toward conversational agents that can express empathy and respond autonomously to client messages.

Prior work has largely focused on empathic response *generation* with limited attention to retrieval based systems that make up the majority of contemporary dialog systems (Laranjo 2018, Montenegro2019, Kearns

2019). Further, while the majority of research papers on empathic dialog focus on generative models these have not been widely adopted in production health dialog systems, due to concerns over unpredictable generation of inappropriate responses or unknown biases causing potential patient harm (Bender 2021). On account of this research focus, there is a lack of evidence for the use of commonsense features for retrieval-based response selection, despite it being safer and more widely-deployed than corresponding generative approaches. This paper is the first to evaluate commonsense reasoning methods to predict both self-reported emotional states and empathic responses within the context of text-based teletherapy, and the pragmatic constraint that responses should be selected from amongst a set of carefully vetted alternatives. I compare a state-of-the-art emotion detection model to a commonsense reasoning model to predict self-reported emotional states and these emotional states are evaluated for their ability to augment a task-specific model to retrieve expert-provided empathic responses within the context of therapy dialog.

My method of data collection differs from prior work in that I collect self-reported emotional state labels in contrast to prior work that asked contributors to enact an assigned emotion category (Rashkin 2019) or post-hoc annotation by third-parties (Demszky 2020). This gives a ground truth representation of the emotional state *in situ* and without the need for additional annotation. Additionally, work on emotion detection has primarily used datasets collected from social media (Demszky 2020) or television show dialog (Busso 2008, Li 2017, Zahiri 2017, Poria 2019). Our work focuses on daily journaling exercises collected as part of a problem-solving therapy intervention. By exploring self-reported reactions to events, I probe more deeply into emotional state inference than is possible using the EmpatheticDialogues or GoEmotions sets. These sets are more concerned with the emotional affect or tone of the text than the speaker's internal emotional state, an important distinction in mental healthcare delivery. By collecting self-reported emotion data, I present a scalable way to train emotion detection systems that account for diversity among cultures and individuals, without imposing the perspective of a third-party annotator.

2.5 Caring for Caregivers Online

This work was completed in the context of developing Caring for Caregivers Online (COCO). COCO is a mobile application with an embedded chatbot providing family caregivers - particularly those with limited resources and high stress and burnout - with on-demand caregiving support and interactive self- and family management skill development in English. COCO supports family caregivers who - regardless of their child's specific chronic condition - experience common symptoms such as stress, anxiety, worry, guilt, and sleep disturbances. The technology is powered by a hybrid model with an AI chatbot and text-based sessions with providers to provide on-demand, personalized, and emotionally intelligent support and health solutions to improve caregiving symptoms and reduce caregiver burnout, thereby promoting their health and well-being. Features and functions include evidence-based intervention components such as daily check-ins, weekly problem-solving therapy (PST) sessions, automated reminders for self-care, caregiving symptom self-tracking, on-demand health and caregiving question-and-answer capabilities, and tailored resource recommendations.

I contributed to this project by developing the knowledge-based conversational AI system (infrastructure, modeling, and development) and contributing to the visual design of the mobile application, wizard-of-oz, and provider platform interfaces. I collaborated with a team of nurses and other care providers who designed the problem-solving therapy intervention (including the structure and content), handled participant recruitment and data collection, and assisted with moderating the usability sessions. A software developer implemented the designs to create the front end of the provider platform.

Chapter 3: Mental Health and COVID-19

While conducting my dissertation research, society experienced a global pandemic that resulted in a shift from in-person treatment to telemedicine. Mental healthcare was no exception. This chapter attempts to synthesize relevant details that occurred during this time to provide an historical account of this fundamental shift in healthcare delivery as well as to present work done to support healthcare efforts during this time that were outside the scope of the original dissertation proposal.

In December 2019, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was discovered in Wuhan, China. The World Health Organization (WHO) officially named the disease, caused by SARS-COV-2, coronavirus disease 2019 (COVID-19) on February 11, 2020. By the end of February, COVID-19 spread throughout the world and would be designated by the WHO as a global pandemic on March 11th. As a result, governments across the world closed their borders and instituted measures to curb the virus's spread including social distancing, lockdown orders, and other prevention measures. As would later be confirmed, the pandemic increased rates of anxiety and depression (Salari 2020).

I collaborated with researchers at University of Cambridge who were interested in tracking the effects of the pandemic on mental health by developing a short message service (SMS) based version of the Office of National Statistics (ONS) survey to be delivered daily by a conversational agent. In addition to this survey and in the interest of supporting the well-being of the population during this time, I included a daily reflection and weekly intervention that asked participants to reflect on how they may improve their situation, what barriers they may face when implementing that solution, and how they may overcome those barriers. This chapter

describes the delivery and results of this intervention and survey technique. Further, this chapter explains a paradigm shift in the delivery of healthcare which has increased the applicability of the findings of this dissertation.

3.1 Changes to the Healthcare Landscape

The COVID-19 pandemic has had a profound impact on behavioral healthcare. The increased levels of anxiety and depression triggered by the pandemic have strained the already limited resources of mental health practitioners. The U.S. healthcare system was already experiencing a shortage of mental health practitioners with 60 percent of U.S. counties without a psychiatrist. This problem was compounded by the aforementioned rising levels of anxiety and depression triggered by the COVID-19 global pandemic and socio-economic consequences of lockdowns.

In the initial stages of the pandemic, public health practitioners faced difficulty in ascertaining the extent to which COVID-19 had spread within communities and hospitals were inundated with calls requesting triage for COVID-19. To address the need to scale these triage services nationally, governments and healthcare organizations turned to conversational agents to field the volume of calls. In the early days of the pandemic, I worked to develop a system that could triage COVID-19 symptoms. The infrastructure for this system would later be used to deploy Cora, a system designed to measure mental well-being in the UK and provide a simplified version of problem-solving therapy.

While the technology for telehealth had been widely available prior to the pandemic, the rollout and adoption of these services had been slow. The Center for Medicare and Medicaid Services reported a 32x increase in telemedicine utilization for behavioral health between 2019 and 2020 (Samson 2021).

3.2 Cora Study

3.2.1 Data Collection

To better understand how the pandemic affected mental wellness, the Cora Wellness study collected survey responses for the well-being survey questions from the U.K. Office of National Statistics. I included two additional daily questions and three additional weekly questions that asked the participant to self-reflect (Table 3.1). The Cora Wellness study was launched in the U.K. and the code was released publicly to support similar efforts elsewhere (including a COVID-19 symptom checker).[3]

Table 3.1: Questions in ONS survey, daily reflections, and weekly reflections

Office of National Statistics Wellbeing Survey (ONS)

Daily Question	Response
Overall, how satisfied are you with your life nowadays?	1-10
Overall, to what extent do you feel that the things you do in your life are worthwhile?	1-10
Overall, how happy did you feel yesterday?	1-10
Overall, how anxious did you feel yesterday?	1-10

Daily Reflections (Cora)

Daily Question	Response
What gave you the most hope today?	Text
What caused you the most anxiety today?	Text

Weekly Reflections (Cora)

Weekly Question	Response
What is a possible solution to help you feel better?	Text
What is a potential barrier to implementing this solution?	Text
How might you overcome this barrier?	Text

[3] Code publicly available: https://github.com/kearnsw/cora/

The team recruited 95 participants residing in the UK (Mean Age=31.52, SD=9.87) using the online recruitment tool Prolific. Recruitment materials asked the participants to complete the ONS well-being survey and short daily reflection for 21 days (dropped to 14 days for participants who joined after the first week, n=66), a longer weekly reflection on goals and obstacles, and take an exit survey, after which they received compensation of £5 for study participation. There were three days in which the system either did not send the survey or did not record the submissions. In total, Cora collected 1,323 survey responses from participants (Figure 3.1).

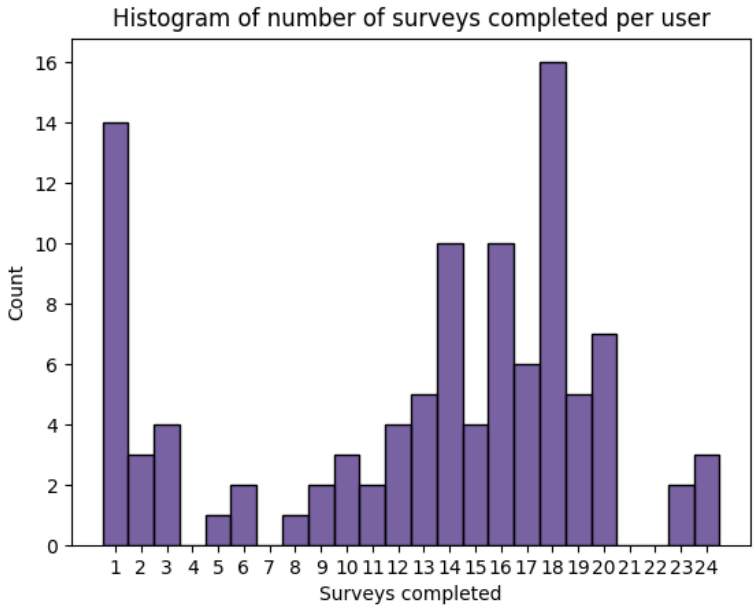


Figure 3.1: Number of surveys completed by participants

At the end of the study the participants answered the following questions as part of the exit survey:

1) What was your favorite thing about the text message survey experience?
2) What was the most challenging thing about the text message survey experience?
3) Over the course of the study, what brought you the most joy?
4) Over the course of the study, what brought you the most anxiety?

5) What helped to reduce your anxiety the most?

6) Any other thoughts you'd like to share with the research team?

3.2.2 Content Analysis

Content analysis (Krippendorff 1980) was used to answer two research questions:

RQ1: What were the leading sources of hope in the UK population during the COVID-19 Pandemic?

RQ2: What are the leading causes of anxiety in the UK population during the COVID-19 Pandemic?

To understand the causes of anxiety and hope, responses were annotated at the utterance level. An initial set of 200 responses of both hope and anxiety were labeled with one or more codes from a codebook of sixty-eight codes (Appendix B). The code book was refined until the inter-annotator agreement reached an average Cohen's K=.68 overall codes. The remaining data were split between the two annotators. The following sections provide a narrative synthesis of participant responses. The narrative is directed by the co-occurrence matrix (Figure 3.2) and frequency of unique participants that submitted messages that contained a code (Table 3.2 and Table 3.3). The full tables are provided in Appendix C.

Table 3.2: Unique participants per cause of anxiety

Code	Count
Nothing	63
WOR:General	37
ENV:Current_Situation	35
ACT:Thinking_about_the_Future	33
ACT:Travel	32
HEA:Other_Condition	32
ACT:Social_Activity	30
HEA:Mental_Wellbeing	30
PER:Child	30
ACT:Shopping	29
ENV:Pandemic	28
WOR:Tasks	27
HEA:Social_Distancing	26
FIN:Financial_Matters	23
PER:Family	21
PER:Partner	21
WOR:Issue_at_Work	21
ACT:News	20

Table 3.3: Unique participants per cause of hope

Code	Count
ACT:Social_Activity	66
PER:Friend	56
ENV:Current_Situation	53
Nothing	44
PER:Family	41
ACT:Outdoor_Activity	38
PER:Child	38
ACT:Thinking_about_the_Future	34
PER:Partner	31
EVE:Good_News	27
HEA:Mental_Wellbeing	27
GOV:Reopening	23
ACT:Home_Activity	22
EVE:Time_Off	22
WOR:General	22
WOR:Tasks	22
ACT:Media_Entertainment	20
ACT:Personal_Growth	20

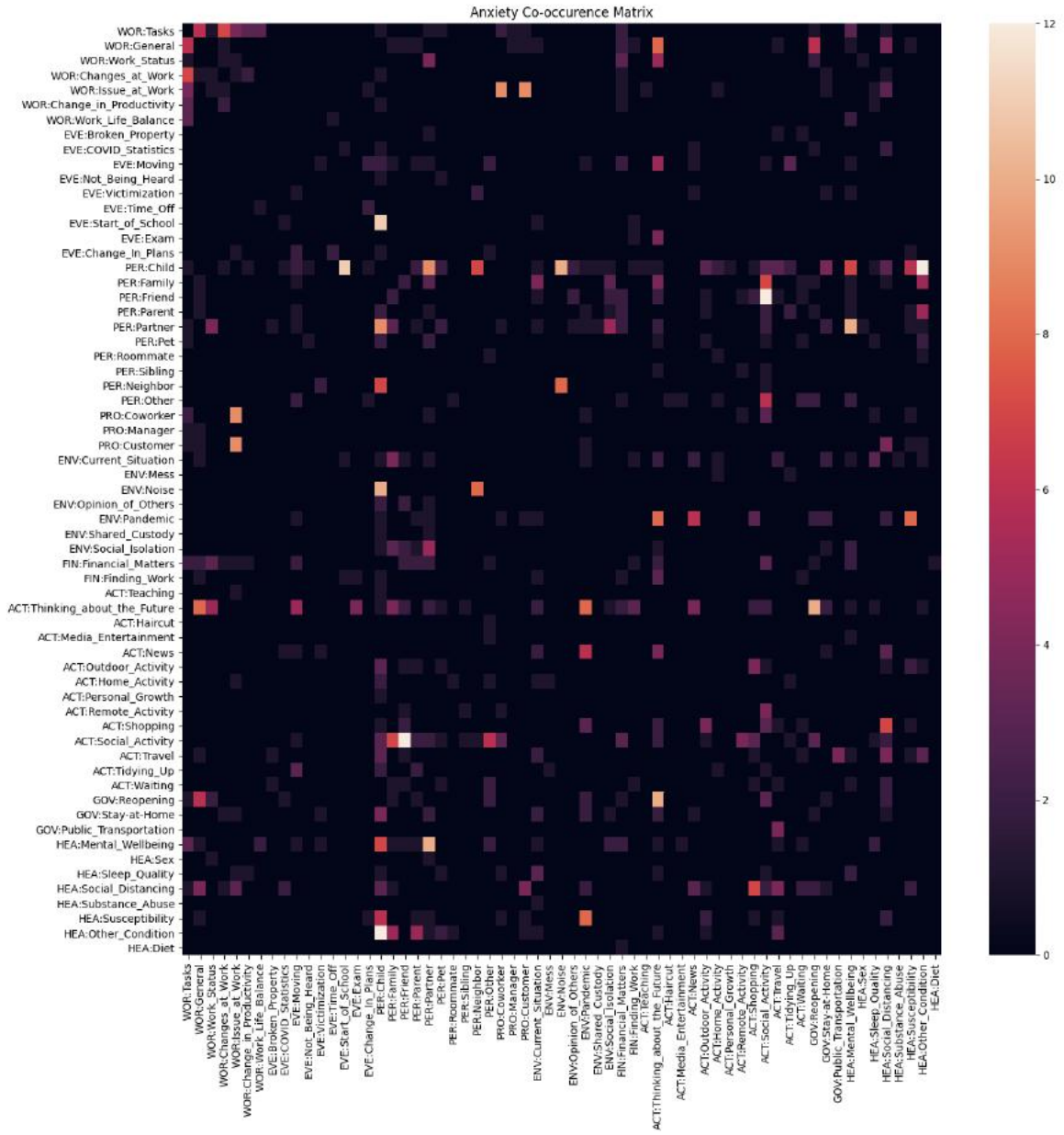


Figure 3.2: Correlation Matrix where one of each label exists in an utterance.

Causes of Anxiety

Work was a major contributor to stress among participants. They shared concerns about returning to work and potential changes to their work status, e.g. hours cut, loss of employment, and pay cuts. These thoughts were often anticipatory. Participants often mentioned feeling too busy or having difficulty completing tasks (n=27).

Participants were anxious about returning to work and what changes there would be at work. There were issues at work with customers not following social distancing rules.

Concern over health and well-being was another major theme of participant responses. Particularly, participants were concerned about the health of their children, families, and parents. Thirty participants mentioned their own mental well-being or that of their partners and children.

Participants were concerned about others not following social distancing guidelines and were hesitant about meeting with friends and family. Social distancing was primarily a concern when shopping and when taking public and chartered transportation.

Causes of Hope

Overwhelmingly, participants reported social activities with children, parents, family, and friends as the main sources of hope (Table 3.2). These activities were more likely to be remote in the case of family, friends, and parents. These activities largely took place outdoors in the case of the immediate family. This is in line with emotional wellness recommendations that social activity and fresh air lead to improved well-being. These opportunities become more difficult as activities shift more indoors in the Winter.

3.2.3 Exit Surveys

I used ATLAS.ti Mac (Version 22.0.6.0) to code the exit survey feedback. I present an informal analysis of the themes with examples from the participants feedback below.

Value

Participants overwhelmingly reported that the prompt to reflect on their day and mental health was their favorite part of the study. Further, participants indicated that they would be interested in continuing to use the

service and that they would recommend it to others. One participant indicated that the study motivated them to go back to therapy.

> *"Thank you. I would definitely use an app based service like this again"*

Usability

Participants found the system easy to use and valued being able to complete the survey whenever was convenient for them.

> *"It made me think about my mental health in a focused and calm way on a daily basis."*

They also valued that the survey did not take long to complete.

> *"It's easy and compatible with a busy life."*

Two participants indicated they found it difficult to provide a numerical value each day to how they felt in relation to the ONS survey questions. Some responses to challenges with the system indicated that the participants had trouble figuring out what caused specific emotions.

Social Presence

Many participants indicated that the messages from the system made them feel like they had someone with whom they could talk with multiple participants likening the system to a "friend". They reported feeling reassured and heard.

"I looked forward to your messages each day. It was like talking to a real person"

"I found that some things bothered me about my situation I didn't have anyone to talk to. It felt liberating telling by text about my problems"

However, some participants indicated that their biggest challenge was not getting a response because the system was a "robot".

Emotional Health Tracking

While not a requirement of the study, several participants indicated that they had used the service to personally keep track of their scores over time. This had both positive and negative effects based on the individual. For example, one participant used the fact that their scores were not increasing to consider ways to improve their life.

"It made me seriously consider why my scores weren't going up and what I could do to improve my life"

Whereas, other participants found their scores staying the same or going down as the most challenging part of the study. Due to the environment in which this data was collected, some participants indicated that they were not doing much at the moment and so found it challenging to answer.

"Thinking of different things each day when I haven't been doing much"

Emotional Categories

There were a large number of responses that indicated that nothing had brought the participant either hope or anxiety that day, which was reflected in some of the exit surveys where participants indicated that they wanted to express a greater variety of emotional states.

> *"My main emotion these days is anger not anxiety, and this was never covered"*

3.3 Discussion

The Cora wellbeing study provided the initial technical infrastructure and several learnings that were later built upon to develop the daily journaling system used to collect emotional state information in Chapter 4. It showed that users are willing to share detailed information about their daily lives with the system and that these can be labeled to contextualize quantitative measures of wellbeing. Further, participants found the exercise of daily reflections to be intrinsically valuable. These findings indicated that this delivery method may be suitable to collect the daily experiences of family caregivers in a low-friction way.

Asking participants to respond to specific emotions (hope and anxiety) was found to be inefficient, as there were a large proportion of answers that indicated nothing had caused hope and/or anxiety, which represent a missed opportunity to understand the user's emotional state. Most interestingly, exit interviews with participants indicated that they found identifying the causes of their emotional states the most difficult. This indicates a flaw in the prevailing order of mood check-ins within digital health interventions, which ask users to first indicate their emotional state and then provide the cause. This led to the swapping of these two questions in the journaling exercise described in the next chapter, i.e. asking first for a positive or negative event to prime the emotional memory of the user and then asking for the emotional state that resulted from that event.

3.4 Conclusion

Prior to this Cora wellbeing study, only the quantitative values were collected to measure wellbeing which did not provide insight into the mechanisms causing changes. Due to the large amount of changes that resulted from the pandemic, it would have been difficult to pinpoint what contributed to the changes in ONS survey results. In contrast, causes of hope and anxiety provided by the participants successfully provided insight into what factors contributed to population level changes in emotional health.

Further, engaging in the daily journaling exercise was seen as not only easy to complete but also inherently valuable to users. This stands in contrast to the current standard method of collecting emotional state labels which relies on annotation, which provides no benefit to the annotator.

Chapter 4: Emotional State Inference from Daily Journaling

In this chapter, I compare methods to predict self-reported mood changes as a result of events. These methods were evaluated on a dataset of *event-emotional state* pairs collected through a daily journaling exercise delivered by a conversational agent.

A systematic review of emotion detection systems found that they have not been tested in critical scenarios such as health care (Acheampong 2020), a gap that this research fills by contributing an open-source emotion detection dataset of self-reported emotional states in response to daily life events representative of the type of dialog in the "emotional check-in" phase of teletherapy. This chapter also describes an evaluation of the relative performance of computational methods for *emotional state inference* to approximate these self-reported emotional states (Example 4.1).

Utterance:	My son being really intense and stressed out	(Event)
Prediction:	Overwhelmed	(Emotion)

Example 4.1: The Emotional State Inference task is to predict the speaker's emotional state resulting from the given event.

As can be seen in this example, this task may require inference when an emotion is not reported explicitly. Generative commonsense reasoning models have not yet been compared to alternative forms of emotion detection for predicting self-reported feelings in daily check-ins. These models may provide a path to predicting a broader range of emotions than models trained on GoEmotions dataset. Whereas there are 27 emotional categories available in the GoEmotions classification dataset, there are over 10k emotional reaction labels included in the ATOMIC 2020 commonsense knowledge graph. This granularity is important in the context of therapy, as recent research has found that 150 distinct emotional categories are required to adequately respond to client needs in therapy settings (Brown 2020). The Cowen emotion categories (used to annotate the GoEmotions dataset) were derived from a factor analysis from data collected using a set of 34 button options given to annotators who were asked to describe their feelings in a laboratory setting in response to video clips (Cowen 2017). With contemporary technology, it is not possible to elicit the entire range of human emotions via short video clips, e.g. anguish or ecstasy, thus limiting the emotional granularity and expressivity of the GoEmotions model. Therefore, this work will explore using generative models to facilitate a more expressive form of emotion recognition and evaluate this ability on self-reported emotional states gathered using free-text rather than button input.

4.1 Data Collection

I built a conversational agent to deliver a daily journaling exercise and recommend goals via text messaging. In total, 179 family caregivers enrolled in the two-week study. The study team collected demographic data on 137 of the family caregivers who completed the exit survey (28 Asian or Pacific-Islander, 4 Black, 7 Hispanic, 1 Middle Eastern, 95 White, 2 reported more than one race) leaving 42 unknown. Of those participants, the majority were female (93%), working full-time (66%). Approximately half had an annual household income above $80,000 (~50%). The study completion rate was 79% with the majority of participants continuing to

interact with the system beyond the 5-day minimum and after an opt-out reminder message on the 14th day (Figure 4.1).

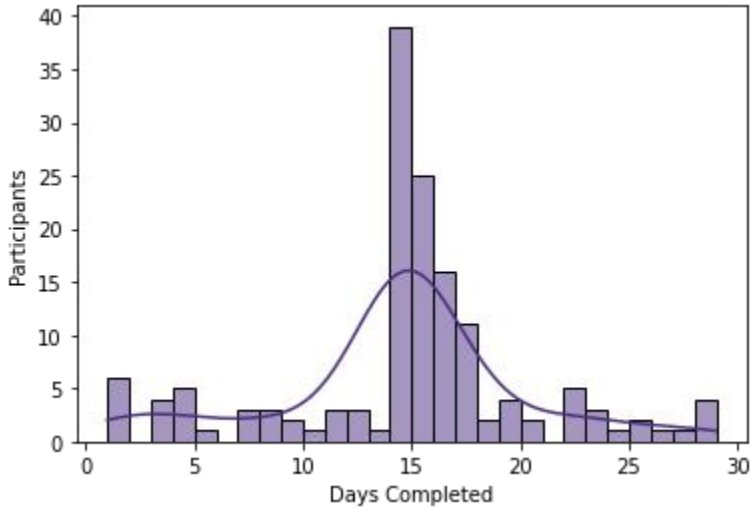


Figure 4.1: Completion of journaling exercises by the number of participants

Participants completed 2,134 daily journaling exercises. The daily exercise asked participants to describe two events, one positive and one negative, and to reflect on their emotional response to those events resulting in 4,268 event-state pairs (Table 4.1). This was reduced to 3,465 event-state pairs in the final dataset after removal of out of scope utterances, e.g. messages indicating technical issues, questions back to the system, etc. I collaborated with care team members to add a feature to the exercise in which half the participants were randomized to select a goal every three days in an effort to increase the benefit to participants and understand their perceptions of this feature. The evaluation of the goal suggestion feature was not evaluated in the scope of this work.

Table 4.1: An example reflection from the daily journaling exercise

Question	Example Response	Type
Tell me about a positive experience that you had today.	My daughter giggled for the first time today.	Event
In a couple of words, how did that make you feel?	Happy	Emotion
Tell me about an experience that could have been better for you today.	I wasn't as focused as I would like to be.	Event
In a couple of words, how did that make you feel?	Frustrated and disappointed	Emotion

Emotional concepts are thought to be compositional (Lindquist 2012), so participants were asked to describe their feelings in a couple of words which many used to provide more than one emotional category. To increase the granularity of self-reported emotions, participants were encouraged, but not required, to use an adaptation of the feeling wheel (Figure 4.2) based on the original that has three tiers of emotions with expanding complexity and seven core emotions (Willcox, 1982). The conversational agent suggested participants start by identifying their inner emotions (emotions from the innermost circle of the wheel) and then working their way outward to describe their emotions in more detail using one or two words.

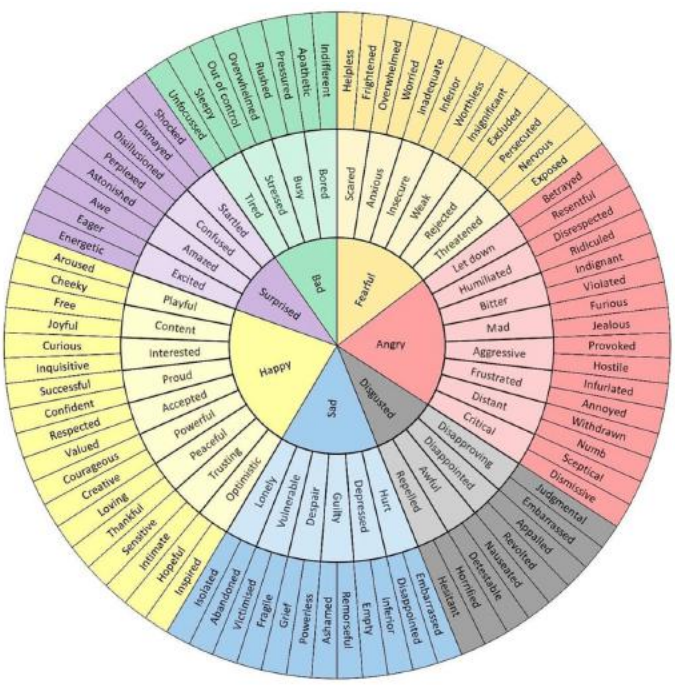


Figure 4.2: Feeling wheel by Geoffrey Roberts (reproduced with permission of content creator)

4.1.1 Emotional Response Analysis

The top 30 self-reported emotional states are clustered based on their co-occurrence in multi-state responses, e.g. "satisfied and accomplished" (Figure 4.3). There are two major clusters, which correspond to positive and negative emotional states. Although occasionally, responses contain mixed polarity emotional states, e.g. productive and worried. Within the positive cluster, joy co-occurred most frequently with reports of being thankful, loved, and content. Happiness was reported alongside feeling relaxed and proud. Participants reported feeling relieved and accomplished alongside feelings of satisfaction and pride. Within the negative cluster, participants reported stress alongside feelings of being overwhelmed and anxious. Reports of sadness co-occurred with reports of feeling worried, angry, or frustrated. Frustration was further reported alongside feelings of disappointment, annoyance, and sadness.

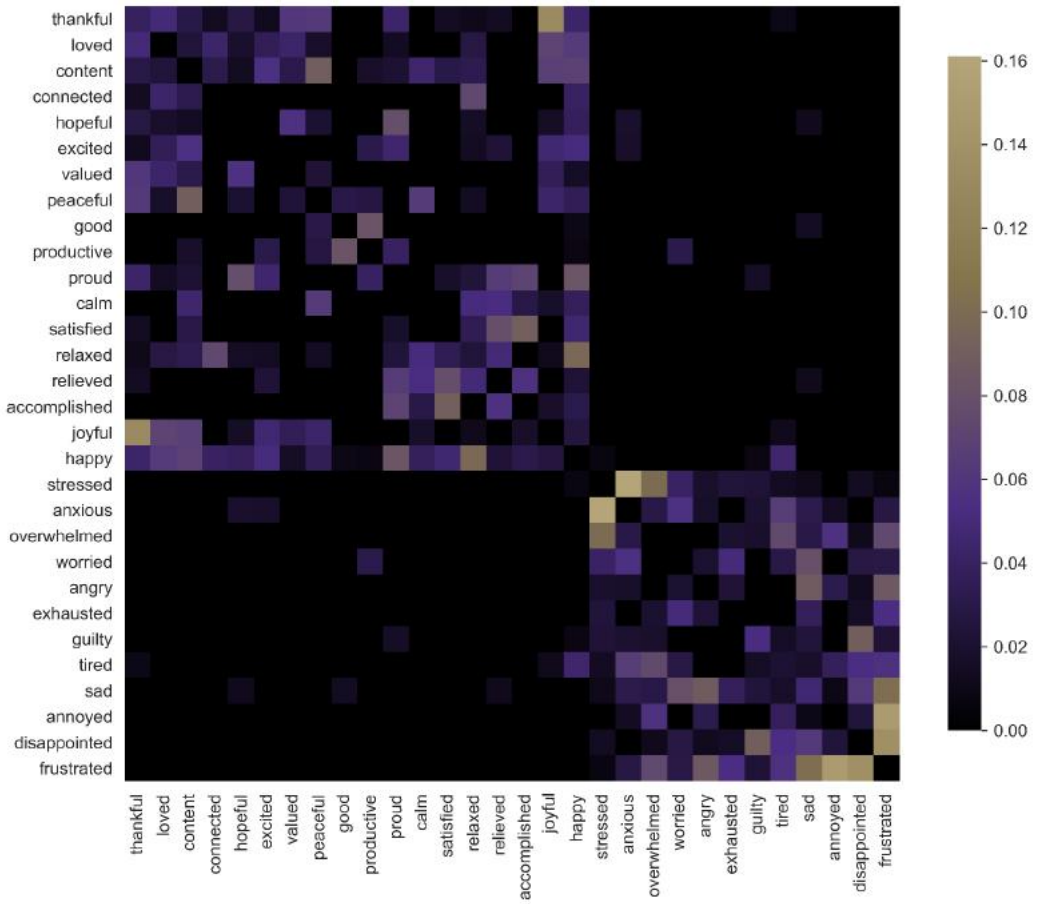


Figure 4.3: Hierarchical co-occurrence matrix of the top-30 self-reported emotional states

4.1.2 Ethical Considerations

The messages were monitored for safety concerns, and the studies were exempted by the IRB. Participants who used the service consented to their de-identified data being shared for research purposes and were compensated for participation.

4.1.3 Qualitative Analysis

Participants were asked during the exit interview if they had any feedback on the study. I coded this data and presented an informal thematic analysis in the remainder of this section.

Participants reported finding value in the exercise:

> *"It made you stop and actually think about how you are feeling, which sometimes I may not truly think about otherwise"*

> *"I liked the daily check ins - it made me more likely to reflect on my day."*

AI was seen as an acceptable delivery method and even reported to have social presence:

> *"It was nice to have someone ask about the good and not as good parts of the day, even if it was AI that was asking."*

> *"it felt nice to have the responses, even though i know it was all AI, like having a secret friend to talk to."*

Text-messaging as a mode of delivery was of high-value:

> *"[my favorite part was] getting a daily reminder to help me monitor my mood and mental health"*

As with most contemporary health dialog systems, the initial prototype used a sentiment analysis approach to respond with either a positive or negative response. This system occasionally made errors and all negative feedback was directed at these errors:

> *"The robot could have provided some better intelligent answers when the feeling that the person states is not in the chart - the robo program needs to get much much sophisticated for it to be useful and helpful."*

> *"Bot occasionally misinterpreted emotions (I'm glad you felt that way when I wasn't!)"*

> *"Also, the responses seemed so robotic and insincere and sometimes it said "I'm sorry to hear that" when I indicated a positive feeling/in response to a positive prompt and sometimes it said "that's great" in resposne to a negative feeling/prompt."*

4.2 Methods

4.2.1 Emotional State Inference Models

I evaluated GoEmotions (Demszky 2020), a BERT-based classifier of Cowen emotion categories, MentalBERT (Ji 2021) a mental health domain-specific masked language model, COMET (Hwang 2020), a commonsense knowledge grounded language model, and Instruct GPT-3 (Ouyang 2022), a large language model which has shown strong performance with few-shot prompting (Figure 4.4). The GoEmotions and MentalBERT models used the *monologg/bert-base-cased-goemotions-original* and *mental/mental-bert-base-uncased* tokenizers and

model versions available through the HuggingFace model repository[4] and were instantiated based on their model cards using the *AutoModelForSequenceClassification* and *AutoModelForMaskedLM* classes from the *transformers* library, respectively. The COMET model used the BART parameters and inference code from the publicly available repository.[5] The Instruct GPT-3 model used the *text-davinci-002* model variant, the latest version at the time of the study, which is privately hosted and served through the OpenAI API.[6]

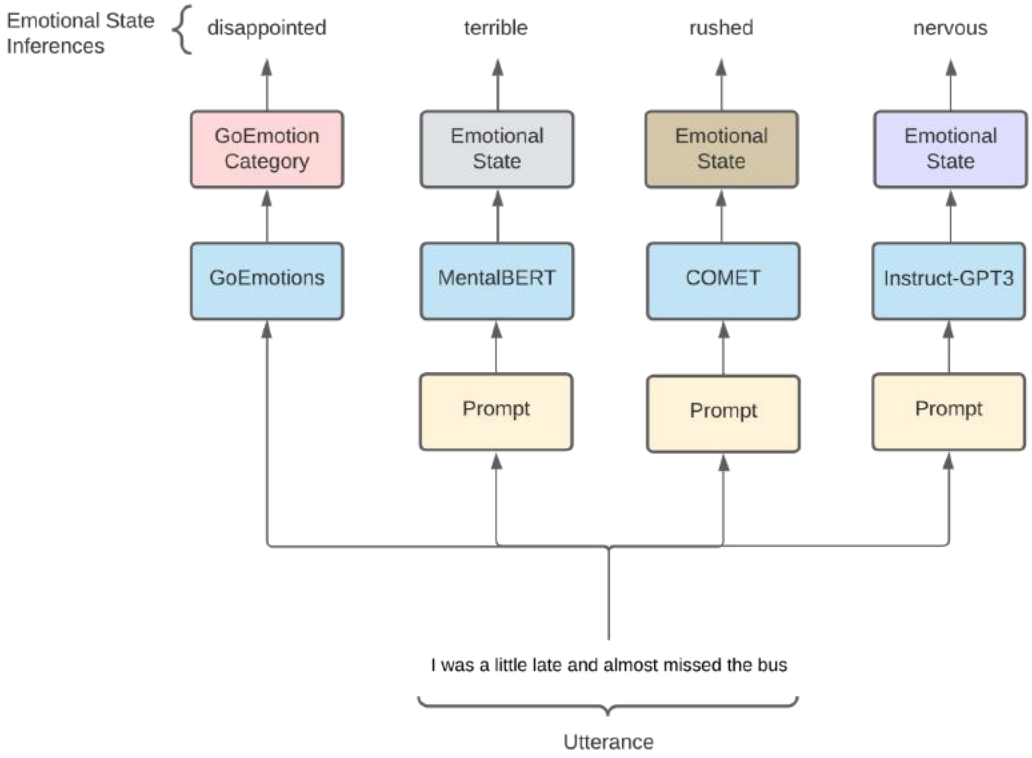


Figure 4.4: Emotional state inference methods examined in this study

[4] https://huggingface.co/models
[5] https://github.com/allenai/comet-atomic-2020
[6] https://openai.com/api/

4.2.2 Model Prompts

As the GoEmotions model is task-specific, no modifications to the events were necessary to predict an emotion category. Since the remaining models are general purpose language models, the input sequence was modified to be compatible with the modeling objective for which they were trained.

The MentalBERT model was prompted with a masked language model task:

Template: <event>. I feel [MASK]

Example: I went to the park with my kids. I feel [MASK]

Response: happy

COMET was prompted as a generation task using the ATOMIC 2020 special tokens with which it was trained:

Template: <event> xReact [GEN]

Example: I went to the park with my kids. xReact [GEN]

Response: happy

The Instruct GPT-3 model was prompted to assign an emotional state to the speaker using "prefixes", with no prior examples in a zero-shot approach:

Template: Provide the emotion felt by the speaker.

Speaker: {event}

Emotion:

Example: Provide the emotion felt by the speaker.

Speaker: I went to the park with my kids.

Emotion:

Response: The speaker felt happy and content.

The output generated by MentalBERT and COMET almost entirely consisted of single words and so was evaluated without any further processing. However, the generated output of the GPT-3 model repeated portions of the prompt with a few variations. These were each identified through manual inspection and used to automatically extract the first emotion word provided by the model. The accuracy of this process was confirmed by manual review.

4.2.3 Speaker Awareness

The COMET model differentiates between *agent* and *patient* attributing different emotional states to each, which have been defined within event semantics as the initiator and recipient of an action respectively (Kroeger 2005). As the point was to ascertain the emotional state of the speaker, I determined the role of the speaker in the described event and used heuristic methods to fill the appropriate ATOMIC relationship. I used the following heuristics to assign emotional states to the speaker based on (Sap 2019) relationships output by the COMET models.

Each event was first passed through an "en_web_core_sm" *spaCy* pipeline to apply part-of-speech tagging and dependency parsing.[7] The speaker was determined to be the agent of the event if any of the following criteria are met:

i. The subject (NSUBJ) of the sentence was a first-person pronoun, e.g. "*I* went to the beach" or

[7] https://spacy.io/

"*We* played a game".

ii. There was no subject of the sentence and the sentence begins with a VERB, e.g. "Went to the store" or "Had breakfast with friends".

4.2.4 Classification

To determine the generalizability of these results to other emotion detection tasks, I evaluated the models on the GoEmotions (Cowen, Ekman) and Cora (Hope/Anxiety) datasets. To achieve this, a transformer-based classifier was trained on top of the underlying hidden-state representations for each model (Figure 4.5). Specifically, the Dual Intent and Entity Transformer (DIET) classifier from Rasa 2.8.5 was used (Bunk 2020), which necessitated the input of sequence and sentence features, *e(u)* . The sequence features were the embeddings of each token for all models. For the BERT based models (BERT, GoEmotions, and MentalBERT), the sentence features were taken to be the [CLS] token (first hidden state). For the BART based models (BART and COMET), the sentence features were taken to be the last hidden state as suggested for classification tasks by the authors of the BART paper. These were then processed by the DIET classifier to predict a distribution, *p(e)*, over all possible emotions for that dataset (28 including *neutral* for GoEmotions and two for Cora).

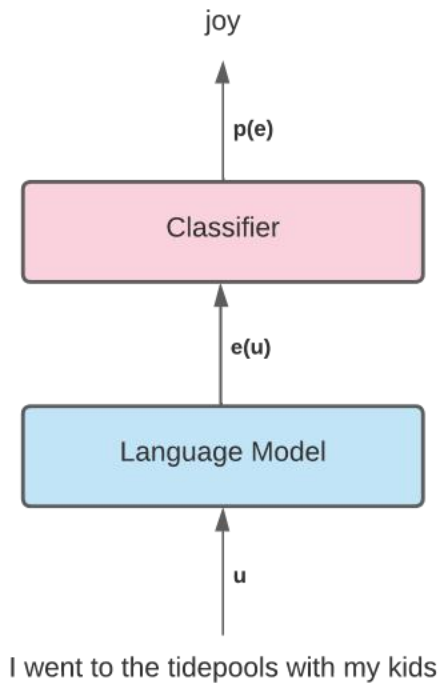


Figure 4.5: Classifier architecture for validation on GoEmotions and Cora datasets

4.3 Experiments

SBERT score is one of a number of soft-matching metrics that uses the cosine similarity between the semantic vectors of a source text and target text to provide a score between 0 and 1 (Reimers 2019). Using SBERT allowed comparison of the classification-based models (selection from a set of possible labels) against the generation models (unconstrained text generation) for prediction of emotional state (Figure 4.6). For the evaluation, I used the sentence-transformer library with the 'bert-large-nli-stsb-mean-tokens' model which I found to better correlate with human judgement of similarity between related emotions than other models.

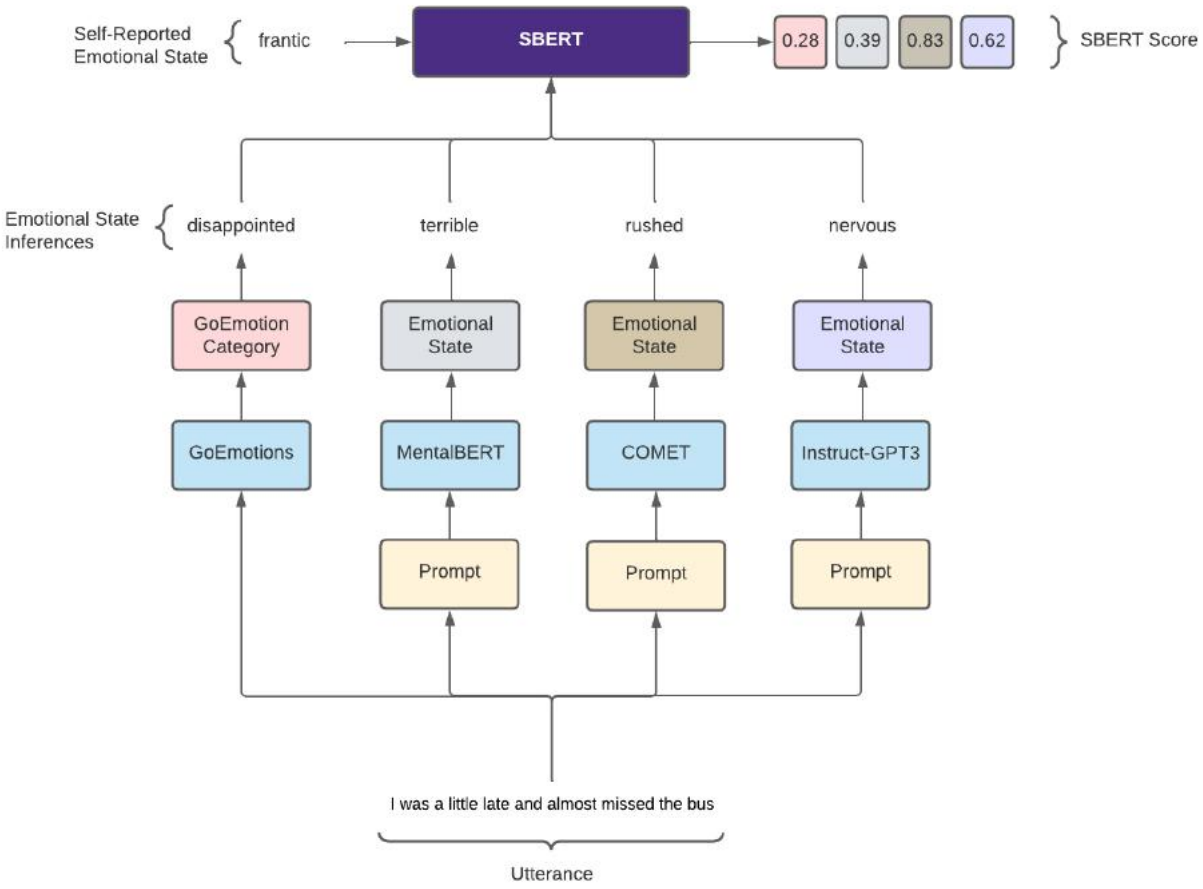


Figure 4.6: SBERT score calculation used to compare the top predicted emotions of each model to the self-reported emotion values.

Each model predicted a single emotion label with similarity estimated by the SBERT model in relation to the self-reported emotional state for an utterance. The average SBERT score over all utterances is used to provide a single score for each model (Table 4.2).

To approximate the upper bound on model performance, I ran an oracle experiment. I report the results alongside the other methods. To compare with the other methods, the oracle is only allowed to select one emotion. The oracle was designed to select the emotion category with highest frequency from the set of emotions in the self-report. For example, "happy and lost" would consist of these two emotion categories

["happy", "lost"], since "happy" is more frequent in the dataset. The oracle selects "happy" as the answer which is then compared to both "happy" (sim=1.0) and "lost" (sim=0.1) for an average of (0.55). The oracle therefore accounts for the highest performance possible given knowledge of how a particular individual responded to a particular event.

Table 4.2: Examples from error analysis of emotional state inference predictions by each model with similarity to the self-reported label measured by the SBERT similarity score in parentheses

Event	Oracle (Self-Report)	COMET	GoEmotions	MentalBERT	Instruct GPT-3
I received positive feedback on a draft of my term paper that I worked very hard to write.	validated	happy (0.421)	**approval (0.767)**	lost (-0.023)	happy (0.421)
I received a call from CPS that my son's biological brother needs placement	nervous	**worried (0.726)**	neutral (0.114)	terrible (0.522)	overwhelmed (0.494)
I was a little late and almost missed the bus	frantic	**rushed (0.830)**	disappointment (0.281)	terrible (0.393)	nervous (0.620)
Conflict with family over activities	frustrated	upset (0.812)	neutral (0.017)	alone (0.274)	**frustrated (1.000)**

4.3.1 Comparison of Emotional State Inference Methods

COMET performs the best at predicting self-reported emotional states with close performance (Δ-0.018) by Instruct GPT-3 with one-third of the parameters (Table 4.3). The top-10 predictions from each model are reported in Table 4.3, with those that have an SBERT score >= 0.8 when compared to the oracle predictions are bolded in **gold** and those with SBERT score >= 0.7 are bolded in **light purple**. The Instruct GPT-3 model produced responses that matched the most common self-reported values. However, the COMET model produced the greatest variety of emotional states at 72, over twice as many as the Instruct GPT-3 model.

Table 4.3: Emotional state inference performance measured with SBERT metric on self-reported labels, and the number of unique emotional states and top-10 predicted categories on the study test set

Method	SBERT	Categ.	Params	Top-10 Categories
GoEmotions	0.332	23	108M	neutral, **disappointment**, admiration, **joy**, **sadness**, **annoyance**, approval, desire, anger, realization
MentalBERT	0.365	25	109M	**terrible**, **bad**, alone, sick, **awful**, you, lost, stupid, **great**, empty
Instruct GPT-3	0.558	35	1.3B	**happy**, **frustrated**, **content**, **proud**, **relieved**, **sad**, **disappointed**, accomplished, **excited**, **regretful**
COMET	0.576*	72	406M	**happy**, **tired**, **sad**, accomplished, **satisfied**, **frustrated**, **angry**, **upset**, **relaxed**, **relieved**
Oracle	0.875	217	-	**happy**, **frustrated**, **sad**, **disappointed**, **tired**, **content**, **relaxed**, **proud**, **joyful**, **overwhelmed**

* Statistically significant improvement from Instruct GPT-3 (p=0.0006) using t-test across all SBERT scores

4.3.3 Speaker Awareness Experiment

To understand the effect of speaker awareness on the COMET model performance, I ran an experiment to add the speaker awareness module (Table 4.4).

Table 4.4: Performance of COMET model with speaker awareness

Method	**SBERT**
COMET	0.576
COMET - Speaker Awareness	0.567 (Δ-0.009)

Results indicate that speaker awareness had minimal to no effect. An explanation for the limited effect of speaker awareness on model performance is related to "emotional mirroring", which is reflected in the ATOMIC 2020 training data where annotator labels for the feelings of the subject (xReact) and others (oReact) for a given utterance had an average SBERT score between them of 0.573 when labels were aligned based on maximum SBERT similarity. Within the diary study dataset, the oReact and xReact values predicted by the COMET are exact matches in 1988/3465 (57.4%) of examples.

4.3.4 Validation on Cora Data

The 2,259 examples in the Cora dataset (presented in Chapter 3) were divided into 5-fold cross-validation splits each with a ratio of 4:1 (train, test). The Cora dataset contains self-reported events but focuses on only two emotional states: hope and anxiety. Results from this experiment follow the same relative performance ranking as those on the self-reported dataset (Table 4.5). The Instruct-GPT evaluation was a late addition to the proceeding section and due to the financial cost and necessity to modify the architecture to incorporate Instruct-GPT embeddings, they were not included in the Cora or GoEmotions validations.

Table 4.5: Results of the model on the binary emotion classification task averaged across folds

Method	Hope (F1)	Anxiety (F1)	Overall (Macro F1)
BERT	0.722	0.753	0.739
GoEmotions	0.760	0.733	0.746
MentalBERT	0.799	0.758	0.780
BART	0.805	0.778	0.792
COMET	**0.815**	**0.841**	**0.829 (+Δ0.090)**

4.3.5 Validation on GoEmotions Data

The GoEmotions dataset was used to validate that the evaluation method proposed in this chapter aligned with the results on an existing publicly available dataset.[8] The results on GoEmotions are lower than that reported by the authors of that paper, which may be the result of any number of factors including architectural differences, their use of hyperparameter tuning and their reporting of the best performing model whereas I report the average across 5 training runs. I made these choices to ensure a fair comparison across models rather than attempting to set a new benchmark. As with the evaluation on the Cora dataset, the relative ranking of the models remained the same (Table 4.6). Taken together, these results provide strong support for the merit and validity of the proposed evaluation method. Also, the results on the self-report data show that there is a clear advantage to using COMET as demonstrated on all three datasets.

[8] https://github.com/google-research/google-research/tree/master/goemotions/data

Table 4.6: Results of models on the GoEmotions dataset

Method	Macro Precision	Macro Recall	Macro F1
BERT	0.321	0.305	0.309
GoEmotions	0.331	0.309	0.317
MentalBERT	0.362	0.340	0.347
BART	0.390	0.371	0.373
COMET	**0.420**	**0.378**	**0.392 (+Δ0.083)**

4.4 Discussion

Emotion recognition from conversation is an emerging field of study. Work to date has focused on third-party annotations which do not represent the ground truth of the speaker, or used actors portraying stereotypes of emotion categories that may not translate to real-world settings where displays of emotions are less overt. Further, these systems have used models of emotion that do not capture the full spectrum of human emotion as described linguistically. This work adds to the literature in this field by testing emotion detection models on self-reported emotions in response to real experiences and allowing for free-form responses to capture a wider variety of emotion categories. Testing this method in the context of a journaling exercise presents a clearer picture of how these methods may perform in a healthcare setting.

On this novel task, the COMET model outperformed all other models at predicting self-reported emotions. COMET outperformed GoEmotions by 0.244 similarity score with SBERT, MentalBERT by 0.211, and InstructGPT by 0.018. This can be interpreted as a relative performance increase of 73.5% relative to GoEmotions (an emotion detection model), 57.8% relative to MentalBERT (a domain-specific language model), and 3.2% relative to InstructGPT. This finding for COMET was further validated on Cora and GoEmotions

datasets. On the binary task for the Cora dataset, COMET outperformed GoEmotions by 8.3%, MentalBERT by 4.9%, and its base language model, BART, by 3.7%. On the 28-category classification task for GoEmotions, using the aforementioned modeling approach, COMET outperformed GoEmotions by 7.5%, MentalBERT by 4.5%, and its base language model, BART, by 1.9%. These findings strongly support the hypothesis that commonsense inference improves emotional state recognition. Further, the zero-shot prompting experiment with Instruct GPT-3 indicates that LLMs learn to infer reactions to events, a basic component required for a theory of mind, and that there is room for improvement when compared with the COMET results (indicating the potential to improve Instruct GPT-3 performance by tuning on commonsense knowledge). The oracle evaluation results indicate that there is still significant room for improvement on this task. For future work on this task, I recommend modeling the speaker state over time to better learn to predict how they individually respond to different environmental stimuli. In contrast, the current work tends toward the average emotional state response not accounting for individual differences, which is a logical next step for this line of research.

The GoEmotions model predicted a large number of *neutral* labels, reflecting the most common category in its training data. This is indicative of the limitations of having emotional states labeled by anyone other than the speaker. Models trained on this type of data learn to pick up on distinct linguistic markers rather than gaining an understanding of event semantics. This may explain the limited performance of the GoEmotions model when transferred to the diary dataset.

Based on participant feedback on the journaling exercise, they enjoyed reflecting on their day via an automated system. However the sentiment-based approach to empathic responses was clearly indicated to be insufficient for developing rapport with study participants, which is a limitation that I address in the next chapter.

4.5 Conclusion

Language models fine-tuned for commonsense reasoning perform better at predicting the self-reported emotional states of users from described events than robust baseline models. This indicates that commonsense reasoning is important for predicting emotional responses to daily life events. Further, it provides a potential method to continuously fine-tune these models with the capacity to infer emotional responses to events by interacting directly with clients by incorporating *event-emotion state* pairs collected through daily check-ins within the training data to predict responses to future events. In the next chapter, we will explore the utility of modeling this information within the context of predicting empathic responses, a task which helps address participant concerns around insincerity of sentiment-based approaches while supporting rapport building and social presence.

Chapter 5: Enhancing Empathic Response Prediction with Emotional State Inference

User feedback presented in the last chapter indicated that some participants viewed the positive and negative sentiment-based responses to their disclosures of information as insincere or invalidating, harming rapport. This finding underscores the well-established importance of empathy in developing a therapeutic relationship (Feller 2003). Recognition of the emotional state of a client in therapy is a key component of establishing empathy with them (Mercer & Reynolds 2002), leading to the hypothesis that incorporating emotional state inference into a system would improve its performance in selecting an appropriate empathic response within a therapy context.

This chapter presents an alternative approach to empathic response prediction that increases empathic expressivity over contemporary dialog systems that use sentiment or emotion detection-based methods. In it, I describe and evaluate a method to predict an appropriate empathic response relevant to a described event, which can be used to suggest a response to a care provider or automatically respond, so patients feel heard.

To develop this task, subject matter experts grouped the daily events described by participants in the daily journaling exercise such that they could use a single empathic message to respond to all messages in a group (Example 5.1). Each group was assigned a label, such that additional empathic messages could be written for that label to introduce variety across multiple sessions of therapy by sampling without replacement from the set of responses for a label to prevent repetition.

User 1:	I needed to get some work done, and I only did a small portion of it.	(Event 1)
User 2:	I wasn't as focused as I would like to be.	(Event 2)
AI:	Staying focused can be hard. Sounds like you're a little frustrated with yourself.	(Response Text)
	[*Work* + *Unproductive*]	(Response Label)

Example 5.1: The Empathic Response Prediction task is to predict the most appropriate response given an event through classification. To identify the appropriate empathic response, it is necessary to recognize the nature of the event (lack of productivity at work). The driving hypothesis of this work is that the additional ability to infer the underlying emotional state of the User (frustration) will benefit task performance.

This process of data labeling is covered in more detail in the next section and is followed by an explanation of the proposed method of incorporating emotional state inference predictions and their evaluation against emotion detection and sentiment-based methods.

5.1 Data

This work leveraged the same data set described in Chapter 4, with all 3,465 events grouped by a team of subject matter experts in psychology and nursing based on how they would emphatically respond in a typical therapy dialogue. The SMEs chose to use Maslow's hierarchy of needs (Maslow 1954) as a guide for the initial assignment of events into response categories, or *topics*, e.g., "family" or "work". Their rationale was that one of the important functions of emotions in this context is to indicate whether a caregiver's needs were met in response to a particular category of events. For example, a family caregiver often feels torn between work and

caregiving, which is hindered by the need to be productive at work. This leads to feelings of *guilt* for being unable to be the best version of themselves in both worlds. Given that emotions also serve as social signals that can elicit responses and facilitate specific behaviors from others (Keltner 2003), a response that accurately reflects the (un)met needs and the subsequent emotions underlying the event indicates to the listener a deeper level of empathy.

The annotation process was conducted using the HumanFirst NLU design tool and followed a directed content analysis approach (Hsieh 2005). First, two team members worked together to apply and adapt the hierarchy of needs and created labels. After a framework for labeling was established, two team members labeled the utterances independently to validate and further develop the code book. Through this process, new labels were continuously added and efforts were made to reduce the number of labels to minimize redundancy. The team members met for 60 minutes per week for eight weeks to discuss the annotations, merge and split clusters, and reach a consensus on the final response label set. This process resulted in a non-uniform distribution (Figure 5.1) of 70 hierarchical response labels/topics with an associated response for each label as demonstrated in Table 5.1. Each label is a combination of topics, e.g., "*family* + *child* + *joy*" consists of three topics, "*family*" and "*child*", and "*joy*". For simplicity, each unit of the label will be referred to as a topic, although this occasionally will include emotions, e.g., "joy".

Table 5.1: Examples from the dataset include the utterance containing an event, the response label, and the response text.

Utterance (Event)	Response Label	Response Text
My daughter giggled for the first time today.	family+child+joy	Aww. Children are such a blessing. They can bring so much joy!
My kid was extra whiny today	family+child+difficult behavior+tantrums	That's frustrating. Kids can be hard to deal with sometimes.
I wasn't as focused as I would like to be.	work+struggling	Staying focused can be hard. Sounds like you're a little frustrated with yourself.

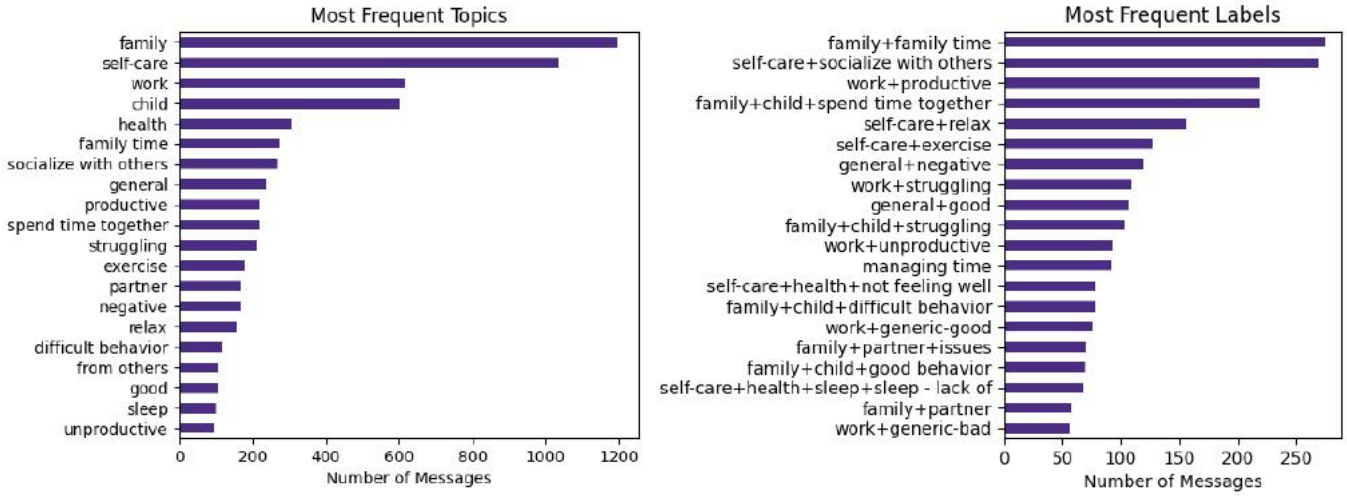


Figure 5.1: Most frequent (n=20) response labels and topics across the entire dataset

5.2 Methods

Figure 5.2 provides a high-level schematic overview of the methods, which is followed by a more detailed account in Sections 5.2.1-5.2.3. To isolate the contribution of representation method variations toward performance on the empathic response prediction task, I tested each variant with this generic classification

architecture and the same hyperparameters. The utterance provided by the caregiver, u, was encoded into an embedding, $e(u)$, using a word-piece tokenizer (Wu 2016) and pretrained transformer-based language model architecture variant, e.g. BERT (Devlin 2019) or BART (Lewis 2019). A randomly-initialized task-specific transformer-based classifier (decoder) was trained from the embedded representations of an utterance to predict the response label assigned to that utterance by the subject matter experts. During this process, all pretrained language model weights were frozen, and only the randomly-initialized transformer classifier was tuned to the task.

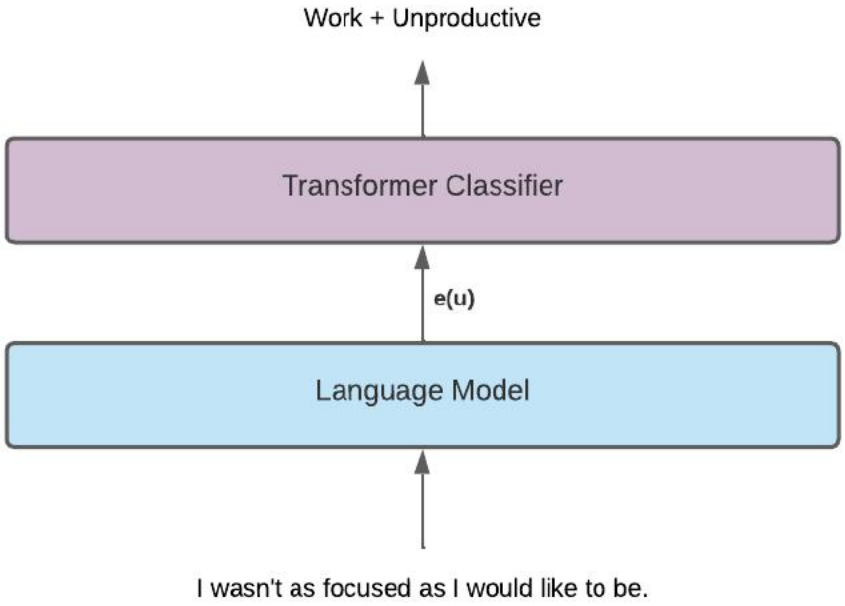


Figure 5.2: The model architecture used for all experiments consisted of a transformer-based language model encoder and a transformer-based classifier that predicted the appropriate class of response given a caregiver utterance

The remainder of this section provides a detailed account of the representation methods tested and details on how the classifier was trained from these representations, including how emotional state inferences were incorporated into model predictions.

5.2.1 Pretrained Language Models

Pretrained language models (PLMs) were used to encode both the utterance (e.g. "My daughter giggled for the first time today.") and the response label (e.g. "*family+child+joy*") with the "+" symbols replaced by spaces. The different variants benchmarked are presented in Table 5.2 and include both MentalBERT, a mental health domain-specific model, and COMET, fine-tuned for commonsense reasoning, along with their base language models BERT and BART, respectively, to understand the effect of these two domain-adaptions to empathic response generation.

Table 5.2: Representation methods tested, their base language model, and on what data they were fine-tuned if applicable

Model	Language Model	Domain Adaptation (Target)	Citation
BERT	bert-base-uncased	-	Devlin 2019
MentalBERT	bert-base-uncased	Mental Health Datasets	Ji 2022
BART	facebook/bart-large	-	Lewis 2020
COMET	facebook/bart-large	ATOMIC 2020	Hwang 2021

Due to the limited size of the dataset (n=3,465 utterances as described in Chap. 4), the PLM weights were frozen and used to encode the input sequence into both a global utterance and individual token embeddings. The global utterance embedding is a single vector that summarizes the information in the input sequence. The number of individual token embeddings varies across utterances and is equal to the number of word-pieces in the input utterance concerned. In accordance with the authors' suggestions within the seminal papers for each model, the BERT-based models use the [CLS] token output to represent the utterance, whereas BART-based models use the last hidden state of the sequence to represent the utterance. The net result for each utterance is a sequence of embeddings, *e(u)*, consisting of one for the utterance as a whole, and one for each word-piece token derived from it. To present an input of uniform size to downstream models, the resulting matrix (one row per embedding) is padded with zero vectors such that the number of rows is equal to one (the utterance embedding) plus the number of tokens derived from the longest utterance in the training set (Fig. 5.3).

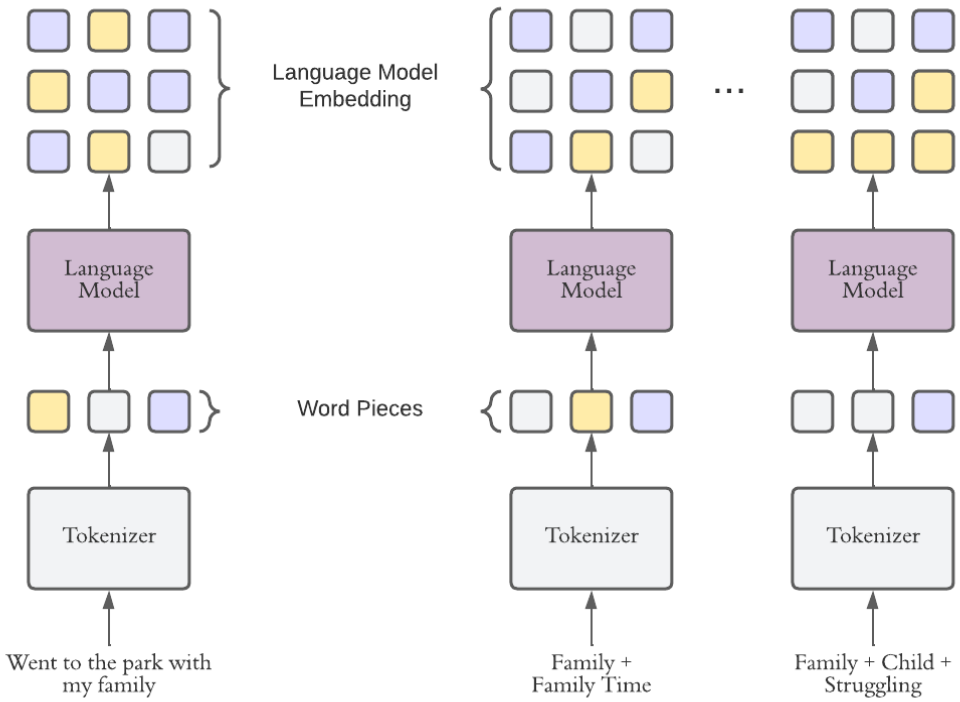


Figure 5.3: Language model embedding process

5.2.2 Transformer Decoder

For each empathic response prediction (ERP) experiment, I provided the language model embedding described in the last section as input to a TED policy (Vlasov 2020) (a transformer-based classification architecture described in detail in Chapter 2) while varying the input representation method between conditions (Figure 5.4). I evaluated the classifier using as input the language model embeddings of the event, $e(u)$, or the language model embedding of the concatenated utterance and emotional state inference model outputs, $e(u + s_{xReact})$. The model is trained to minimize the cross entropy loss between a $\mathbb{D}^{20}$ projection of the language model output and $\mathbb{D}^{20}$ projections of the embedded response labels. The cross-entropy loss is backpropagated through the feed-forward and transformer layers to update the model weights stopping at the frozen language model weights.

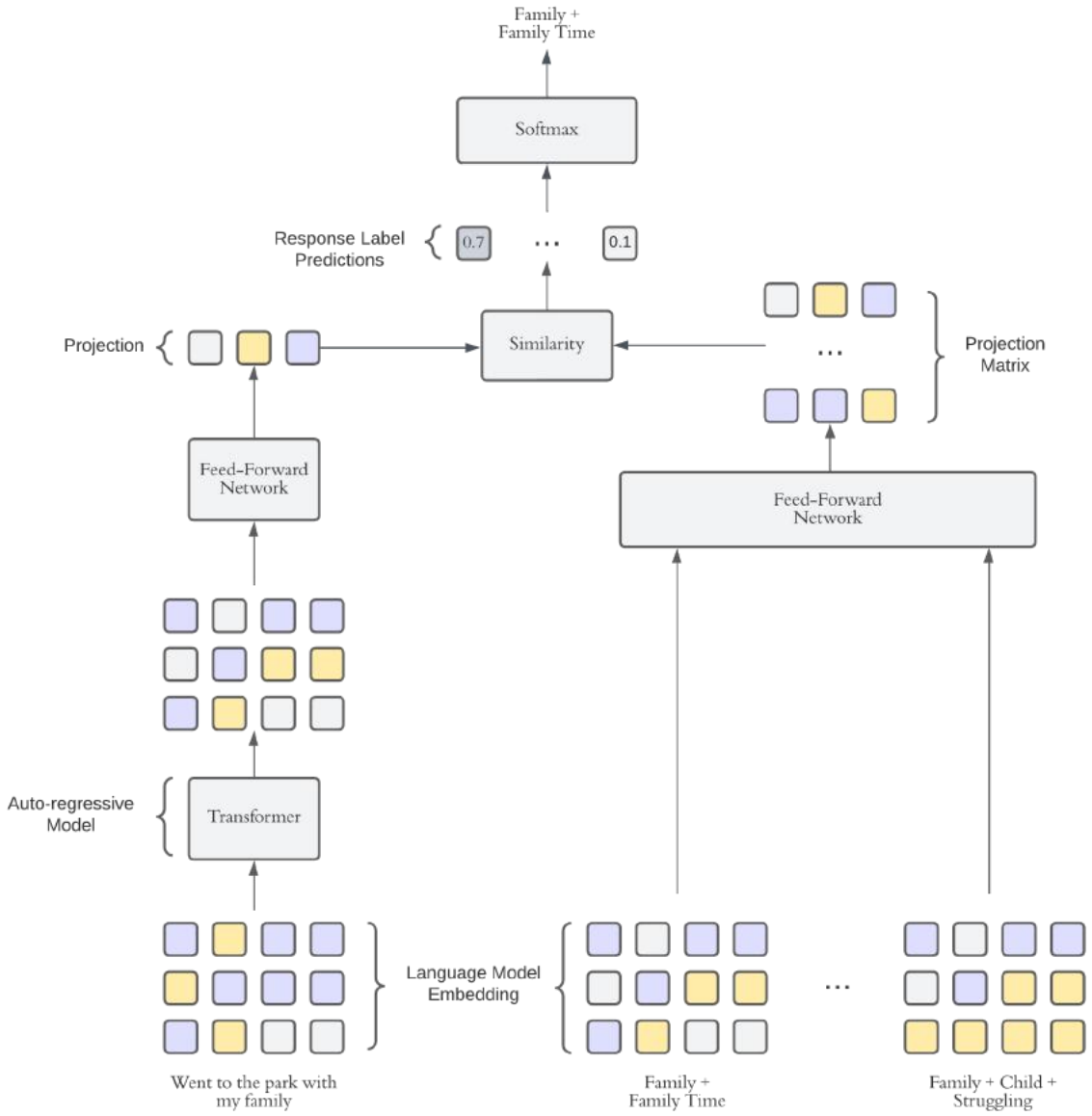


Figure 5.4: TED Policy to predict a distribution over candidate response labels given an utterance

5.2.3 Emotional State Inference

Building from the emotional state inference experiments in the last chapter, the COMET model was used to infer the emotional state of how the caregiver felt, *xReact*, as a result of the described event. This inference was generated in the same way as described in Chapter 4. The resultant inference was appended to the utterance to create what I will refer to as the ESI-augmented utterance (Figure 5.5). The ESI-augmented utterance was then

encoded using the language model and passed to the generic classification architecture as described in the last two subsections.

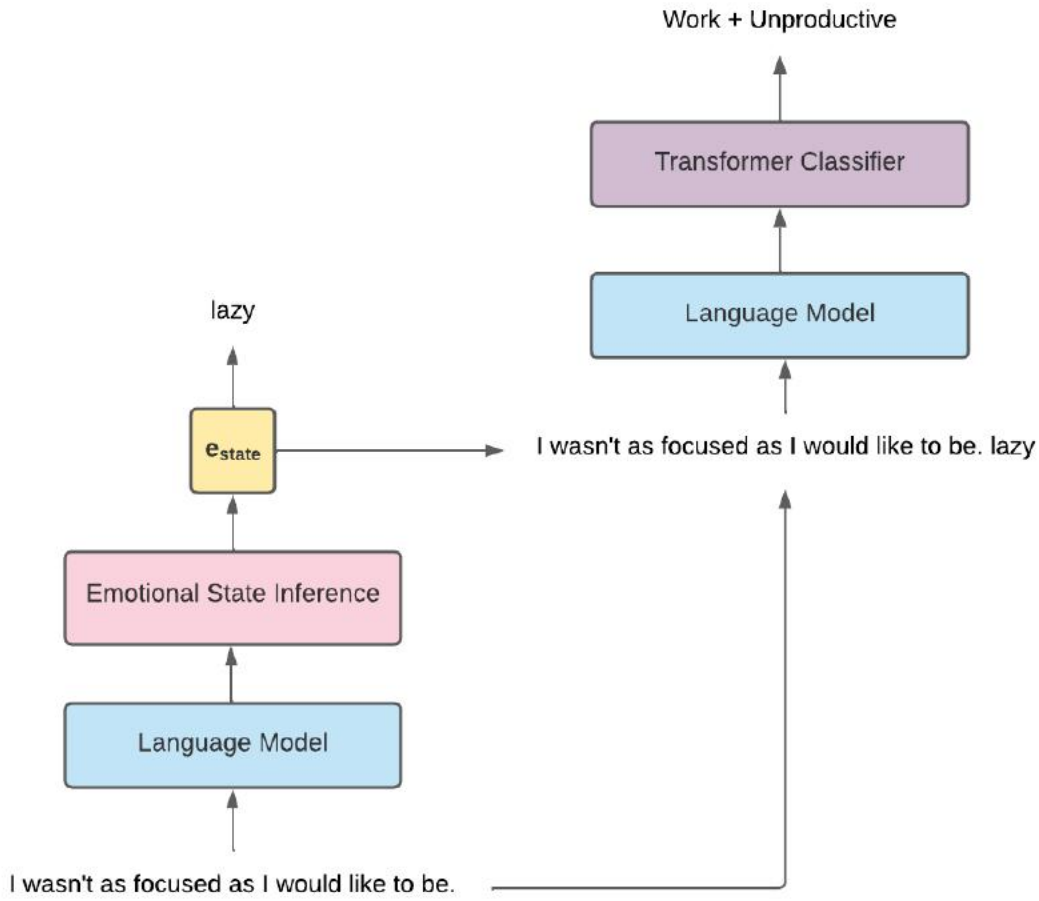


Figure 5.5: The architecture used to evaluate the impact of ESI on ERP.[9]

[9] **N.B.** Adding the emotional states as slots did not increase performance with the TED policy in Rasa v2.8.5. Indicating a potential bug in the software. To avoid this impacting the results, the emotional state was appended to the utterance instead.

5.3 Experiments

The representation methods described in sections 5.2.1 and 5.2.3 were evaluated using the weighted precision, recall, and F1 of their resultant empathic response predictions across 5-fold cross-validation. To account for differences in performance due to the underlying language models of GoEmotions (BERT) and COMET (BART), I report empathic response selection for their base models as a baseline and also include MentalBERT, a domain-specific BERT model.

All experiments were run on a 64-core Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz. Training and evaluation were done using the Rasa Open Source framework.[10] The TED policy was configured with the defaults in Rasa v2.8.5, which consisted of a single transformer layer that was optimized using cross-entropy loss. Each training run consisted of 100 epochs using a learning rate of .001 and took approximately 1 hour, and inference takes 3-6 seconds per example depending on model architecture.

5.3.1 Comparative Analysis of Pretrained Language Models

I first evaluated the pretrained language models on unmodified utterances. This gave a measure of how much information was captured latently in the pretrained language model and set a baseline to compare the addition of emotional state inferences against in the next subsection (§5.3.2). The difference between the BART model and the COMET model is that the latter was trained on the ATOMIC-2020 commonsense knowledge graph which includes mental state inferences (§2.2.3).

[10] https://github.com/RasaHQ/rasa

Table 5.3: Language model performance on the non-augmented utterance

Language Model	Model Params	Domain-Adaptation	Precision	Recall	F1
BERT	108M	-	0.395	0.404	0.399
MentalBERT	109M	Mental Health Data	0.429	0.441	0.434
BART	406M	-	0.457	0.467	0.462
COMET	406M	Commonsense Reasoning	**0.535**	**0.543**	**0.539**

As shown in Table 5.3, the COMET model outperformed MentalBERT (+Δ0.105) the mental health domain-specific language model as well as the baseline models BERT (+Δ0.140) and BART (+Δ0.075). This shows the benefit for empathic response prediction of fine-tuning language models on commonsense reasoning knowledge. For the remaining experiments, I compared COMET, the best performing model, to BART the model from which it was fine-tuned on the ATOMIC-2020 knowledge graph. The following sections describe analyses probing into the factors contributing to the COMET model's improved performance.

5.3.2 Evaluation of Emotional State Inference

Example 5.2 provides an illustration of the ESI-augmented utterances used in experiments to understand the performance of the model with varying degrees of granularity for emotional state information. To recall, the COMET model generated 72 unique emotional states whereas the GoEmotions model predicted 23 with a high bias toward "neutral" labels. To evaluate the impact of the model knowing if the prompt asked for a positive or negative event, a sentiment-augmented utterance was created by appending "positive" or "negative" to the described event in the same way as creating the ESI-augmented utterance. Penultimately, the self-reported values were added to the message.

Message: I got to confirm my baby daughter.

Sentiment: I got to confirm my baby daughter. *positive*

GoEmotions: I got to confirm my baby daughter. *neutral*

xReact (n=1): I got to confirm my baby daughter. *happy*

xReact (n=5): I got to confirm my baby daughter. *good proud excited satisfied happy*

Oracle: I got to confirm my baby daughter. *happy and proud*

Examples 5.2: Examples of ESI-Augmented utterances appending different ESI method output (in italics) for the same message

The BART and COMET models were then trained on the ESI-augmented utterances using the same procedure as in the other experiments and evaluated (Table 5.4).

Table 5.4: Language model performance on variants of the ESI-augmented utterances

Method	Precision	Recall	F1
BART	0.457	0.467	0.462
BART + GoEmotions	0.457	0.463	0.460 (-Δ0.002)
BART + $\text{Sentiment}_{\text{oracle}}$	0.474	0.479	0.476 (+Δ0.014)
BART + xReact (n=1)	0.492	0.496	0.494 (+Δ0.032)
BART + xReact (n=5)	**0.525**	**0.521**	**0.523 (+Δ0.061)**
COMET	0.535	0.543	0.539
COMET + GoEmotions	0.521	0.520	0.521 (-Δ0.018)
COMET + $\text{Sentiment}_{\text{oracle}}$	0.547	0.556	0.552 (+Δ0.013)
COMET + xReact (n=1)	0.545	**0.558**	0.551 (+Δ0.012)
COMET + xReact (n=5)	**0.553**	0.554	**0.554** (+Δ0.015)

The addition of the top-*n* emotional state inference from the COMET model to the utterance improved BART language model F1 score between 6.9% (*n*=1) and 13.2% (*n*=5). Surprisingly, COMET's inferences also improved its own F1 score by 2.8% in what could be referred to as a self-augmentation process. The COMET model with the ESI-augmented utterance performed 20% better than BART, the model from which it was fine-tuned. This indicates that emotional state inference is latently captured by the COMET model and that fine-tuning on the emotional state inferences in ATOMIC accounts for the majority (79%) of the improvement seen by using the hidden state of COMET over that of its base model, BART.

Appending the GoEmotion predictions had a slight negative impact on model performance likely due to the higher percentage of neutral responses (60%) than all other emotion categories (Figure 5.6) with the next closest, disappointment (6%). This phenomenon could be viewed as a result of the GoEmotions dataset leading to more ambiguity, as well as greater likelihood of an utterance being labeled as *neutral* with third-party annotators than self-reported data.

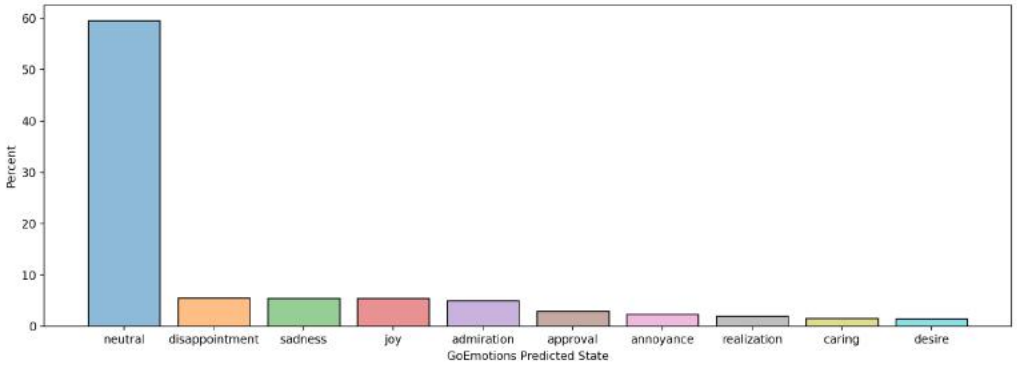


Figure 5.6: GoEmotion predicted emotion categories

The sentiment experiment gives additional data points as context for the results of the primary experiment and hypothesis. These results indicate that knowing the ground truth sentiment (*positive* for the

prompt "Tell me about a positive experience that you had today" or *negative* for the prompt "Tell me about an experience that could have been better for you today") improves F1 score in BART by 3% and COMET by 2.4%. This is less than a quarter of the improvement of adding *xReact* labels to BART (13.2%), but approximately equivalent to the improvement of adding *xReact* labels to COMET (2.8%). This shows a much stronger effect of adding xReact label predictions from COMET than the Oracle sentiment, which is the best that could be expected from a sentiment analysis component.

Based on these results, it would be reasonable to assume that the inclusion of the more expressive and presumably more accurate self-report data would lead to similar gains in model performance. However, The results of the oracle evaluation indicated no improvement in empathic response prediction when using the self-reported emotional states of the user for both BART and COMET models (Table 5.5).

Table 5.5: Oracle evaluation by using self-reported emotional state information

Method	**Precision**	**Recall**	**F1**
BART + Oracle	0.464 (+Δ0.007)	0.465 (-Δ0.002)	0.464 (+Δ0.002)
COMET + Oracle	0.521 (-Δ0.014)	0.530 (-Δ0.013)	0.525 (-Δ0.014)

This is a surprising result that requires further research that is outside the scope of the current work. One potential explanation is due to the team labeling the responses not having access to the self-reported emotional states and thus incorrectly assigning a response label to the utterance. To explore this further, I calculated the negative and positive examples for each response label and present those with the highest entropy in Figure 5.7. These response labels could provide a partial explanation for the results in Table 5.5, in that its conceivable that there either was not enough information in the event or that the person's reaction to the event was unexpected in comparison to other events grouped together in that response label.

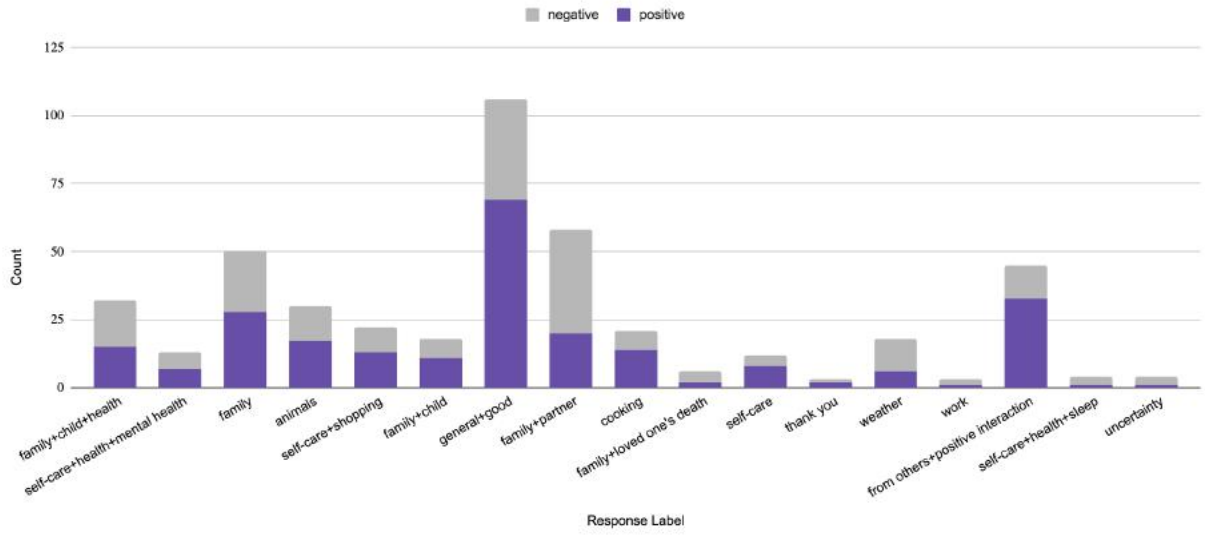


Figure 5.7: Response labels with the highest entropy in sentiment

User:	I went to the park with family	(Event)
	[*Worried* / *Happy*]	(Self-Report / Predicted Emotional State)
Bot:	Spending time with family can make us feel	
	more connected with our loved ones.	(Response)
	[*Family + Family Time*]	(Response Label)

Example 5.3: Illustration of the type of error which may contribute to lack of improvement from appending the self-reported emotional states (Oracle)

As shown in Example 5.3, the self-reported emotion likely takes into account other environmental and personal factors that contribute to their feelings of worry in relation to going to the park with their family. This

illustrates the limitations of attempting to "mind-read" emotional states. The present work uses emotional state inference to inform empathy prediction in direct response to the message, but the recommendation is to ask the person how they actually felt. So a follow-up to the empathic response would be to ask "how did that make you feel?". If the answer is contrary to the expected emotion type, this can also be an indication of a potential cognitive distortion or area to focus the session. In this example, the care provider or automated system can follow up on why the patient is worried and identify ways to help them worry less.

Another potential explanation would be that the emotional state inference system predicts emotional states more consistently across participants experiencing similar events which provided a stronger signal to the task-specific TED policy. These could also complement each other in that the model is more capable of predicting a signal and the TED policy is able to map this to the consistent signal of human labelers.

5.3.3 Error Analysis

Confusion Matrices

This section presents the confusion matrices of each model variant across a single split that had the highest performance for the BART model. Reviewing the differences between the confusion matrices (Figures 5.8-5.10) indicates that by adding ESI the BART model was able to better differentiate between otherwise topically related response labels, e.g. *family + child + spending time together* and *family + child + struggling* or *family + family conflict* (Table 5.6).

Table 5.6: Examples for a sample of responses that share a top-level topic

Event	Response Label	Response
We had a visit with our family (for the first time since COVID started).	family + family time	Spending time with family can make us feel more connected with our loved ones.
Taking kids to the tide pools	family + child + spend time together	Sounds like you had a good time with your kids. Spending quality time together is a great way to strengthen family bonds.
My son had a nightmare diaper and I am exhausted from cleaning him, the house and myself.	family + child + struggling	That sounds difficult. Kids bring us joy but sometimes can be hard to deal with.
Family argument about watching kids while working	family + conflict	Sounds like you're having some issues with your family. That tension can be hard to deal with on a daily basis.

The confusion matrix is clustered by topic so that shaded regions represent intra-topic cluster agreement, for example there is a shaded region around family topics indicating confusion was centralized to other family topics. Intra-cluster disagreement is less problematic in the case of dialog systems, since various techniques exist to deal with intra-cluster uncertainty, for example, fall-back to a higher level intent or detecting polarity (positive or negative) or valence (neutral, intense). A rule could be written to fall-back to a higher level intent when there is high entropy in the predicted probability distribution of the model.

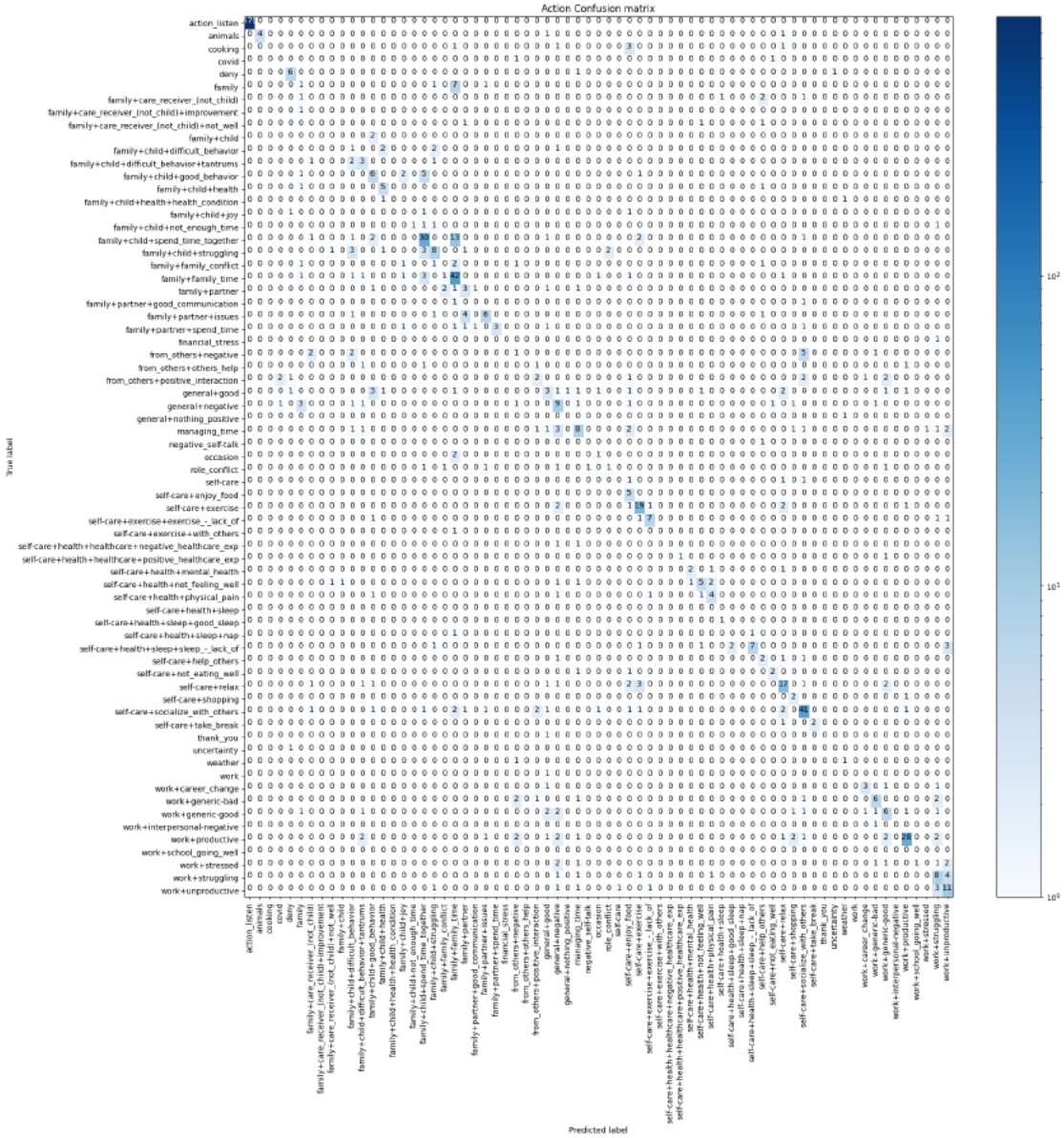


Figure 5.8: Confusion Matrix for BART on un-augmented utterances (split 3)

However, the BART + ESI model (Figure 5.9) is worse at differentiating between *family + child + struggling* and *family + family time* which shares similar ESI distributions. The model may have overfit the signal from the emotional state distribution of these two since it is different from the norm and the dataset size is limited.

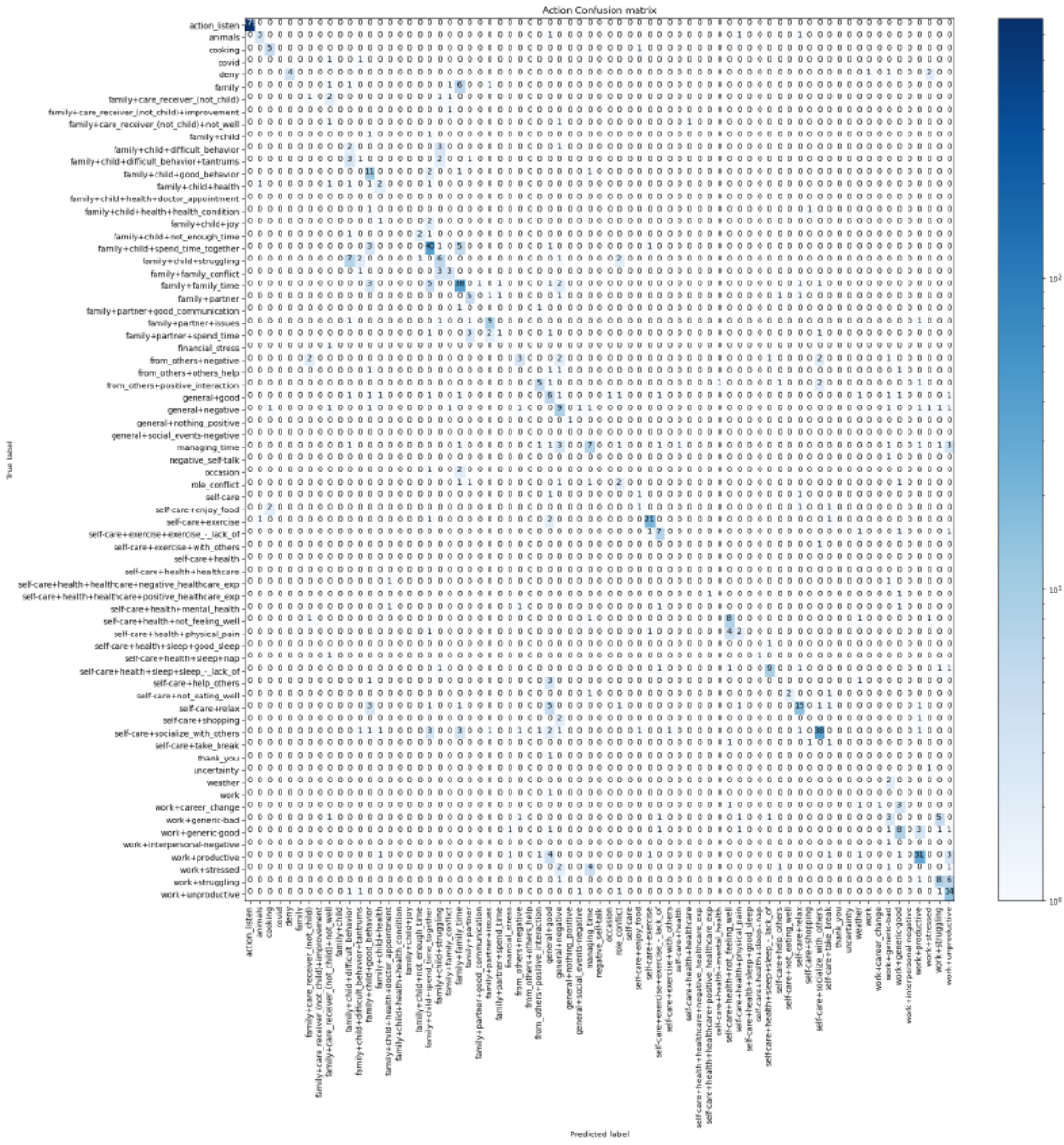


Figure 5.9: Confusion Matrix for BART + ESI on un-augmented utterances (split 3)

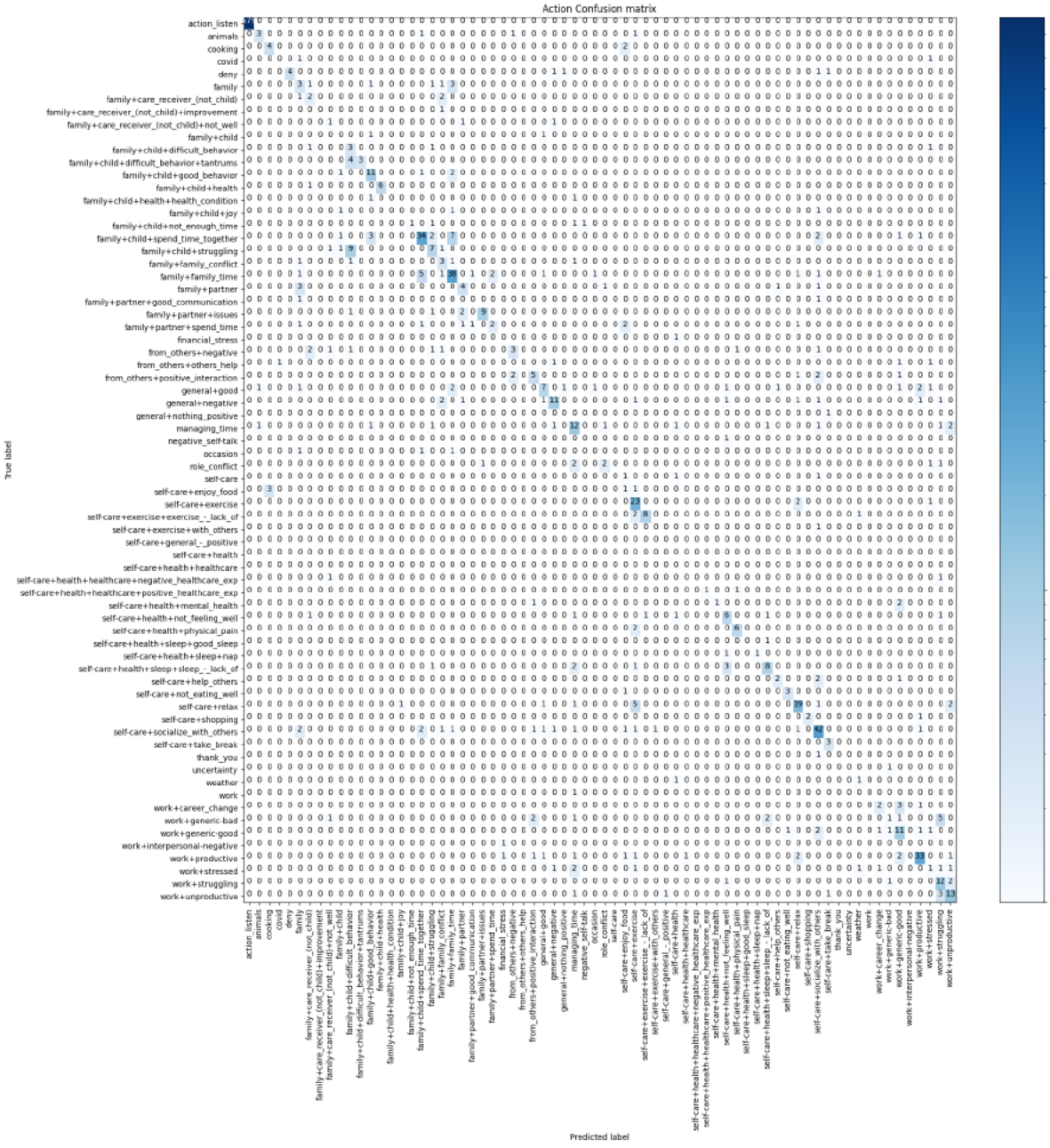


Figure 5.10: Confusion Matrix for COMET on un-augmented utterances (split 3)

ESI Probing

I explore how emotional state inference may contribute to the differences noted between the confusion matrices by looking at the emotional state inferences of the COMET model on the examples in the same split and plotting the distribution of emotional states across topically related response labels. The ESI distributions where COMET outperforms BART tend to be skewed toward specific emotion categories providing a strong signal, whereas those where BART outperforms COMET tend toward more generic emotion categories (e.g. happy) or spread uniformly across multiple emotional states such that little or no signal is present for machine learning. These signals help the model to distinguish between otherwise topically related categories (Figures 5.11 and 5.12).

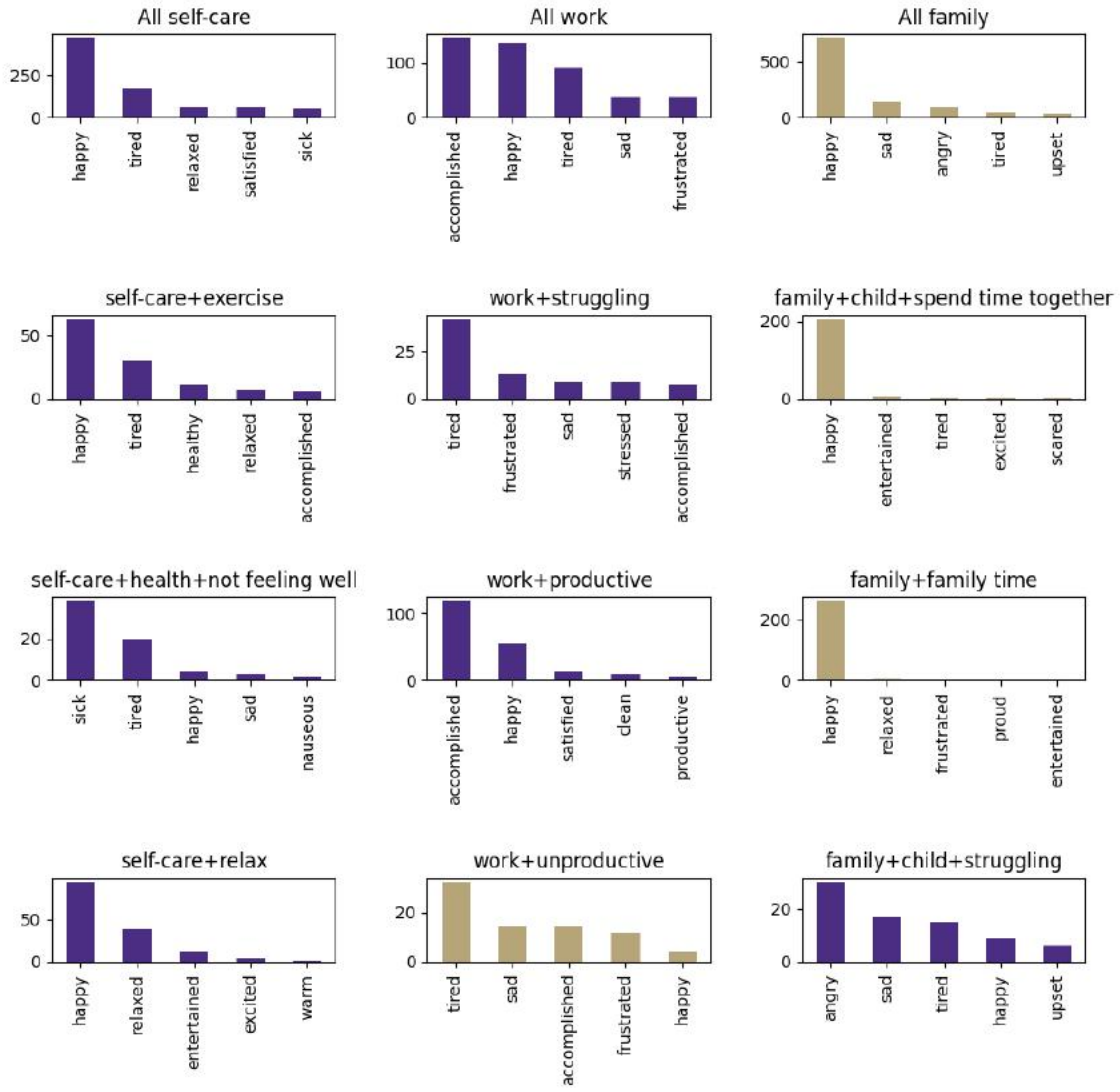


Figure 5.11: Emotional state distributions (n=1) for the top contributors to the F1 difference between the BART and COMET models. Classes where BART outperformed COMET have gold bars and those where COMET outperformed BART have purple bars.

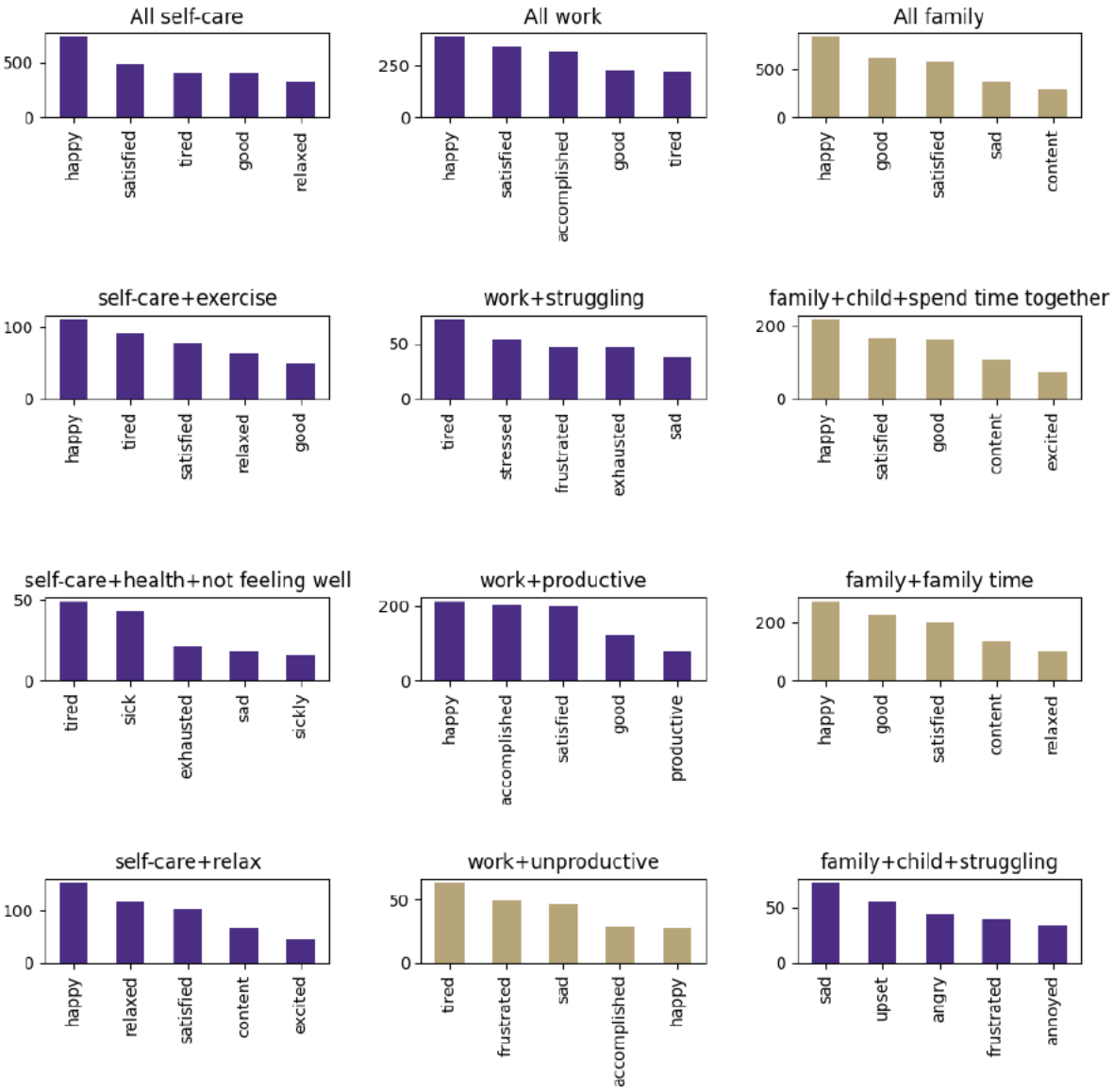


Figure 5.12: Emotional state distributions COMET (n=5) for the labels that most contributed to the F1 difference between the BART and COMET models. Classes where BART outperformed COMET have gold bars and those where COMET outperformed BART have purple bars.

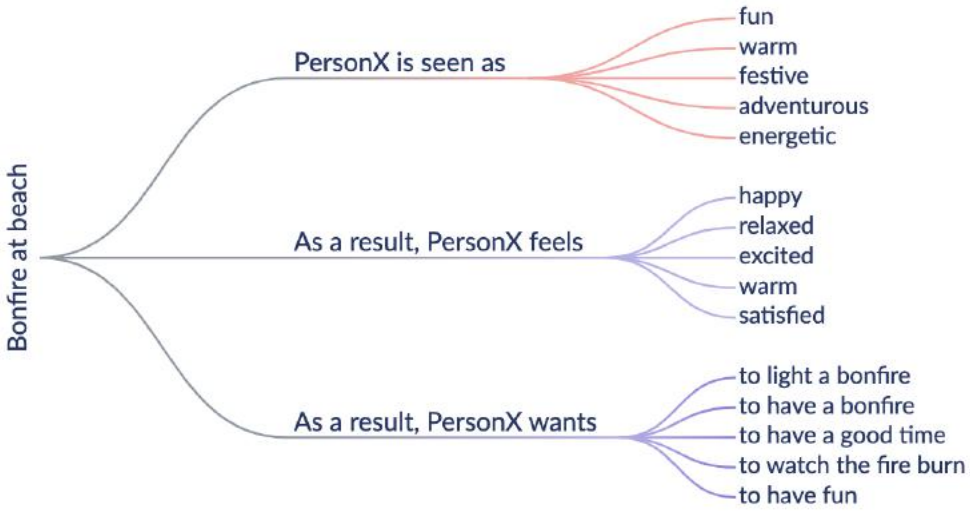
To understand what may be contributing to the remaining difference between BART + ESI_{COMET} and the COMET language model on its own, I present additional inferences made by COMET that were not provided

to the BART + ESI_{COMET} model, including "PersonX is seen as" (xAttr) and "As a result, PersonX wants" (xWant), in addition to the ESI_{COMET} of "As a result, PersonX feels" (xReact).[11]

Utterance: Bonfire at beach

COMET Prediction: self-care+relax (correct)

BART Prediction: self-care+exercise (incorrect)



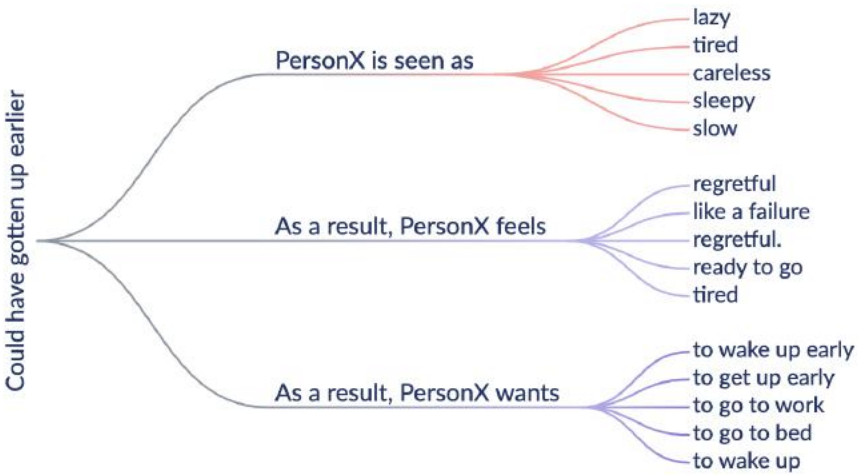
The COMET model inferred that the person feels happy and relaxed and is seen by others as having fun, which likely contributed to its prediction that the person is engaging in a relaxing form of self-care.

Utterance: Could have gotten up earlier

COMET Prediction: managing_time (correct)

BART Prediction: self-care+health+sleep+sleep-lack-of (incorrect)

[11] Generated from the publicly available deployment at https://mosaickg.apps.allenai.org/



The BART model prediction suggests that the user did not get enough sleep, when in reality they got too much sleep. The COMET model on the other hand infers that the user wanted to get up early, feels regretful, and may be seen as lazy, which may contribute to the correct prediction of "managing time".

5.4 Discussion

I have shown that by emulating aspects of how humans empathize, i.e. the addition of emotional state inferences and commonsense reasoning to language models, improves the ability of pretrained language models to retrieve the appropriate empathic response. This was found with COMET, a pre-trained language model fine-tuned on commonsense reasoning data, which outperformed other language models. It was confirmed by the improvement of the BART model through the addition of emotional state inferences predicted by COMET. This finding suggests a path to improve human-AI collaboration systems. Inferences used by humans can be provided to machines to improve AI model performance (Example 5.4) and the AI model can explain why it made predictions by providing the intermediate inferences to the provider. In future work, it may then be

possible for the provider to correct or otherwise modify these inferences to control the empathic response generated by the model.

Client:	I am having trouble being productive.
Provider:	Empathize with this client who feels unproductive.
AI:	Being productive is important for a lot of us. It sounds like you're feeling guilty and maybe frustrated right now.

Example 5.4: Hypothetical interaction between a provider and an AI system to respond emphatically to a client message

This work further validates and extends to the teletherapy setting, emotional state inference, and BART-based models the learnings from Rashkin et al.'s work with the EmpatheticDialogues dataset: the performance of empathic response prediction systems can be improved through the addition of emotion detection to the input text (Rashkin 2019). The computation time scales linearly with the number of emotional state inferences. Since the run-time of the language model to predict emotional states is linear in the number of generations while the beam size is kept constant, it would take ~5 times as long to run inference for the n=5 configuration. Depending on compute capacity, it may be beneficial instead to rely on the hidden states of the COMET language model alone since they capture much of the information latently with only a minor improvement from generating emotional state inferences explicitly.

5.5 Limitations

As is true for annotated datasets generally, the quality can continue to be improved by identifying examples that are ambiguous and disambiguating the labels, e.g. family + family time and family + child + spend time together.

While the emotional state data in the prior chapter were self-reported data, the empathic responses were written by a team of clinicians and how these messages are received likely varies based on the cultural context and individual preferences. So, there remains a need to evaluate the perception of the empathic responses with those who would be interacting with the system.

Further, there are many types of ways to respond to a message and not every message requires an empathic response. The next natural step is to extend these experiments to multi-turn dialog which requires the model to determine when it is appropriate to use empathic responses and when to use other actions including other therapy techniques such as asking open-ended questions or motivating change talk.

5.6 Conclusion

Commonsense reasoning from transformers (COMET) significantly improves empathic response prediction task performance on the daily journaling dataset over widely used language models in the mental health domain (27%) and its base language model (20%). Upon exploration, the mental states inferred by the COMET model appear consistent with human judgment and reasoning required to predict the appropriate response category. By including the emotional state inferences, performance increased by 13.2% for BART and 2.8% for COMET relative to their performance without emotional state inference.

Chapter 6: AI-Assisted Provider Platform Evaluation

The AI-Assisted provider platform is a text-based virtual therapy interface used by a care provider to communicate more efficiently with a client asynchronously. In the work described in this chapter, I used a mixed-methods approach to evaluate the usability, acceptability, and preliminary efficacy of the provider platform and response suggestion feature to support providers in delivering protocolized therapies emphatically. User interviews were conducted to understand perceptions of the technology and areas where this technology may be of immediate use in a healthcare setting. By analyzing these interviews and click data from interactions with the platform, I sought to measure the effects of AI-assistive features on care providers and any differences in these effects between experts and non-experts.

The response suggestion feature is separated into two complementary parts: empathic response suggestions and therapeutic response suggestions. The empathic response suggestion system is the same system described in Chapter 5. The therapeutic response selector recognizes common caregiving symptoms and surfaces these to the operator through symptom slot filling and recommending client goals and solutions.

First, I describe the design and development of the provider platform system starting from a Wizard-of-Oz (WOZ) interface for health dialog collection and transitioning to an AI-augmented teletherapy platform. Next, I describe the specific prototype used for evaluation and the methodology used to collect the data. Then, I describe the statistical analyses used to evaluate the efficiency and empathy of the care providers.

Finally, I present and analyze the results of the study in relation to the hypothesis that the addition of AI-assistance will improve the quality and efficiency of text-based teletherapy.

6.1 Prototype Development

The provider platform was developed from an earlier WOZ prototype for health dialog data collection. This section details the learnings from testing these early prototypes. This section was adapted from the relevant sections of the following work with modifications for clarity:

Kearns, W.R., Kaura, N., Divina, M., Vo, C.V., Si, D., Ward, T., & Yuwen, W. (2020). A Wizard-of-Oz Interface and Persona-based Methodology for Collecting Health Counseling Dialog. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*.

This section is divided into three parts. First, I introduce the motivating clinical scenario and describe the protocol for the intervention. Next, I describe a novel method for health counseling dialog collection using caregiver personas and standardized patients. This section is followed by a description of the WOZ interface.

6.1.1 Intervention Protocol

The protocolized therapy delivered through this system is a self-management intervention based on the Social Problem-Solving Model (D'Zurilla 1999) and Social Cognitive Theory (Bandura1986). The approach of this therapy is to provide caregivers with self-management skills e.g., activation, motivation, and self-efficacy through the Problem-Solving Therapy (PST) process (D'Zurilla 1999). PST gives a global coping process and supports assessing an individual's stressful events and behavioral ability to resolve problems (Toseland1982, Sahler 2002, Malcarne 2019, Teasdale 2021). The research team, with expertise in nursing, psychology, and health

informatics, created a seven-step intervention protocol based on the aforementioned theories and evidence-based treatments (Figure 6.1). As discussed in Chapter 2, the COCO system targets common caregiving symptoms (e.g., fatigue, disturbed sleep, depressive symptoms, and anxiety) through an on-demand health dialog system that provides caregivers with the tools to self-monitor symptoms, problem solve, and take appropriate actions through a series of 4-5 sessions. Skill building is scaffolded in these sessions so that caregivers start with simple goals and solutions before progressing to more complicated ones.

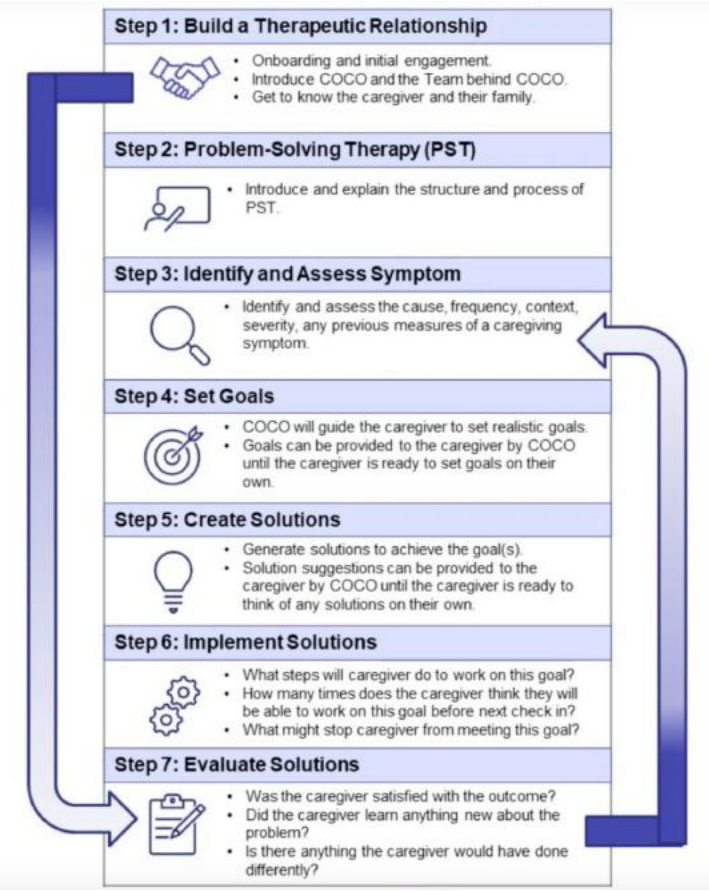


Figure 6.1: Steps of the Problem Solving Therapy intervention

6.1.2 Persona-based Dialog Collection

To collect health dialogs in the context of child caregiving, I collaborated with subject matter experts to assemble a set of caregiver personas based on existing literature on the experiences of caregivers of children with asthma (Belin 2017, Fagnano 2012, Laster 2009), the subject matter expert's prior qualitative and quantitative studies (Yuwen 2016, 2017) and two participatory design sessions with family caregivers of children with asthma. The caregiver personas have a set of characteristics such as age, gender, work type, work hours, main caregiving symptom(s), child characteristics, etc. (Figure 6.2).

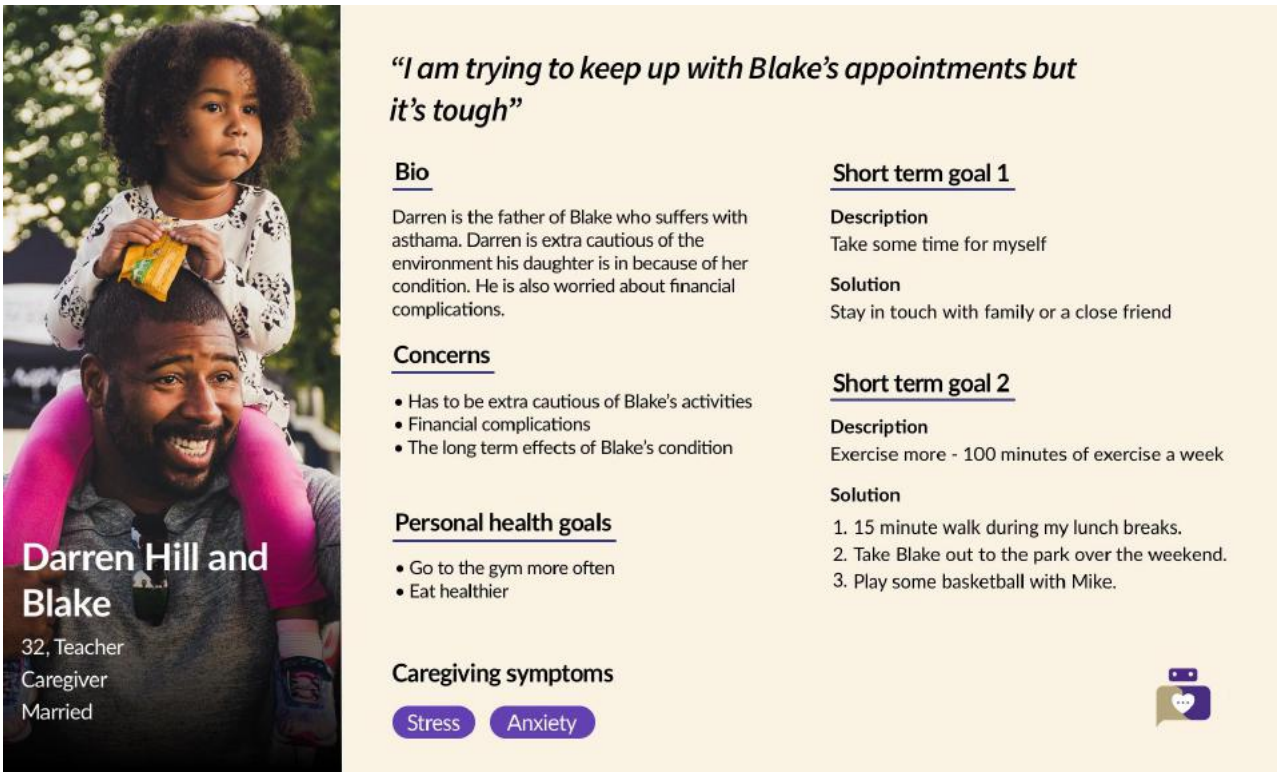


Figure 6.2: Caregiver persona as presented to the standardized patient actor

Standardized patient actors were hired to portray caregiver persona in a five-week program of approximately 10-15 minutes per session. During these sessions, the standardized patient actors and the WOZ operator (the "wizard") worked together to identify, address, and measure caregiving symptoms.

6.1.3 System Description

The goal of the WOZ interface was to augment the wizard's ability to quickly respond and label the standardized patient personas' mental state. This was achieved by mental state inference predictions and empathic and therapeutic responses that draw on values stored in a conversational state tracker, a component that maintains a memory of slots (e.g. *wants*, *needs*, *symptoms*, *ratings*) recognized by the system or input by the operator.

The interface is divided into three sections (Figure 6.3):

1) Chat Interface (leftmost panel)
2) Affective Grounding (tab labeled "GROUNDING RESPONSE")
3) Therapeutic Response Predictor (rightmost panel, labeled "RESPONSE CANDIDATES")

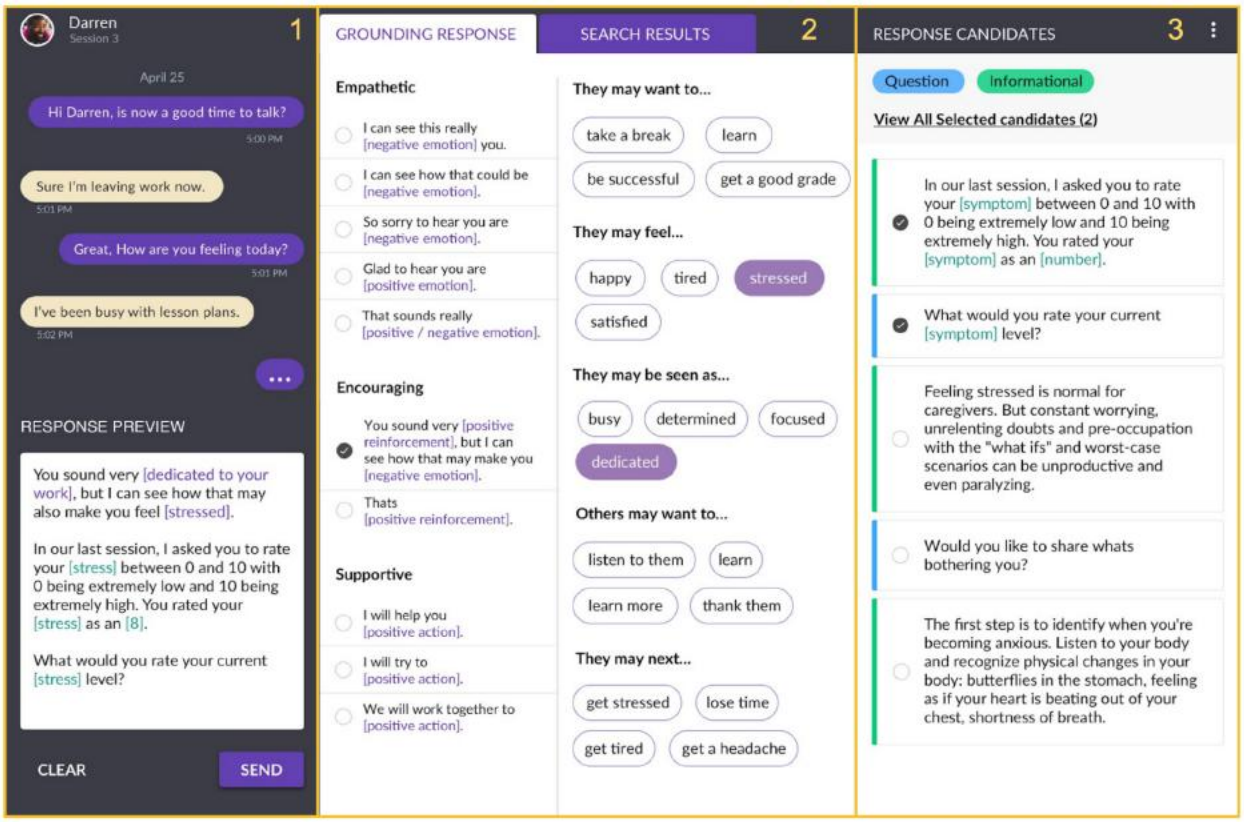


Figure 6.3: A screenshot from the Wizard-of-Oz interface during a Problem Solving Therapy dialog.

Chat Interface

The chat interface is a component of both the WOZ interface and a standalone client for participants. The Wizard can stage multiple responses within the response window to prevent interruption or a shift in the topic of conversation prior to the completion of their turn. Whenever the Wizard submits their response, a snapshot of the interface state is saved alongside the dialog.

Affective Grounding

Grounding is the action of seeking to establish a shared understanding with an interlocutor by communicating understanding to reach common ground, a task fundamental to models of social communication (Clark 1991, Cahn 1999). To ensure the accuracy of its predictions, the system must either explicitly ask for confirmation

(explicit grounding) from the patient or reference its predictions in the construction of an empathic response with the expectation that the patient will correct any errors (implicit grounding). To standardize grounding responses from the interface operators, they were provided with a set of templates that each contained slots to be filled as they interacted with the family caregiver.

The operator is first asked to select a template based on their clinical understanding of an appropriate affective response. The operator can then either select from a set of predicted keywords (e.g. “stressed”, “dedicated”) generated by a commonsense reasoning model conditioned on the described events or input a novel value for each slot (e.g. “negative emotion”, “positive reinforcement”). Finally, the operator is asked to edit their selected slots to fit grammatically and colloquially into the template.

Therapeutic Response Predictor

A database of response candidates related to the intervention was constructed from a seed set of dialogs between the clinical team members. They each generated 4-5 sessions of dialog according to the PST protocol based on an assigned caregiver persona. After the dialog act and slots of each response were annotated, the team reviewed the dialogs together to cross-validate and reach consensus on which series of dialog acts were appropriate in each context. Lastly, the therapeutic response candidates were adjusted to express the persona of the bot consistently.

A ranked list of responses was predicted by a mixture of experts for dialog management that combined transfer learning using conversational representations from transformers (Henderson 2019), a state-of-the-art response selection model, with a frame-based finite-state dialog model. The latter is suitable for PST dialog due to its reliable structure and focus on form filling. These two methods were chosen to improve the performance of the response selector early in the training process.

6.1.4 Usability Testing

The platform was tested by an outside team of designers who recruited care providers to use the system. I used a Tobi Pro eye tracker during the studies to capture participant gaze information to understand where they focused their attention within the interface and for how long (Figure 6.4). Unfortunately, eye tracking could no longer be performed due to pandemic related safety procedures. So the data is limited to the early prototype that has a radically different design than the final system. The eye tracking data indicates that the participants primarily focused on the last message in the dialog history and focused on the area directly near the chat window including the text box and empathetic responses. When they spent time viewing the response candidates they tended to spend more time viewing the top ranked responses. Not much time was spent engaging with the task of providing emotional state annotations.

Figure 6.4: Eye tracking results for WOZ system

This led to v2 of the interface and not presenting so much information to the provider, e.g. hiding the emotional state information and using the COMET predictions without further tuning. The key takeaways from the rapid prototyping process are presented in the next section.

6.1.6 Key Takeaways

Stay-at-home orders disproportionately affected those with caregiving responsibilities increasing their stress. To address this need, we started enrolling family caregivers and redesigned the system to support Human-AI collaboration rather than interactive training of health dialog systems. To make this shift, the system went through rapid prototyping phases that resulted in learnings that led to system improvements:

- Whereas it was expected that dialogs of care providers could be used to train an empathic system directly, high-quality reflections were sparse leading to inefficient data collection. In fact, care providers commonly used variants of simple affirmation such as “I’m sorry” or “That’s great” to caregivers disclosing information which indicated the methods presented in Chapter 5 could be used to assist humans in addition to machines in practicing empathy.

- Eye tracking data indicates that care providers were primarily focused on the last message of the conversation, the text input box, and the simple grounding responses. For this reason, the provider platform included tabs that auto-advance after selection, rather than the three-column layout of the WOZ interface to focus perception and accelerate decision-making.

- Although common in WOZ methodology, not informing participants they are speaking to a human was counterproductive as it set the expectation with standardized patients that they were interacting

with an automated system contributing to psychological stress on the care providers who felt pressure to provide quick responses. This factor coupled with an unfamiliar interface led to cognitive overload, which reduced data quality and led to sparse mental state inference labels since the operators were primarily focused on providing timely responses.

- Care providers had high variance in the structure of their sessions which resulted in pathways that were not represented in the initial training data lowering model performance and increasing response times. This led to the development of a visual PST step tracker that care providers could use to control the step of the dialog to reorient the system and locate their next intended responses.

6.2 System Description of the Provider Platform Prototype

Building upon the learnings from the WOZ system, the technology was redesigned into a more user-centered provider platform and evaluated for efficiency with scripted virtual patients. A simplified version was used for this study to focus the evaluation on the effect of the response suggestion system (Figure 6.5).

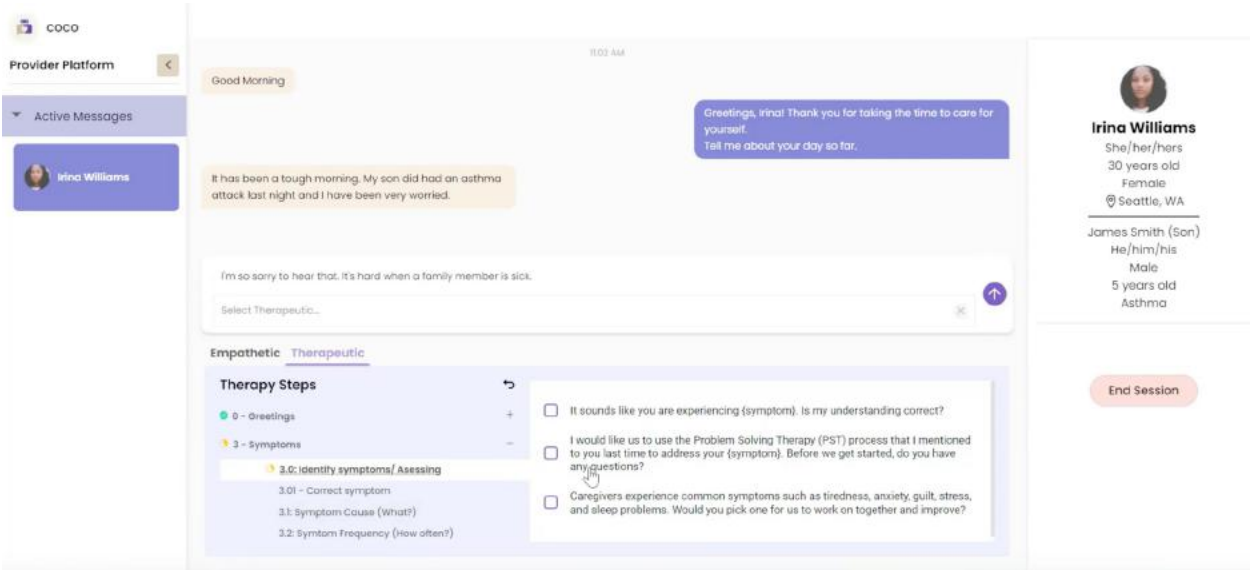


Figure 6.5: Study version of the provider platform

6.2.1 Active Messages

To maximize the number of messages to which a care provider can respond requires they communicate with multiple clients simultaneously to utilize the downtime while waiting for each client to respond. To support multi-client workflows, a message queue was added with the ability to switch between conversations while saving state information for each conversation.

6.2.2 Client Profile

The client profile was added to the right panel to assist the care provider in maintaining context between sessions, context switches, and to surface information collected by automated check-ins. An end session button was added to mark a session as complete, removing it from the message queue.

6.2.1 Conversational State Tracker

A natural language understanding component was developed and incorporated the language model features described in Chapter 4 as a custom language model featurizer in a Rasa NLU pipeline to recognize common caregiving symptoms and store these in a dialog state tracker. These are used to fill response templates with the appropriate slots, e.g. *name, time of day, emotion*, *symptom*, *goal*, or *problem*. When recognized, these are stored in the client record and stored in a knowledge graph. The slots are used to recommend *goal* and *solution* options.

Utterance: I kept thinking about my child having an asthma attack.

...

Template: Earlier you mentioned that you were [symptom]

Response: Earlier you mentioned that you were worried.

When a care provider orients themselves to the context of a particular session they are provided with a checklist of problem solving therapy steps depending on the session number (Table 6.1). A checkmark serves as a visual indicator of which steps have been completed and which are remaining. The current dialog step is set through a finite state automaton that sequentially follows the checklist. At any time, the care provider can select a therapy step from the checklist to bring up all responses for that step. The interface sends requests for the currently selected step within the checklist, which is fulfilled by the conversational AI system which retrieves the templates for that step and fills the slots within the template using the conversational state tracker as outlined above.

6.2.2 Problem-Solving Therapy Steps

The COCO care team developed the following checklist for a five-session PST protocol (Table 6.1). The first session focuses on identifying one of the caregiver's symptoms, what problems it's causing, and developing a potential solution to the problem that the member will try out for that week. The remaining sessions follow the same pattern of checking in on how a solution is working, co-developing a new solution if the current solution is not working, and psychoeducation.

Table 6.1: PST checklist items with order within the checklist by session number

		SESSION	1	2	3	4	5
Problem Solving Therapy							
Step	Sub-Step	Response Label					
0.0		Start/ Greeting	1	1	1	1	1
0.5		Check-in for follow-up session		2	2	2	2
1.0		Therapeutic Relationship (onboarding)					
2.0		Explain the PST Structure	3				
3.0		Identify symptoms/ Assessment	2				
	3.1	Symptom cause (What?)	4				
	3.2	Symptom frequency (How often?)	5				
	3.3	Symptom context (Where and who?)	6				
	3.4	Symptom severity	7				
	3.5	Previous measures	8				

	3.99	Ask more questions					
4.0		Recommending goals for the first time	9				
	4.1	Guidance on setting goals		11	11	11	11
	4.9	More about the goals					
	4.99	Extra information on goals					
5.0		Recommending solutions for the first time	10				
	5.1	Guidance for solutions	11	13	13	13	13
	5.9	More about the solutions					
6.0		Solution Implementation		15	15	15	15
	6.1	What steps will you do to work on this solution?					
	6.2	How many times do you think you will be able to work on this solution before we check in next time?					
	6.3	What are one or two things that might stop you from doing this solution?					
	6.4	What will you do to make sure you do this solution?					
7.0		Re-Evaluate Symptom		3	3	3	3
	7.1	Were you satisfied with the solution?		4	4	4	4
	7.2	Did you learn anything new about the problem?		5	5	5	5
	7.3	Is there anything you would have done differently?		6	6	6	6
	7.39	Check if this solution helps		7	7	7	7
	7.4	Decide to continue or terminate solution		8	8	8	8
	7.9	Confirm to continue the session		9	9	9	9
	7.94	Try more goals		10	10	10	10
	7.95	Request more solutions		12	12	12	12
	7.96	Request more implementation		14	14	14	14
8.0		Ask to setup a reminder	12	16	16	16	16
	8.1	Confirm with the reminder	13	17	17	17	17
9.0		Summarize the session	14	18	18	18	18
	9.1	Ask to setup a follow up session	15	19	19	19	19
	9.2	Confirm with the follow up session	16	20	20	20	20
10.0		Additional Issues					
11.0							
	11.1	Request Confirmation - All General response calls for participants confirmation can be sorted here					
	11.2	Continue the talk - All general connection utterances that try to keep on chatting					
99.0		Goodbye	17	21	21	21	21

6.2.3 Clinical Knowledge-Based Recommendation

I developed a schema for a clinical knowledge graph based on the problem solving therapy process (Figure 6.6). The knowledge graph was constructed by subject matter experts who added and associated 23 goals with 12 caregiving symptoms, and 21 solutions which each have one associated resource. A solution can help with more than one goal and there are 56 connections between these two node types. Similarly, a goal can address more than one symptom so there are 119 connections between these two node types. Caregivers are linked to these symptoms, goals, and solutions within the knowledge graph with these links updated through the PST steps based on values in the conversational state tracker. In this way, the system is capable of making up-to-date recommendations to care providers based on information known about the client from their conversational history.

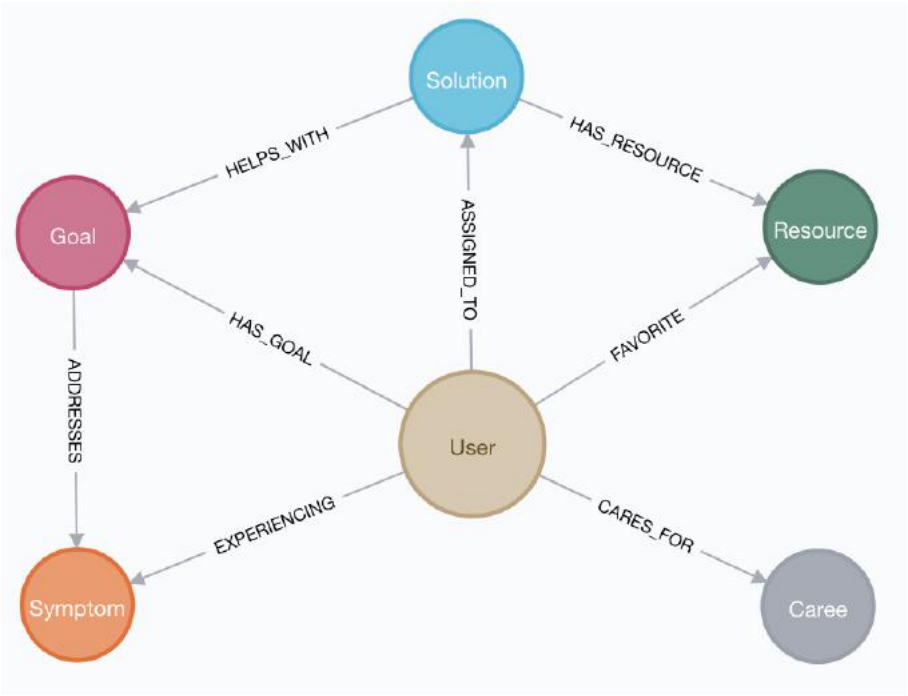


Figure 6.6: Schema of the Knowledge Graph used to store user data for predictions

A profile includes information about the current emotional state and recent events. When a caregiver or care provider indicates that the caregiver is experiencing a particular symptom, this symptom is mappable to a set of goals to address that symptom. Once the caregiver selects a goal, this goal then maps to a number of solutions for achieving that goal and each solution has a set of resources for the caregiver to engage with between visits.

Each node on the graph is an object with associated metadata. An example of each type of node is provided below (Figures 6.7-6.10) in sequence from client ("user") to their symptom, goal selection, solution selection, and resource associated with that solution.

Figure 6.7: An example client ("user") profile

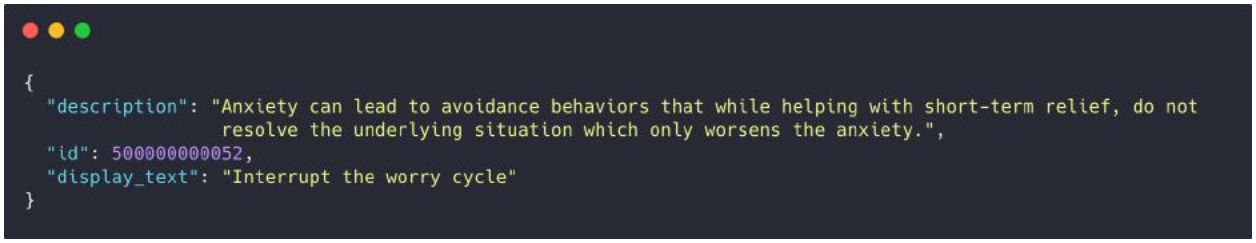


Figure 6.8: An example goal for the client above

```
{
  "how": "Start writing for five to 15 minutes, and write about whatever is on your mind. Keep going until
          you feel you have written what needs to be said but haven't delved into rumination. Describe the
          events that are currently causing difficulties for you. Remember that sometimes it isn't what is
          currently happening that causes stress, but the concerns you have about what could happen.",
  "what": "Keeping a journal or simply writing down what is in your mind and how that makes you feel.",
  "cost": "FREE",
  "supplies": ["Notebook", "Pen"],
  "name": "Write down your feelings",
  "why": "Numerous studies have demonstrated the effectiveness of writing down your feelings for health,
          happiness, and stress management. It's not just a simple technique but an enjoyable one.",
  "description": "Keeping a journal or simply writing down the things you are thankful for can be
                  helpful.",
  "locations": ["WORK", "HOME"],
  "id": 600000000005
}
```


Figure 6.9: An example solution for the goal above


```
{
  "long_text": "Worries, doubts, and anxieties are a normal part of life. It's natural to worry about an
               unpaid bill, an upcoming job interview, or a first date. But "normal" worry becomes
               excessive when it's persistent and uncontrollable.",
  "file_type": "HTML",
  "short_text": "Are you plagued by constant worries and anxious thoughts? These tips can help calm your
                worried mind and ease anxiety.",
  "id": 700000000055,
  "title": "How to Stop Worrying",
  "url": "https://www.helpguide.org/articles/anxiety/how-to-stop-worrying.htm",
  "tags": ['self-care']
}
```


Figure 6.10: An example resource for the solution above

6.3 Study Design

I implemented a 2x2 study design to compare performance between clinical and sub-clinical care providers; whereby, each participant completed one session without response suggestion and one with response suggestion. The order of these two conditions was randomized to control for practice effects in the evaluation. Both sessions were treated as the first session of PST.

Participants were invited to join an online session via video telecommunications software. Once the participant joined the session, the moderator followed a script (Appendix F) to have the participant complete the following steps:

1. Tutorial
2. Session 1
3. Session 2
4. Exit Interview
5. Exit Survey

Each session consists of an exchange between a research team member following the script of a virtual patient with the role of a family caregiver and a study participant with varying levels of behavioral health knowledge in the role of the operator (Appendix G).

6.3.1 Study Recruitment

The study was exempted as human subjects research by the UW IRB (STUDY00013541). The study team sent invitation emails to listservs (e.g., University of Washington (UW) Medicine Psychiatry trainee listserv, UW Tacoma RN-BSN program student listserv) and snowball sampling. In total, 29 care providers responded to the screening survey and 20 care providers completed the study. Of these, 11 were clinical psychologists or psychiatrists and 9 were nurses without mental health experience. All study participants met the following inclusion criteria:

Inclusion Criteria

- Self-described English proficiency
- Internet access (as the study was conducted during the COVID-19 pandemic)
- Additional criteria for specific groups:
    - Novice Group: Registered Nurses *without* extensive mental health training, or the experience of having worked in a behavioral/psych unit for more than 6 months
    - Expert Group: Clinical Psychology or Psychiatry residents at UW

These criteria were checked using an intake-survey and verified by study team members who grouped the participants into two groups (either "expert" or "novice") based on their role and experience in mental health. The exact questions asked during the initial intake survey are given in Appendix H and the group characteristics collected in response to the survey are presented in the quantitative analysis (§6.6.0).

6.5 Methods

Excerpts from the script are used in the subsections below to illustrate the preferred operator behavior.

6.5.1 Therapeutic Responses and Symptom Identification

At each step of the dialog, therapeutic responses were suggested to the operator to advance through the PST steps. During certain steps as part of the therapeutic response, the operator must identify the symptom that the caregiver is experiencing and indicate this understanding to the caregiver.

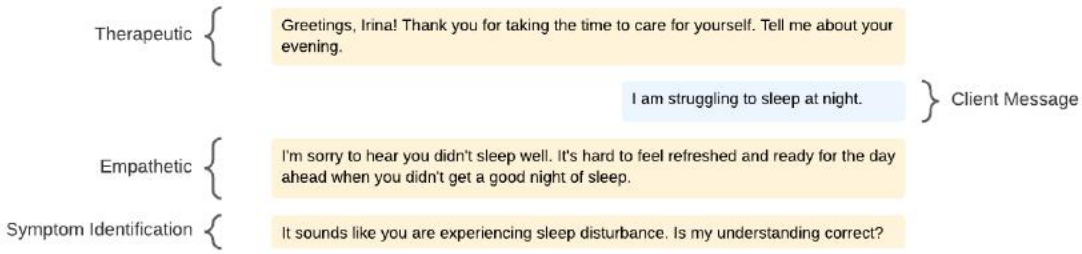


Figure 6.11: Excerpt from the sleep disturbance session showing the symptom identification step and giving an example of how empathic responses are paired with therapeutic responses within a single care provider turn of dialog.

6.5.2 Empathic Responses

Additionally within each session, there were two opportunities for the participant to provide a high-empathy response (after Step 0.0 and after Step 3.1). These sections lend themselves to more open-ended responses that in turn are likely to reference events that have occurred in the client's environment. The gold label responses for both scenarios are provided below:

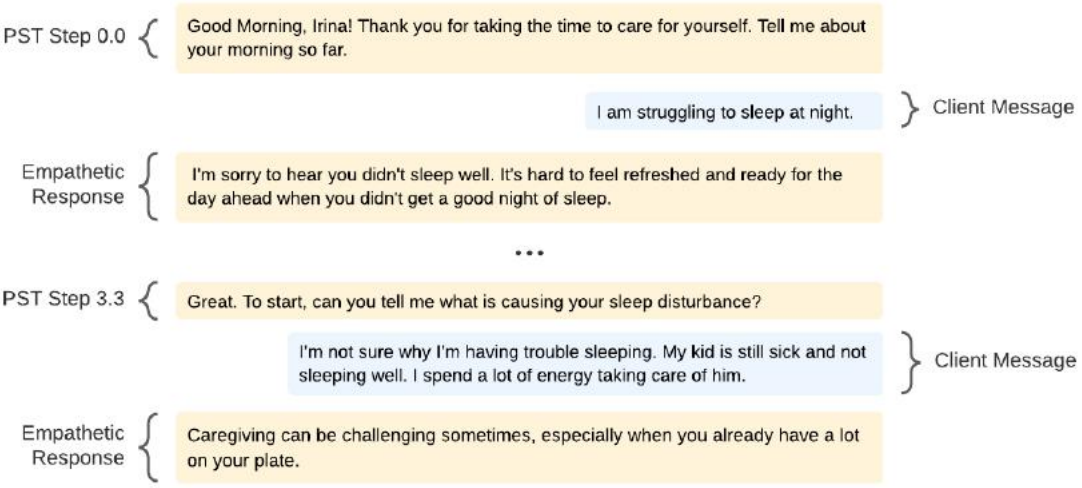


Figure 6.12: Excerpts from the sleep disturbance session showing the two opportunities for empathic responses

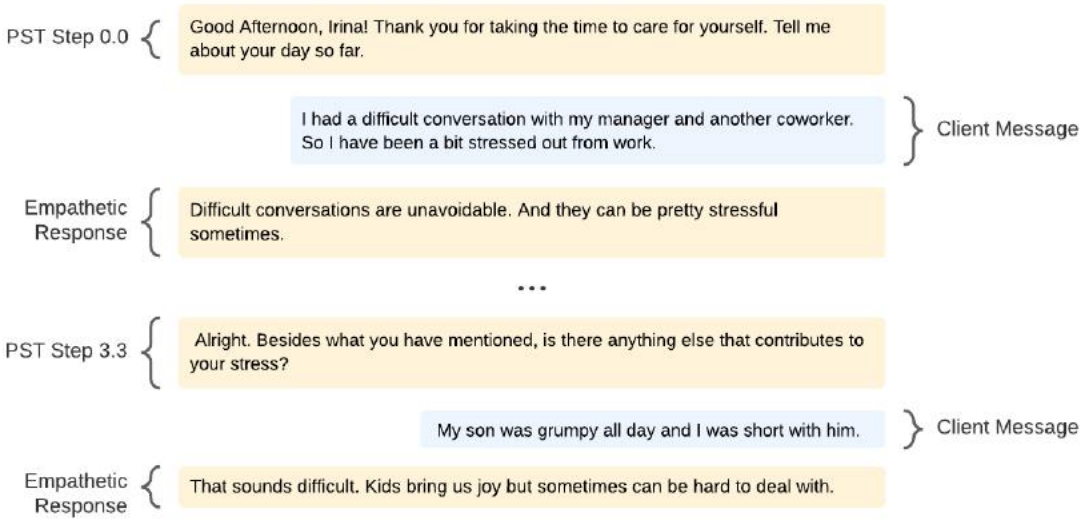


Figure 6.13: Excerpts from the stress session showing the two opportunities for empathic responses

For the remainder of the turns the system suggested simple responses, e.g. "Got it.", which were less context dependent and thus interchangeable so were not evaluated for accuracy or speed.

6.5.3 Goal Selection

In the control group, participants were asked to select goals for the client based on their symptoms (Step 4.0). They were presented with five options, one for each symptom from the knowledge graph that is exclusive to that symptom. In the intervention condition, the system provided only one set of goals based on the symptom under discussion.

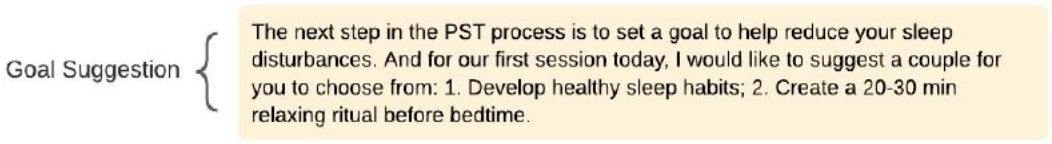


Figure 6.14: Goal suggestion for sleep disturbance

6.6 Quantitative Analysis

This section presents the quantitative analysis results of the AI-recommendation system (intervention) and a baseline system that presents all response options (control). Due to the data retention policy of the communications vendor, data was only stored for seven days and then automatically deleted for nine participants unbeknownst to the research team. After discovery of this limitation, an offline backup procedure was put in place to preserve the data of the remaining twenty participants which was then used for the following analysis.

The quantitative evaluation of the selected responses was based on four distinct hypotheses. It is hypothesized that care providers will make:

(H1) Faster responses when presented with a ranked list of responses than a random list that contains the correct answer.

(H2) Faster and more accurate *empathic responses* when presented with a ranked list of empathic responses than a random list that contains the correct answer.

(H3) More accurate *goal recommendations* when presented with goals that match the client's goal than when presented with all goal choices.

(H4) More accurate *predictions of a caregiver's symptoms* when presented with a ranked list of symptoms than with an unranked list of symptoms.

Each hypothesis was evaluated using permutation testing to compare either the intervention and control conditions or the group differences between experts and non-experts. The effect sizes between groups and conditions were calculated using Cohen's D.

6.6.0 Group Characteristics

Participants were asked to complete an initial survey on the perceived usefulness of the COCO platform and caregiver burnout. The following demographic information was collected as part of the initial survey and used to screen participants for the evaluation study in accordance with the inclusion and exclusion criteria (see 6.3.1):

1. Please provide your current status at UW. If you are at a different institution, please select "other" and share your institution name and your education (e.g., BS, MS)
2. Please share your age in years
3. Please provide your current field of study
4. How many years have you been a nurse working directly with patients and families?
5. Do you have prior mental health training beyond what was included in your undergraduate nursing education?
6. Please explain your mental health training (e.g., through what program and for how long?)
7. Have you worked in a psychiatric/behavioral/mental health setting for more than 6 months in total?
8. What is your current training level?

The nursing group (n=9) was composed of a majority of nurses with more than 5 years of experience (n=6), two nurses with 3-5 years of experience, and one nurse with less than 1 year of experience working directly with patients and families (Figure 6.15).

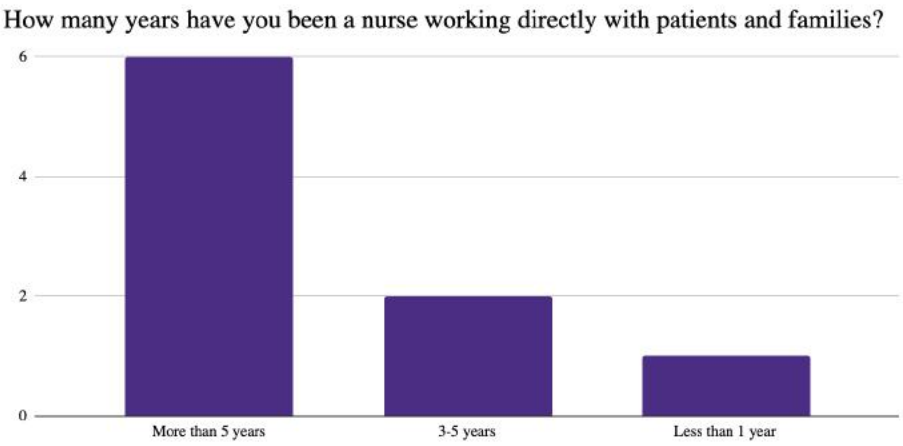


Figure 6.15: Distribution of nursing experience for non-expert group

Clinical psychologists and psychiatrists also tended to be more experienced in their field with the majority (n=6) in their third year of residency or beyond (Figure 6.16).

Figure 6.16: Distribution of clinical experience for expert group with training levels R1-R5 relating to years of residency

6.6.1 Overall Response Time

Response times were measured through the interface and used to evaluate **H1**. The response times averaged across all conversation turns in the session are normally distributed across participants in both the control and intervention groups (Figure 6.17). A permutation test was used to determine the statistical significance of the reduction (29.34%) in response time between the intervention and control conditions (Table 6.2).

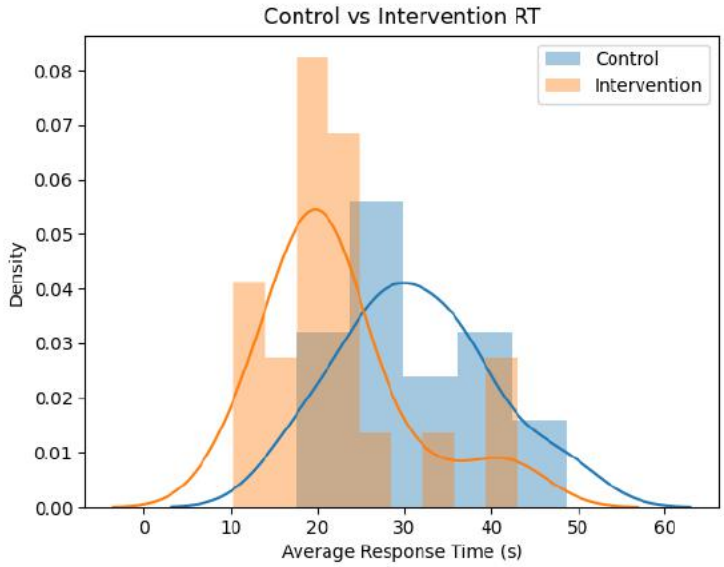


Figure 6.17: Probability distribution plot of average response times in seconds

Table 6.2: Permutation test results for average response times per message across all participants in the novice, expert, and combined groups

	All (n=20)	Novice (n=9)	Expert (n=11)
Intervention Average RT (s)	22.0894	21.548	22.532
Control Average RT (s)	31.2625	32.148	30.5378
Reduction	29.34%	32.97%	26.22%
p-value	0.00202	0.02748	0.04931
Cohen's D	1.08	1.21	0.92

The reduction in average response time between the intervention and control conditions was found to be statistically significant ($p = 0.002$) with a large effect size ($d = 1.08$). Applying the system reduced average response times by 29.34%.

6.6.2 Relative Response Time Reduction

As shown in Table 6.2 and Figure 6.18, non-experts had a greater mean reduction in response time than experts to the point of surpassing expert response times. However, this difference in relative reduction between the two groups was not significant ($p=0.577$).

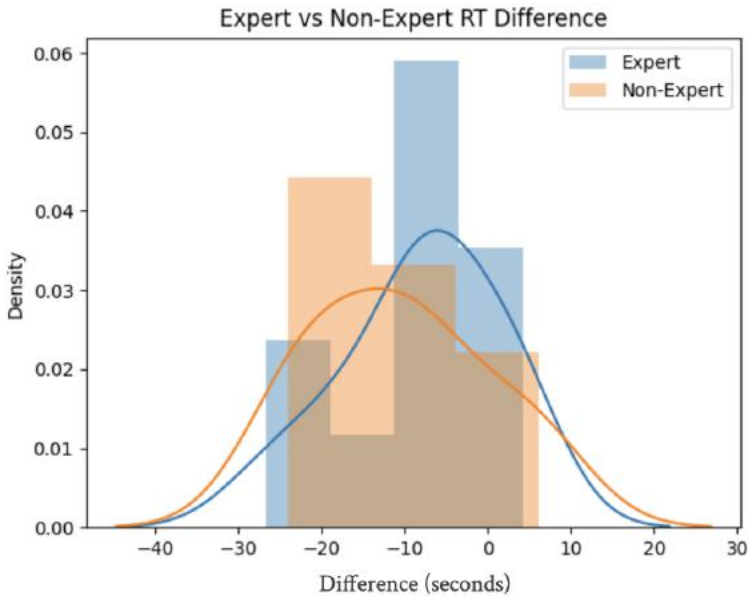


Figure 6.18: Probability distribution plot of the average difference in response times (in seconds) between the intervention and control for experts and non-experts.

6.6.3 Empathic Response Time

The response time for the two turns which included high-empathy responses were compared between the intervention and control conditions to evaluate the speed portion of **H2**. The response time distributions are presented (Figure 6.19).

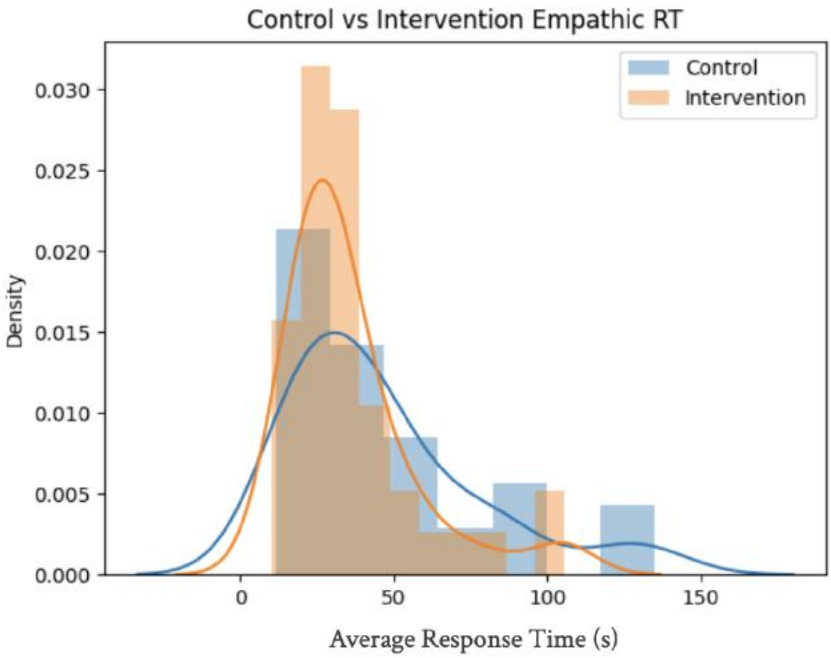


Figure 6.19: Probability distribution plot of the average empathic response times per participant

A permutation test was used to determine the statistical significance of the decrease in response time (23.09%) between the intervention and control (Table 6.3). The reduction in response time was not found to be statistically significant (p=0.082) between the intervention and control conditions. This may in part be due to the operators quickly selecting simple responses such as "I'm sorry to hear that" and not searching for higher-empathy responses.

Table 6.3: Permutation test results for empathic response time

	All (n=20)	Novice (n=9)	Expert (n=11)
Intervention Average	35.7964	30.359	40.245
Control Average	46.5417	43.036	49.409
Decrease	23.09%	29.46%	18.55%
p-value	0.08173	0.08896	0.33938
Cohen's D	0.40	0.59	0.30

6.6.4 Relative Empathic Response Time Reduction

On average, the novice group decreased their response times by 29.46%, while the expert group decreased their response times by 18.55% (Table 6.3, Figure 6.20). However, the difference in relative reduction between these two groups was not statistically significant (p=0.751).

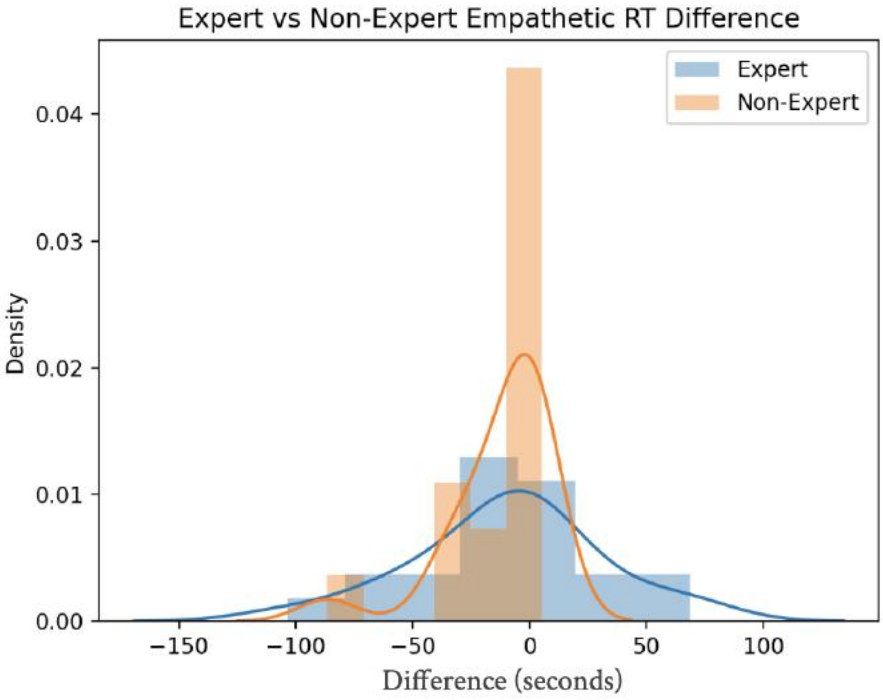


Figure 6.20: Probability distribution plot of the average difference in empathic response times between the control and intervention conditions for experts and non-experts

6.6.5 Empathic Response Accuracy

During two steps of the session, participants were prompted to provide an empathic reflection. The number of correct reflections were counted and presented in the contingency table below. Based on a Fisher's Exact Test, the association between the AI-intervention (intervention vs control) and number of correct empathic responses (0, 1, 2) as defined in **H2** was found to be statistically significant (Table 6.4).

Table 6.4: Contingency Table for Empathic Response Accuracy

	Control	**Intervention**
Zero correct	12	2
One correct	7	9
Two correct	1	9

p-value	0.000952

The most common empathic response suggestions are presented in Tables 6.5 and 6.6 to provide further context into the differences in performance between the intervention and control groups.

Table 6.5: Control Responses

Count	**Empathic Response**
18	I'm sorry to hear that.
4	I'm sorry to hear you haven't been sleeping well. It's hard to feel refreshed and ready for the day ahead when you didn't get a good night of sleep.
3	That sounds difficult. I'm sorry you had to go through that.
3	That sounds difficult. Kids bring us joy but sometimes can be hard to deal with.
2	Work can be very stressful and overwhelming. Taking care of ourselves is especially important right now.
1	It's hard to not be worried when a loved one is not doing well. I imagine that worry impacts your ability to fall asleep at night.
1	Difficult conversations are unavoidable. And they can be pretty stressful sometimes.
1	I'm sorry that you're not sleeping well. It's hard to enjoy life when you're not feeling your best.
1	I'm sorrry to hear that. You must be very worried. about your son.
1	I'm sorry to hear that your child is still sick.
1	I'm sorry to hear that.That sounds difficult. Kids bring us joy but sometimes can be hard to deal with.
1	Caregiving can be challenging sometimes, especially when you already have a lot on your plate.
1	Seems like you're feeling stretched too thin with demands from different areas of your life. That can be exhausting.
1	I am here to help.
1	I'm sorry to hear you haven't been sleeping well. It's hard to sleep well when your child is struggling.

Table 6.6: Intervention Responses

Count	Empathic Response
9	That sounds difficult. Kids bring us joy but sometimes can be hard to deal with.
8	I'm sorry to hear you haven't been sleeping well. It's hard to feel refreshed and ready for the day ahead when you didn't get a good night of sleep.
8	Caregiving can be challenging sometimes, especially when you already have a lot on your plate.
6	Work can be very stressful and overwhelming. Taking care of ourselves is especially important right now.
4	Difficult conversations are unavoidable. And they can be pretty stressful sometimes.
1	Difficult conversations are unavoidable, on top of that work can be very stressful and overwhelming. Taking care of ourselves is especially important right now.
1	I hear you. That sounds difficult.
1	I'm sorry to hear that.
1	That sounds difficult. I'm sorry you had to go through that.
1	It's hard to not be worried when a loved one is not doing well.

6.6.7 Goal Selection Accuracy

The novice group selected the correct response 66% of the time whereas the expert group selected the correct response 54.5% of the time. Moderator notes on this phenomenon point to experts having strong opinions against certain goals selected by the panel that created the knowledge graph, e.g. journaling. This indicates that care recommendations for providers are not one-size-fits-all and would require personalization without strong evidence of efficacy present to override clinician intuition. These results show statistical significance (p=0.001) supporting **H3** (Table 6.7).

Table 6.7: Permutation test for goal selection accuracy

	All (n=20)	Novice (n=9)	Expert (n=11)
Intervention Average	1	1	1
Control Average	0.6	0.6666666667	0.54545
Increase	66.67%	50.00%	83.33%
p-value	0.00129	0.0798	0.01361
Cohen's D	1.13	0.94	1.23

6.6.8 Symptom Identification Accuracy

There was no variability between the intervention and control groups in their ability to identify the symptom experienced by the caregiver in each scenario. For this reason, this experiment failed to support **H4**.

6.6.9 System Usability Score

The exit survey found that the system had excellent usability with a system usability score of 79.5 (Bangor 2019) with full survey data provided in Appendix I. Of the positively-framed questions from the system usability survey, participants ranked highest that users could learn the platform quickly (4.3, between agreement and strong agreement) and ranked lowest that they would use the platform frequently (3.6, between neutral and agreement). Of the negatively-framed questions, participants ranked all the questions relatively the same (1.75, between disagreement and strong disagreement).

For the statement, *"I needed to learn a lot of things before I could get going with the COCO Provider Platform."*, nurses provided an average score of 2.3 (between disagreement and neutral), whereas clinical psychologists provided an average score of 1.1 (indicating strong disagreement). This may be an indication that the nurses were less familiar with the delivery of problem-solving therapy, which is to be expected by the inclusion criteria into expert and non-expert groups.

Similarly, for the statement, *"I think I would need the support of a technical person to use the COCO Provider Platform."*, nurses provided an average score of 2.1 (indicating disagreement), whereas clinical psychologists provided an average score of 1.4 (between strong disagreement and disagreement). This may indicate that the clinical psychologist cohort was more accustomed to the use of technology in care than the nursing cohort.

Scores for the statement, *"I think I would like to use the COCO Provider Platform frequently."*, decreased with the intensity of care provided by each type of provider, with nurses having an average score of 4.0

(indicating agreement), clinical psychologists having an average score of 3.4 (between neutral and agreement), and psychiatrists having an average score of 3.2 (between neutral and agreement).

6.7 Qualitative Analysis

Following the session, the moderator asked participants the following questions:

1. What's your overall impression of the platform?
    a. What (else) do you like about this platform?
    b. What (else) do you not like about the platform?
2. You interacted with a different system for each session, an AI-driven version and a version without AI. You might have noticed that in one of the sessions, the responses were recommended by AI on the top, and in the other one, the responses were listed in an unordered response bank. What do you think about the AI feature?
3. How might a tool like this fit within your workflow or that of a colleague?
4. What did you think of the empathetic responses?
5. What did you think of the therapeutic responses?
6. Towards the end, you were prompted to choose some health goals for the caregiver. What did you think of that function?
7. What would you recommend us change or add to the platform?

Participants were asked if the exit interview could be recorded, based on their response the answers to the questions were either transcribed through a transcription service or taken as notes by the moderator or a notetaker. From these transcripts and notes, I coded the responses through deductive content analysis (Elo &

Kyngäs 2008) based on the categories of interest (Table 6.8). The codes were iteratively refined through peer debriefing to arrive at the final set.

Table 6.8: Categories from qualitative coding

Category	Subcategory	Definition	Example
Workflow	Delivery Method	How do participants view text-based delivered care fitting in their workflow	*I like the idea of it being really accessible for caregivers. Not everyone has time to schedule a phone call or privacy to process what's going on. So I think that having it be in a text form could be really beneficial because I feel like the barrier to entry is low for someone who's in distress to like get some level of support.*
Workflow	End Users	Who might it be best suited to support	*I guess when I think about primary care folks and community health workers and other kinds of folks, I guess thinking about that, maybe it would be too overwhelming to have the full list. I don't know. So having fewer options might actually be better.*
Workflow	Scenarios	Where might this be applied in care delivery	*Would be good for check-ins for clients between sessions, maybe for those that are at higher risk.*
AI	Efficiency	Reports of reduction or increase in response time	*I definitely felt like it was quicker to find a response, which then in my mind would leave more time to tailor the message to the individual.*
AI	Acceptability and Trust	Descriptions of acceptability and trust in the AI system	*Even though there are automatic responses, if they are a AI-driven, they would probably take everything into account, including what I'm going to look up.*
UX	Usability	Descriptions of the usability of the system	*It seems like something that I could be working with other providers that maybe don't need the same level of training I have on all the different therapies and different drugs. I mean, I feel like you could have other folks working on this where if it got to a sticking point, they could be referred to me, certainly. But it would be wonderful to have some staff being able to do this for caregivers in a clinic that I worked at. And that would be*

			a really nice service to have 20... actually really 24/7.
UX	Updates	Mentions of the checklist, response tabs, and other updates from the WOZ prototype learnings	*I like that they're separated, so you're able to give an empathetic response to help the patient feel heard. And then I like that there was a therapeutic response so that you can intervene and offer a solution or a help to what they're complaining or having issues with.*
General	Improvements	Suggestions from the participant	*Yeah, I think that would be nice just because, yeah, I would want it to sound like my own voice as a therapist. And so, I'm sure that there are little utterances that I use a ton that I could program in, and that could be helpful if you find yourself going back to saying some of the same things over and over.*

6.7.1 Workflow

A key question this work seeks to answer is where this technology is applicable within care delivery. The majority of the answers to this question came when care providers were asked where they thought this technology may fit within their workflow or that of a colleague. These responses were coded into three subcategories: delivery method, end users, and scenarios (Table 6.8).

Delivery Method

Accessibility was a major theme of the care providers. They indicated that text-based treatment could provide a private channel for caregivers to communicate asynchronously whenever they are available and lower the barrier to entry.

> **Clinical Psychologist:** "*I like the idea of it being really accessible for caregivers. Not everyone has time to schedule a phone call or privacy to process what's going on. So I think that having it be in a text form could*

be really beneficial because I feel like the barrier to entry is low for someone who's in distress to like get some level of support."

Psychiatrist: "*I think with these tools, it's always exciting, right, because we don't do a very good job, right, of doing any of this stuff, taking care of caregivers or sometimes just people that we can't reach. So I think if it's implementable, then that's cool. And if it can be actually used in the community, that would be really great. And everyone has smartphones now, right, so."*

Providers noted that the asynchronicity could also fit better within their workflow as they are moving in between patient visits and would not have time to

Clinical Psychologist: "*I think we're often quite busy running around from patient to patient and administrative item to administrative item. And so I don't know the feasibility of having a full, continuous session like I did with Irina in one sitting, but it would allow me to send these messages maybe in between sessions with more expediency, which would be nice. So I think it could fit in that way."*

Nurse: *"Okay. So let's see that. I feel like where I'm working now at the plasma center, I feel like it would actually be very helpful if the people on the other end know how it works because there would be situations where someone who wants to donate, they might say, "Oh. I'm on diabetic..." or they have a certain medical condition that's complicated. And I would have to look things up. It's better than being on the phone and me telling them, "Okay. I'm going to put you on hold. I'm going to look you up. I'll be right back," because I might be taken away to do another duty."*

A clinical psychologist saw this intervention as useful in the early stages of therapy but underscored the importance of escalating intense cases to human-expert care as part of longer-term interventions. An asynchronous communication channel over text was seen as saving time when contacting patients.

End Users

Participants listed many use cases for this technology including follow-up for patients in "high-risk scenarios", pediatrics, and oncology care which are family-focused, and used by subclinical staff to follow up with patients.

> **Psychiatrist:** "*When I think about primary care folks and community health workers and other kinds of folks, [...] having fewer options might actually be better."*

> **Psychiatrist:** "*The current challenge to [having staff follow-up with patients and caregivers] is just staffing and continuity, right? And maybe this tool actually helps, right, because it makes it so that you don't necessarily need the same person running the algorithm from the person side. So maybe it's actually kind of helpful because right now, the situation is... When I walk into a clinic, [..] I don't know if there will be staff or how many staff there will be. And I think there's a loss of cohesion just because of job stuff and work stuff, so maybe this helps bridge some of that I think potentially for caregivers."*

> **Nurse:** ***"****For our transplant patients, those caregivers are with them 24/7, and they literally have to be with them at all times."*

Check-ins between visits and screenings were suggested by four participants. Participants mentioned screening at primary care and use by other care professionals that are not trained in providing therapy for example staff at a clinic or a warmline for people to text to talk when needed based on their work experience in addiction and recovery.

> **Clinical Psychologist: "***Because obviously, yeah, because parents are so busy, and I think that there's something... Implementation is really something I found is difficult of like, okay, you know the skill, you know what to do in the moment. That's why like phone coaching is so cool for DBT. So, this feels like a phone coaching kind of alternative that you could probably even get someone who's not even your therapist to do."*

The system was viewed by clinicians as part of a larger system where the user of the system would likely be care providers who would require mental health training which would increase the reach of mental health services and access to 24/7:

> **Clinical Psychologist:**
>
> *"I mean, I felt like I was learning it just by doing it. [...] It could even be like an adjunct thing for primary care, which could be really cool, of just like an opportunity for people who aren't actually more connected in typical therapy context."*

> **Psychiatrist: "***It seems like something that I could be working with other providers that maybe don't need the same level of training I have on all the different therapies and different drugs. [...] But it would be*

wonderful to have some staff being able to do this for caregivers in a clinic that I worked at. And that would be a really nice service to have [...] 24/7."

Psychiatrist: "*I used to run a warmline, which was for people for addiction and recovery needs, and they would just call to talk about whatever. It wasn't for ER stuff or crises. But it was such a great thing for people to have just available as needed. And the vast majority of all those calls didn't need to go any farther. It would be an awesome service for a clinic to have that."*

6.7.2 Artificial Intelligence

In response to open-ended questions that asked for opinions on the AI-driven responses, participants spoke to two top sub-categories, efficiency and acceptability. Efficiency related to the speed at which they were able to respond to the messages and acceptance related to their trust in the system as a means to deliver care to patients.

Efficiency

Time savings was seen as one of the largest benefits of the system. Shortening the time required for care providers to respond was seen as benefiting the patient by increasing the time available to tailor the message to the individual and increasing the patient response time to shorten the session duration. The status quo in some cases is "not getting a response for a day or two later."

Trust in the relevance of system-suggested responses was indicated as a factor in the speed at which they were able to respond. This points to a link between model performance and efficiency improvement. Familiarity with the response bank was indicated as another factor in reducing the time required to respond to messages (reducing reading time and improving trust).

Acceptability

A clinical psychologist who did not use text-based messaging with patients indicated a desire to use the software to increase email response speed, “especially to crises”, and reported delaying responding to messages due to “having to think of the right thing to say.” Other providers echoed this sentiment:

> **Nurse:** *“I like giving us a foundation for expressing concerns or empathy. It helps us stay professional sometimes when emotions are high.”*

Others shared their concern of using a response recommendation system without the AI feature to ensure the slots in templates were accurately filled. This example response indicates the provider's fear to select the wrong response from the full list of responses, and feeling safer knowing the AI had screened the responses:

> **Clinical Psychologist:** *“And I really did like the AI feature. I think it would be pretty invalidating to accidentally select a response that's pretty off. So I think, obviously there's a flexibility of altering the responses with text, but I really liked that the AI kind of felt like it screened for certain keywords. Then I felt safer that I was going to pick something that was closer to what the client was saying rather than like, "Oh, you're stressed. Oh, what if I accidentally said that you feel guilty?" Then I'm really going to feel like it's a pre-filled response and I'm going to be like, "What am I doing?"”*

One provider asked for an AI assistant similar to “Clippy”, an assistant featured in earlier versions of Microsoft Word that helped improve writing, within their workspace and another asked for pop-ups to prevent mistakes:

> **Clinical Psychologist:** "*As a beginner, it could be useful to have some helpful popups. I'm imagining the little paper clip for Microsoft Word that's like, "You are taking a long time to answer. Do you need some help?" You're taking a while to remember that you can edit if none of these feel good"*

Another care provider shared feedback they had heard from patients that similar systems made them feel like they were not treated as an individual and the responses felt canned.

6.7.3 User Experience

This work represents an initial pilot of this technology and can present learnings for future iterations of this type of assistive technology. Two types of feedback are presented in this session: 1) general feedback on factors impacting usability and 2) feedback specific to updates made based on user interviews and usability issues addressed when updating the design from the WOZ prototype.

Usability

Care providers found the interface familiar to other patient communication systems they had used and commented on the ease of use.

> **Psychiatrist: "***I thought it was pretty straightforward to use in terms of the way it was laid out. It was pretty intuitive, what to click next, how to proceed, what the options were. I think just visually and in terms of working with it, it was fairly intuitive, unlike a lot of things on my work EMR."*

Nurse: "*I think especially nowadays, everyone mostly has a smartphone. So I think like the closer, those features replicate that experience, especially with the little ellipses I think is really helpful too. [...] Does that show the patient on the provider side?"*

Updates

The updates made to the interface to add a controllable checklist of problem solving therapy steps was well received by care providers.

Psychiatrist: "*And I liked, again, how there was the problem solving therapy structure was built into it in a way that seemed to make sense for somebody who's never done problem solving therapy to work through those."*

Psychiatrist: "*So I'm thinking about if you need to go backwards or if you're skipping a step for some reason, it would be nice to be able to skip around once you're familiar with the process. So if that's available, then I think that would be helpful."*

The participants indicated a desire for personalization of the responses and indicated appreciation for the inclusion of the slots within the templated responses, e.g. "Good [time of day], [name]!" → "Good Morning, Irina!".

Nurse: "*I like how personalized they were, that they had like, hello, good afternoon depending on what kind of day. It just seems more personalized. It doesn't sound like a generic response, besides if you just use greetings."*

Although they were not a large part of the evaluation, the addition of the "Active Messages Queue" and the "Client Profile" information to the interface was appreciated:

> **Clinical Psychologist: "***I liked that it had the homepage like built into the columns on the side. So I would see multiple clients on the left and then whoever was talking to, it had relevant details about them just on the right, without having to click into anything else. I appreciate that."*

The participants also appreciated the tabs that separated the "empathetic" and "therapeutic" response types. They also found that they automatically switched between the tabs when a response was selected.

6.7.4 General

This subsection shares general feedback and suggestions for improvements which would be worth exploring in future work as added features.

Improvements

Care providers indicated a need for additional responses and shared a desire to personalize the response bank to include their own variations to so the patient would not feel as if they were talking with a bot:

> **Psychiatrist:** *"If I was typing in my own responses, if it would save them for me so that there's a quicker way to do that. And again, that would just make me feel more natural."*

There was also a desire for additional types of empathic responses including affirmations and validations:

Nurse: *"I think maybe having a response that affirms like they're trying to do something or that maybe if they haven't been able to address their issue to be something about life being really busy and you're dealing with this stress."*

Clinical Psychologist: "*It's kind of like a bridging, like "I hear you and I feel what you're saying. I understand what you're saying." Or like, "Man you're so right. That sounds like that would really be hard." I just feel like that's something we do a lot in therapy is validate the valid, right? Like, "Man, it makes sense you're so tired. You have a lot on your plate.""*

To incorporate these into the system, it was recommended to divide them into subcategories to make them easier to choose, as care providers found it difficult when presented with many choices. It was also suggested by several participants to remove response suggestions that had been used in prior sessions, so that the patient did not feel like the responses were canned.

Psychiatrist: "*If it's going to be the long list of things, dividing it more so we can choose, and then making sure that it's not repetitive."*

6.8 Discussion

Quantitative results indicate that response times are reduced and the accuracy of empathic response selection is increased in both non-experts and experts in mental health delivery by using an AI-assistive interface confirming H1 (§6.6.1-2) and the empathic accuracy component of H2 (§6.6.3-5). They also indicated that the technology

had excellent usability (SUS=79.5). Exit interviews validated these results showing that care providers perceived the technology as improving their efficiency and accuracy as well as being easy to use. Additionally, the results indicate that goal selection accuracy was improved through the AI system validating H3 (§6.6.7). However, there was no improvement in symptom identification (H4 §6.6.8). This is likely due to the simplicity of the symptom identification task due to two factors: 1) the symptoms of *fatigue*, *stress*, *anxiety*, *sleep disturbance*, and *grief* are easily distinguishable, 2) the scripts used for the virtual patients did not include language at the boundary of closely related symptoms, e.g. *fatigue* and *sleep disturbance*.

In the initial development of the platform for WOZ data collection, the hypothesis was that empathic response data could be efficiently collected by providers interacting with standardized patients. However, providers would often use generic phrases such as "sorry to hear that", which indicated a need to support providers in delivering empathic responses also. The results from the provider platform evaluation indicate that without AI-intervention providers continued to use simple responses such as "I'm sorry to hear that" (the top response). Whereas, the AI-intervention increased the use of higher empathy responses.

Qualitative feedback provided several insights as to provider perceptions of the technology that indicate future directions for the improvement and application of similar systems. Participants in the expert group supported the system being used by individuals without mental health training to complete check-ins or low-intensity sessions with clients. This recommendation is supported by the quantitative result that the quality gap was closed between experts and non-experts using the system in the AI-condition. Both experts and non-experts reported learning during the process, indicating an opportunity for the system as a training tool. One nuance related to acceptability was that providers were concerned about sounding robotic and wanted the opportunity to personalize the response pool to sound more like themselves. By reducing cognitive load and the time required to respond to messages, providers felt they would have more time to personalize their messages for the patient.

6.9 Limitations

The prototype evaluated in this study used responses that were written by study team members for a bot-delivered intervention. They were not collected from natural therapy dialog, and at times were perceived as robotic and impersonal. Unevaluated methods to increase naturalness include increasing the response pool to include previously used responses and selecting from each of the top-*k* predicted labels (maximizing recall) ranked based on the perplexity of the response given the last utterance, or alternatively prompting generative models with the dialog state (including ESI and other relevant information from the knowledge graph) and the response retrieved by the system to generate more fluent responses.

6.10 Conclusion

Care providers who used the system found the system to be intuitive and were more empathic in their message selection and responded faster to virtual patient messages. Most found the technology acceptable for use with family caregivers and by individuals without mental health training. Several suggested that the system could be used to train health workers without therapy training to deliver problem-solving therapy. Together these results show promise for incorporating AI-assistive technology into conversations with patients.

Chapter 7: Summary

The central hypothesis of this work is that both the quality and efficiency of text-based telehealth can be improved through recent advances in conversational AI. This hypothesis was evaluated with three aims: (Aim 1) explored the ability of computational methods to infer high-fidelity representations of emotional states as a precursor to empathy, (Aim 2) evaluated these representations as features for a transformer-based empathic response predictor, (Aim 3) piloted this system as a component of a teletherapy platform for the delivery of problem-solving therapy by nurses and psychologists. The results of these aims validate this core hypothesis by successfully collecting emotional health information through an automated SMS-based intervention and by significantly improving empathic accuracy and reducing response times of human care providers using an AI-augmented chat interface.

Together the components of this dissertation (Figure 7.1) provide a unified solution that can help to increase access to mental health care by automating the remote monitoring of emotional health, expanding the number of individuals who can provide protocolized care, and enhancing the efficiency and empathy of the care provided. During the course of this work, I developed a novel evaluation paradigm to better measure how emotion recognition systems can help to track emotional health through automated journaling exercises, applied these measures to predict empathic responses, and evaluated a support tool to assist care providers in delivering problem-solving therapy. I describe each contribution in greater detail in the sections below.

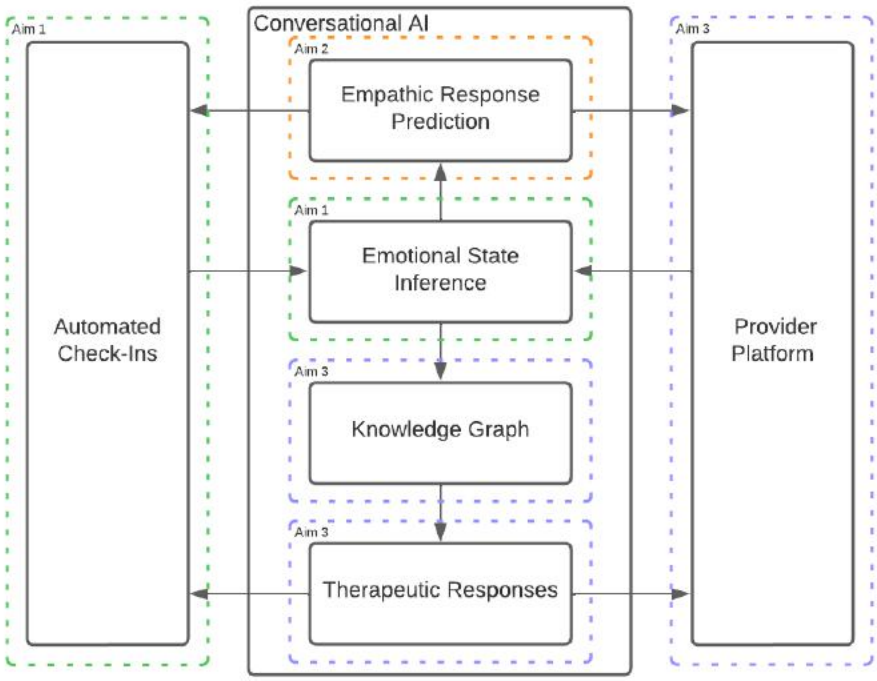


Figure 7.1: Human-AI collaborative system supported by this work with arrows indicating information flow

7.1 Summary of Contributions

Several scientific contributions resulted from the work completed in this dissertation. This section summarizes these for the reader.

C1: Acceptability of EMA delivered by a conversational agent over SMS

One contribution of this work is the demonstration of the acceptability of automated methods of collecting behavioral health data through a chatbot. In response to the COVID-19 global health pandemic, conversational agents played a pivotal role in responding to reports of physical symptoms related to COVID-19. Similarly, the Cora Wellness study (Chapter 3) showed promise in assessing the effects of public policy guidelines on emotional health due to strong positive feedback and high completion rates from participants. Similar feedback was received during the journaling exercise (Chapter 4).

C2: A novel evaluation paradigm and corpus for more granular emotion recognition using self-reported emotions

Prior to this work, emotion recognition from conversation (ERC) methods had only been tested on crowdsourced data labels that (a) were inferred by annotators rather than self-described and (b) did not cover the breadth of emotional states experienced as a result of daily events. This prior work was insufficient to assess the applicability of these methods to identify emotions that approximate self-reported emotional states in the context of automated mood check-ins. To address this gap, I developed a novel paradigm to collect emotional state information for ERC.

I positioned emotional state inference as a similar but distinct task to prior work in emotion detection from text, one that is more relevant to the delivery of care. As discussed previously (Background §2.3), researchers have trained emotion *detection* systems on third-party annotated data leading these models to pick up on surface-level markers of emotion that are salient to the data labelers. In contrast, emotional state *inference* systems derive their predictions from event semantics to reason beyond what is explicitly mentioned in the text. Results on empathic response generation indicate that emotion detection models predict "neutral" too frequently to improve the performance of downstream models on this task. In comparison, the emotional state inferences generated by language models fine-tuned for commonsense reasoning can improve model performance.

The collection of an emotion recognition dataset using self-reported emotional states through an SMS-based intervention presents a future direction for emotion recognition research aligned with the Ecological Momentary Assessment (EMA) methods commonly used in medical research to collect ground truth emotional state information that is scalable, cost-effective, and able to capture greater expressivity than third-party annotation approaches. Experiments comparing utterances augmented with emotional state inference from

COMET to those augmented with GoEmotions predictions or sentiment oracle values demonstrated the value of capturing more emotional nuance than current methods to improve empathic response suggestions. Models with a broader range of emotion labels led to more accurate ERP. In addition, qualitative feedback from participants in the SMS-based interventions showed that people find value in completing daily reflections, which presents an opportunity to collect this information in a way that benefits those providing the data.

C3: Establishing the utility of emotional state inference models for the prediction of self-reported states

I evaluated the relative performance of emotion detection and emotional state inference methods to predict 3,465 *event-emotion* pairs representing 217 self-reported emotions collected from family caregivers. I found that the emotional states generated from COMET, a language model fine-tuned for commonsense reasoning (including social inferences, e.g. the wants, needs, reactions, and how someone is viewed by others as a result of events), were more closely aligned with the self-reported emotional states of the client with an improvement of 73.5% relative to the GoEmotions classifier (an emotion detection model), 57.8% relative to MentalBERT (a domain-specific language model), and 3.2% relative to InstructGPT as measured by the SBERT similarity metric. These improvement gains are extended to the GoEmotions dataset, a publicly available emotion detection dataset of social media posts annotated with 28 emotion categories, where the COMET model improved performance by 23.7% relative to the GoEmotions model, 13% relative to MentalBERT, and 5.1% relative to BART. On the Cora dataset, where the task involves a binary decision between “hope” and “anxiety”, the COMET model improved performance by 11.1% relative to the GoEmotions model, 6.3% relative to MentalBERT, and 4.7% relative to BART. Based on these results, it is clear that fine-tuning language models on commonsense reasoning data can significantly improve model performance on ERC. This finding opens up new

opportunities to automate the detection of emotional states from conversations that are more consistent with self-reports.

C4: Establishing the utility of mental state representations for empathic response prediction

Prior work in empathic response prediction has focused on generation-based methods. In contrast, most dialog systems in production are retrieval-based due to the unpredictability and unknown biases of generative models, which is particularly important in a healthcare setting (Korngiebel 2021). In this work, I developed an approach through which to use mental state inference methods to improve the selection of possible responses to a user utterance from amongst a curated set of responses improving the predictability and safety of the system. Representational choices were shown to impact the performance of predicting appropriate empathic responses. Using the hidden state output of COMET increased weighted-F1 performance by 24.2% relative to MentalBERT and 16.7% relative to BART. Further augmenting the utterance with the top-5 ESIs from COMET increased the weighted-F1 performance of the BART model by 13.2% and the COMET model by 2.8% relative to their performance without ESI. These findings establish the utility of ESI for empathic response prediction, and the subsequent analyses provide insight into the mechanism of action underlying improved performance to inform future work.

C5: A validated approach for AI-augmented teletherapy

Prior work on conversational AI-augmented communication in teletherapy has focused on peer-to-peer communication (Sharma 2022) and post-session feedback (Creed 2022). This is the first system to evaluate the efficiency and accuracy gains of providing empathic and therapeutic response suggestions to care providers.

In this work, I devised, implemented, and evaluated an AI-augmented teletherapy platform in which ESI is used alongside other conversational AI technology to assist therapists in their selection of potential

responses to a client. Nine clinical psychologists/psychiatrists (experts) and eleven nurses (non-experts) completed two problem-solving therapy sessions with a virtual patient whose responses were delivered according to a script by a study team member. The participants were randomly assigned to receive a list of responses in a ranked order suggested by the conversational AI system, or the same list of responses unranked. Quantitative results indicate that the system significantly decreased response times by (+29.34%; p=0.002), tripled empathic response accuracy (+200%; p=0.0001), and increased goal recommendation accuracy (+66.67%; p=0.0013) across both experts and non-experts. Both groups rated the system as having excellent usability (SUS 79.5). Empathic response accuracy increased twice as much in the non-expert group compared to the expert group resulting in equal accuracy in the conversational AI condition. Total response time also decreased more rapidly in the non-expert group removing the difference between groups. Structured qualitative interviews indicate that the participants felt the system would make providing the therapy more effortless, efficient, and accessible to care providers without mental health training. These findings validate the utility of AI-based support for empathic response selection in the context of simulated telehealth sessions and affirm the usability and acceptability of the system concerned.

7.2 Generalizability

Empathy is a skill applicable as much in human-human conversation as in human-bot conversation. Affect-aware conversational intelligence can assist with both, and the system can be deployed as a stand-alone chatbot or human-AI collaborative system. In the early stages of development, it may be beneficial to develop these approaches with human-AI collaboration where the AI system prompts the user with empathic response suggestions, and the provider uses their judgment to choose the most appropriate response and modify it as necessary. Then, as system predictions converge with human selections, the system can be allowed to communicate directly with patients in the scenarios in which it was supervised.

While this dissertation focused on applications in teletherapy, these methods may increase performance in any conversation that would benefit from increased empathy. Indeed, my analysis of the exit interviews of care providers who interacted with the system identified several areas where there may be an immediate need for this technology, including applications within healthcare delivery beyond problem-solving therapy.

Telehealth utilization has grown tremendously during the pandemic; however, this growth has also mostly taken place in urban areas where there is access to broadband internet connections, allowing for an emphasis on live, synchronous video chat. This development has exacerbated the pre-pandemic differences in behavioral health utilization rates between rural and urban communities. Furthermore, because of its focus on text-based messaging, which has minimal data transfer requirements, this work can be applied to regions with limited bandwidth and to populations with connectivity constraints.

Global access to mental health is the primary goal of this line of research. It is now possible to reach nearly everyone on the planet through SMS or USSD (a cost-effective solution for individuals without smartphones and pay-per-message pricing that opens a persistent communication channel for the length of a session). However, developing culturally appropriate interventions will require partnerships with members of the communities the system will serve. Fortunately, through interfaces similar to the AI-augmented provider platform tested in this study, community health workers and peers can be supported in providing protocolized care. Adopters of this system can customize the conversational agent to deliver different assessments, e.g., the patient-facing systems listed in the overview of health dialog systems (Section 2.1) through a configuration file in YAML format[12].

The data the model was trained on comes from family caregivers who on average may discuss a distribution of topics and emotions that differ from the general population. For example, relationship issues may be more prevalent in younger age groups than the population in this study. As a result, the findings of this study

[12] https://yaml.org/

may vary between populations and the model should be retrained if applied to different data distributions. The annotation of empathic response categories was completed by a care team that focused on a taxonomy based on Maslow's hierarchy of needs. This is only one possible approach to developing a hierarchical classification structure for empathic response prediction and there may be other taxonomies that are better suited for an individual or a treatment type. For this reason, the models developed and evaluated in this work attempt to approximate this specific approach to empathy and further research is needed to understand what approaches may be best received by clients. Further, the virtual patient conversations in the provider platform evaluation were scripted. This limits their generalizability to real-patient interactions which is subject to future evaluation.

7.3 Future Work

There are many future directions based on the findings and methods developed in this work. In this section, I describe a few of these opportunities.

Prompt Engineering

Prompting is a new paradigm shift in natural language processing where instead of fine-tuning a model to a specific task, tasks are formatted in a way that aligns with generation from large language models, which has given rise to prompt engineering, i.e. selecting the right prompt to solicit the expected model behavior, as an area of exploration (Liu 2021). My work in this dissertation showed that appending emotional state information to the input sequence improves language model performance on the task of empathic response prediction in a more traditional task-specific architecture. In light of recent advances in NLP, this points to the potential to use chain-of-thought prompting (Wei 2022) for empathic response prediction, a method to improve large language model performance on arithmetic by providing intermediate reasoning steps. This method can be extended to empathic response generation by incorporating mental state inference to generate or retrieve responses. This

method was tested with Instruct GPT-3 via the OpenAI API with (Example 7.1) and without (Example 7.2) chain-of-thought prompting below.

Speaker: I can't keep up with my work.

Provide a response to the speaker:

You should try to find a way to better manage your time so that you can keep up with your work.

Example 7.1: Prompt with the response generated by Instruct GPT-3 through the OpenAI API in red

Speaker: I can't keep up with my work.

Mental State Inferences: The speaker feels overwhelmed. The speaker is seen as busy. The speaker needs to take a break.

Provide a response to the speaker:

I'm sorry to hear that you are feeling overwhelmed with your work. It sounds like you could really use a break. Maybe you can take some time for yourself this weekend and relax a bit.

Example 7.2: Chain of thought prompt created by adding mental state inferences with the response generated by Instruct GPT-3 through the OpenAI API output in green

Idiographic Assessment

Emotional responses to the same stimuli vary from individual to individual and temporally within the same individual. Therefore, it is important to verify emotional state inference with the individual. For this reason, it is advisable that these inferences be used to empathize with the user and verified before storage. For example, a system may ask what is new for a client and use events described by the client to infer a mental state. This mental

state can then be used to select an empathic response, but the system should still ask how the events made the client feel to verify the emotional state inference.

The system can incorporate differences between individuals and their relationships with work, family, or others when generating empathic responses. This would require acquiring an understanding of an individual's relationship to their environment through longitudinal engagement. Since emotional variance to the same stimulus may be less within the individual, it may be possible to infer the individual's emotional state more accurately when collecting this information longitudinally. Additionally, based on the common ground shared between the client and the system. The client may know that the system is aware of a particular emotional reaction to a particular stimulus and in the case that the emotional reaction differs make this clear to the system, making the utterance as informative as necessary based on Grice's maxim of quantity[13].

The end-goal of this system is to improve clinical decision support when it comes to behavioral health and emotional state inference is an intermediary toward that goal. When it comes to the decision making process regarding what solutions would work best for an individual, information collected through the knowledge graph presents a unique opportunity. The system includes an "ASSIGNED_TO" relationship between users and solutions which is updated with timestamps and ratings collected during follow-up sessions. This presents an opportunity for collaborative filtering personalized based on what has worked for others like them in the past. Information collected from journaling exercises can be used to match similar profiles and suggest goals, solutions, and resources that were highly rated by users with similar profiles.

[13] Grice's maxim of quantity states that, in cooperative dialog, a speaker should try to be as informative as they possibly can and give as much information as is needed, but no more than is needed.

Emotional Grounding

Another future direction would be to use natural language inference to determine if the client indicates that their feelings differ from the inferred feelings. Alongside empathic response generation, this would complete the second empathy criterion, "to communicate that understanding and **check its accuracy**", (Mercer & Reynolds 2002). In some contexts the system may not have enough information to confidently predict the emotional state of the individual. In these situations, it would be natural to ask for additional information to better understand the relationship between the individual and the event. The client may then respond to this with a message that indicates that this assumption was incorrect (Example 7.3).

Bot:	What's on your mind this week?	(Ask Event)
Client:	I had a long chat with my daughter.	(Event)
CAI:	Happy	(Emotional State Inference)
Bot:	Talking with your family can be a great way to destress. Whether it's catching up on what's going on in each other's lives or just having a laugh, spending time with family can help you relax and have a good time.	(Empathic Response)
Bot:	How did this conversation make you feel?	(Explicit Empathic Grounding)
Client:	Worried	(Emotional State)
CAI:	Contradiction	(Natural Language Inference)

Example 7.3: Natural language inference for emotional state verification of implicit emotional grounding

Dialog Management

This work has demonstrated the initial ability of conversational AI to express empathy in teletherapy. However, not every turn of dialog calls for an empathic response. One area for improvement of the present work is that it does not look at multiturn dialog. Incorporating this system into an unstructured dialog would require a system capable of determining when to empathize and when not. This would require larger scale data collection than was possible in this work, but could follow the motivational interviewing technique coding method (a transtheoretical approach commonly used in therapy) to label dialog acts.

Multimodal Learning

Whereas, until recently, models for processing natural language, computer vision, and audio signals have had different state-of-the-art architectures, the transformer architecture (Vaswani 2017) has proven to be a state-of-the-art generalized model for each and has shown incredible promise in handling multimodal input. Therefore, it would be a natural next step to test the findings from this work into datasets that include video or audio signals. In addition, such systems could incorporate gaze detection and emotion recognition from voice and facial expressions to create a singular model to predict empathic responses.

7.4 Implications for Health

This dissertation showed the potential for human-AI collaboration to help address the supply-demand imbalance in mental health. Automated systems can engage with individuals whenever is convenient for them and share that information with their care provider to reference during scheduled sessions which can be augmented through the same underlying conversational AI layer. Further, this technology can support a greater number of individuals to provide care.

How we process our emotional states influences our behavioral health decisions (Ferrer 2019), so automating the prediction and monitoring of these states could help to support digital behavioral health interventions. A study of users on a text-based social media platform identified posts that included a variant of the term "I feel __" where the blank is filled by an emotion label (Fan et al. 2019). The researchers analyzed the sentiment of the user's posts within six hours of the emotion labeling and found that the emotional intensity of user's posts increased up to the point of labeling, followed by a dramatic return to their emotional baseline. For example, neuroimaging researchers have shown that the simple process of labeling our emotional states can enable more mindful action by shifting brain activity from the amygdala, an area of the brain involved in maintaining affective states, toward the ventrolateral prefrontal cortex, an area responsible for goal achievement and long-term planning (Torre et al. 2018). Taken together, these results suggest that conscious interpretation of our emotional states can reduce their perceived intensity, and lead to more rational health-related decision-making.

One in ten individuals has difficulty articulating their emotions, a condition known as *alexithymia*. *Theory of mind* is the ability to infer the mental states of oneself and others, encompassing not only the emotional states inferred through empathy but also imputing states of knowledge, belief, desire, and intention (Premack and Woodruff 1978). Children typically develop this skill in the first four years of life (Perner et al. 1987), and recent research with children on the autism spectrum (AS) has shown that deficiencies in developing particular aspects of a *theory of mind* place strain on their social-communicative function (Mazza et al. 2017). Augmentative and alternative communication (AAC) devices are used by some people on the AS to assist with communication. This research has the potential to inform the development of systems to assist individuals both in the recognition of emotional states of others and also to communicate their empathy with others.

This technology was applied to a protocolized therapy designed to assist family caregivers. Over 43 million family caregivers provided unpaid care in the United States in 2019. This number rose significantly

during the pandemic as having professional caregivers enter the home was seen as a risk for spreading of COVID-19. Unfortunately, family caregivers are more likely than those without caregiving responsibilities to experience mental health symptoms related to anxiety and depression. These symptoms can reduce the ability of caregivers to provide treatment according to clinical recommendations, leading to poorer health outcomes and increased risk of hospitalization for those in their care. These mental health conditions often go undiagnosed and untreated since the family caregiver's mental health status is not collected as part of standard practice. This lack of emotional health tracking contributes to the health disparity between families with regular access to mental health care and those without access. While telemedicine solutions can potentially increase access to mental health therapy, the cost of these services may be prohibitive for many patients.

Tools to support individuals with less traditional training in providing behavioral health care, these technologies can support individuals in communities to deliver protocolized therapies with empathy. This may lead to localized care from within a community that is scalable globally, which in turn may lead to greater access to culturally competent care.

Applications of this technology include conversational agents for health (CAH), clinical decision support (CDS), patient relationship management (PRM), clinical training, and augmentative and alternative communication (AAC) devices.

7.5 Conclusions

While conversational AI is often thought about in healthcare through the lens of automated health dialog systems, this work has shown the potential for affect-aware conversational AI to improve the expression of empathy in care provider communication. Further, it has shown that emotional state inference improves empathic response prediction, using methods that are interpretable. Findings from experiments in which the resulting system was embedded in a decision support platform for text-based therapy show these approaches

improve the efficiency and empathy of patient-provider communication. The overarching system proposed in this work presents an opportunity to improve access to mental health care.

Appendix A. Search Strategy

PubMed	**ACM**
("User-Computer Interface[Mesh] AND ((virtual OR automated) (counselor OR advisor OR agent OR therapist OR nurse OR patient)))"	(+(virtual automated) +(counselor advisor nurse therapist patient) +(health healthcare))
(conversational OR relational) agent	(+(conversational relational) +(agent) +(health healthcare))
("User-Computer Interface[Mesh] AND ((dialog OR dialogue) system))"	(+(dialog dialogue) +(system) +(health healthcare))

Appendix B: Code Book for Cora Study

Code	Description
Work	**Mentions of the work environment including changes to the work environment, adjusting to WFH, struggles at work, productivity loss or gain**
WOR:Tasks	Participant mentions working on, being busy with or completing a work based task or effort.
WOR:General	Generally related to working or thinking about work.
WOR:Work_Status	The participant references job status (ex: job title), accomplishments, promotions, demotions.
WOR:Changes_at_Work	Changes in the workplace, excluding changes to the job status of the individual
WOR:Issue_at_Work	Participant experiences a general issue with the workplace (could involve a person or work directly), includes mistakes, being late, angry customer/manager
WOR:Change_in_Productivity	Mention of changes in ability to perform work either positively or negatively
WOR:Work_Life_Balance	Participant references work being too much or that it is interupting a healthy life balance.
Events	**Events that occur at a point in time**
EVE:Broken_Property	Mention of damage or malfunctioning to a possession of the participant, e.g. a car breaking down
EVE:COVID_Statistics	Mention of either worsening or improving statistics related to COVID, e.g. less deaths, more cases, etc.
EVE:Good_News	The participant mentions receiving good news, this is not to be confused with the news media but more likely news from a non-media source
EVE:Moving	The participant describes that they will be moving, are moving, or have moved from one domicile to another
EVE:Not_Being_Heard	Feeling that others are not listening or reacting to the participant
EVE:Victimization	The participants was the target of a crime, e.g. stalking, robbery, etc.
EVE:Time_Off	The participant mentions taking any time away from work, e.g. as a vacation or the weekend
EVE:Start_of_School	The participant mentions the start of school, or home schooling
EVE:Exam	The participant mentions needing to complete or having completed an exam
EVE:Change_In_Plans	Changes in long-term plans that were adjusted as a result of COVID-19
Personal Relationships	**Relationships that are within a household, extended family, friends, and neighbors**

PER:Child	Any mention of children that need not belong to the participant, e.g. the neighborhood children or my daughter
PER:Family	A non-specific reference to the nuclear or extended family, this code should not be used in combination with parent, child, partner, pet, sibling
PER:Friend	A friendly acquaintance of the participant
PER:Parent	The parent of the participant
PER:Partner	A person to whom the participant is married or with whom they are having a romantic or sexual relationship
PER:Pet	An animal with whom the participant has a relationship
PER:Roommate	A person with which the participant shares a domicile
PER:Sibling	The sibling of the participant
PER:Neighbor	A person who lives in near proximity to the participant, but is not a roommate
PER:Other	A catch all category for all other personal relationships
Professional Relationships	**Relationships that exist in a work setting**
PRO:Coworker	Anyone with whom the participant interacts in a peer capacity
PRO:Manager	Anyone who manages the participant directly or indirectly, e.g. CEO, Boss, etc
PRO:Customer	Including students, customers, etc
Environment	**Aspects of the users environment**
ENV:Current_Situation	Any mention that the user is happy or unhappy with the way things are for them at the present moment
ENV:Mess	Any mention of cleanliness or tidyness
ENV:Noise	Any mention of noise disturbances, e.g. children screaming
ENV:Opinion_of_Others	Any mention of the opinion of others, e.g. what other people said about me or what my friends may think about me if I wear this
ENV:Pandemic	Any reference to the pandemic, coronavirus, COVID-19 explicitly causing anxiety.
ENV:Shared_Custody	Any mention of shared custody of children or pets, i.e. a child or pet needing to spend time in multiple households
ENV:Social_Isolation	Any mention of feeling alone or cut off from friends or family
ENV:Food_Security	Any concerns related to being able to find or afford food
Fiscal	**Financial concerns**
FIN:Financial_Matters	Changes in someone's financial standing (increases or decreases).
FIN:Finding_Work	Any description of looking for work, interviewing for a job, getting a job offer/rejection, not finding work, etc.

FIN:Other	Catch all category related to all other financial concerns
Activities	**Actions that are completed by the participant**
ACT:Teaching	Instructing or training a co-worker, customer, or child
ACT:Thinking_about_the_Future	Thinking of any future events, e.g. going to work tomorrow, waking up tomorrow, things getting worse
ACT:Haircut	Mention of haircuts
ACT:Making_Appointment	Mention of making an appointment
ACT:Media_Entertainment	Mention of watching or otherwise consuming media entertainment including sports, tv, etc. excluding the news
ACT:News	Mention of watching or otherwise consuming news media, or indirectly mentioning something that would be understood as coming from news media
ACT:Outdoor_Activity	Mention of getting fresh air, running, exercising, hiking, sailing, driving etc., excluding shopping
ACT:Home_Activity	Any description of doing things inside the home
ACT:Personal_Growth	Any description of improvement or achievement this can be in relation to self or that of a child commonly
ACT:Remote_Activity	An acitivty that is not done in person, e.g. playing games online with friends or family. This code should be used in addition to another activity is the activity is mentioned to be remote, e.g. Talking with a friend online. Would be both Talking and Remote_Activity.
ACT:Shopping	Any description of the participant going outdoors to shop
ACT:Social_Activity	Mention of an activity that consists of multiple individuals communicating or participating in an activity, e.g. helping others, talking with somone, gaming with friends.
ACT:Travel	Mention of going outside of the home, but not as part of an outdoor activity or shopping.
ACT:Tidying_Up	Mention of tidying up the home
ACT:Waiting	Mention of the experience of waiting on something or someone
Government_Response	**Concepts that are under government control**
GOV:Reopening	Government actions related to reopening or easing covid restrictions.
GOV:General	Government changing or adjusting laws or policies, or general issue with government response.
GOV:Stay-at-Home	Government mandated stay at home.
GOV:Public_Transportation	Any mention of the use of public transport or changes to public transportation options

Health	**Health related topics**
HEA:Mental_Wellbeing	Mention of mental wellness
HEA:Sex	Mention of sex
HEA:Sleep_Quality	Mention of sleep quality
HEA:Social_Distancing	Mention of social distancing
HEA:Substance_Abuse	Mention of substance abuse
HEA:Susceptibility	User or someone they're speaking about is perceived to have a high risk for covid.
HEA:Other_Condition	Mention of any other health condition
HEA:Diet	Mention of dieting
Nothing	**The user specifically states that nothing has caused them hope or anxiety**

Appendix C: Causes of Anxiety and Hope in Cora Study

Table C.1: Causes of anxiety with number of unique participants who shared a response in that category

Code	Count
Nothing	63
WOR:General	37
ENV:Current_Situation	35
ACT:Thinking_about_the_Future	33
ACT:Travel	32
HEA:Other_Condition	32
ACT:Social_Activity	30
HEA:Mental_Wellbeing	30
PER:Child	30
ACT:Shopping	29
ENV:Pandemic	28
WOR:Tasks	27
HEA:Social_Distancing	26
FIN:Financial_Matters	23
PER:Family	21
PER:Partner	21
WOR:Issue_at_Work	21
ACT:News	20
PER:Friend	19
GOV:Stay-at-Home	18
PER:Other	18
PER:Parent	17
ACT:Outdoor_Activity	15
GOV:Reopening	15
HEA:Susceptibility	15
HEA:Sleep_Quality	14
ACT:Waiting	13

ENV:Social_Isolation	13
EVE:Moving	13
FIN:Finding_Work	13
ACT:Home_Activity	11
PRO:Coworker	11
ACT:Media_Entertainment	10
EVE:Change_In_Plans	10
EVE:COVID_Statistics	10
EVE:Start_of_School	10
GOV:General	10
WOR:Change_in_Productivity	10
WOR:Work_Status	10
ACT:Tidying_Up	9
ENV:Opinion_of_Others	9
WOR:Changes_at_Work	9
EVE:Broken_Property	8
PRO:Customer	8
PER:Pet	7
WOR:Work_Life_Balance	7
ENV:Mess	5
EVE:Victimization	5
HEA:Diet	5
ACT:Remote_Activity	4
EVE:Exam	4
EVE:Time_Off	4
GOV:Public_Transportation	4
PRO:Manager	4
ENV:Noise	3
PER:Neighbor	3
ACT:Making_Appointment	2
ACT:Teaching	2

ENV:Food_Security	2
EVE:Not_Being_Heard	2
FIN:Other	2
HEA:Substance_Abuse	2
PER:Sibling	2
ACT:Haircut	1
ACT:Personal_Growth	1
ENV:Shared_Custody	1
HEA:Sex	1
PER:Roommate	1

Table C.2: Causes of hope with number of unique participants who shared a response in that category

Code	Count
ACT:Social_Activity	66
PER:Friend	56
ENV:Current_Situation	53
Nothing	44
PER:Family	41
ACT:Outdoor_Activity	38
PER:Child	38
ACT:Thinking_about_the_Future	34
PER:Partner	31
EVE:Good_News	27
HEA:Mental_Wellbeing	27
GOV:Reopening	23
ACT:Home_Activity	22
EVE:Time_Off	22
WOR:General	22
WOR:Tasks	22
ACT:Media_Entertainment	20

ACT:Personal_Growth	20
FIN:Financial_Matters	18
ENV:Pandemic	17
PER:Parent	16
FIN:Finding_Work	15
ACT:Remote_Activity	13
ACT:Shopping	11
HEA:Sleep_Quality	11
WOR:Change_in_Productivity	11
EVE:Moving	10
PER:Pet	9
PRO:Coworker	9
ACT:News	8
EVE:Start_of_School	8
Government_Response	8
HEA:Other_Condition	8
ACT:Haircut	7
ACT:Teaching	7
ACT:Tidying_Up	7
EVE:COVID_Statistics	6
HEA:Diet	6
PER:Other	6
HEA:Social_Distancing	4
PRO:Customer	4
ENV:Opinion_of_Others	3
FIN:Other	3
PER:Roommate	3
WOR:Changes_at_Work	3
ACT:Making_Appointment	2
ACT:Travel	2
EVE:Broken_Property	2

Activities	1
ENV:Noise	1
ENV:Shared_Custody	1
EVE:Change_In_Plans	1
GOV:General	1
GOV:Stay-at-Home	1
HEA:Substance_Abuse	1
PER:Neighbor	1
PER:Sibling	1
PRO:Manager	1

Appendix D. Cluster Example

Table D.1: Two example utterance for each response label (cluster) under the "family" topic

Utterance (Event)	Response Label
Had a nice play date with friends family	family
My son and partners' flight home got cancelled. I'm really missing them and worrying about their safety. I also had to arrange a ride from the airport for them.	family
My grandma isn't feeling well	family+care receiver (not child)
I found out my friend's cancer came back and has spread to her lungs	family+care receiver (not child)
My mom moved out of an isolation room.	family+care receiver (not child)+improvement
If my husband had not had a dilerium episode and ended up in hospital er and now it is not safe for me to bring him home. Our lives are turned upside down once again.	family+care receiver (not child)+not well
My mom went to ER today	family+care receiver (not child)+not well
I took my daughter to have her ears pierced	family+child
I bought Halloween costumes with my daughter	family+child
Meltdowns from my child	family+child+difficult behavior
My child is refusing go to summer camp	family+child+difficult behavior
My son had tantrum again this afternoon because my husband said something he doesn't like.	family+child+difficult behavior+tantrums
My 5 year old having an extended melt down at bedtime	family+child+difficult behavior+tantrums
My son took a new medication without any arguments tonight	family+child+good behavior
My son prepared me a gift in the morning	family+child+good behavior
My daughter was given a healthy medical report.	family+child+health
My 9 year old is in the ICU still.	family+child+health
Talking with my sons VA doctor	family+child+health+doctor appointment
Great news at Dr. Appointment with daughter	family+child+health+doctor appointment
My child sleeping through the night without pain would be nice.	family+child+health+health condition
More worry about kids returning to school next week	family+child+health+health condition
I enjoyed returning home to happy kids after running some errands.	family+child+joy
Kids excited for school	family+child+joy
Didn't get to spend time with my kids	family+child+not enough time
I could have spent more time teaching my daughter	family+child+not enough time

Watching Olympics with my daughter	family+child+spend time together
I went to the new uwajimaya with my son today	family+child+spend time together
Potty training fail	family+child+struggling
Taking care of condition with child and hard to see her having difficult time	family+child+struggling
Family problems	family+family conflict
Ruminating about mom	family+family conflict
I made great food for my family and friends	family+family time
Having more family time	family+family time
The strange thing is that learning about/talking with the sister who's son-in-law is on the verge of death served as a reminder of how many "families" (biological; legal; OTJ; In the condo; etc) I have for love & support through grief AND for celebrating all things wonderful.	family+loved one's death
Just found out my children's paternal grandfather passed away today, at 100 years and 5 months.	family+loved one's death
Talking about finances with my husband	family+partner
My husband was talking to me today after we had some tough discussions	family+partner
Had a good, open conversation with my husband to talk about our communication issues	family+partner+good communication
I was able to have a good conversation with my partner	family+partner+good communication
My husband is upset with me and is not talking to me	family+partner+issues
Fighting with my husband	family+partner+issues
Went on a date with my husband.	family+partner+spend time
Had anniversary dinner with my wife	family+partner+spend time

Appendix E. Results by Response

Table E.1: BART versus BART + ESI (xReact n=5) versus COMET model performance by response label measured by the F1-scores for that model averaged across all 5-folds with total support > 10 and sorted by based on improvement in BART + ESI performance which correlates with the response labels where COMET outperformed BART

Response Label	Support	BART	BART w/ ESI	COMET	BART w/ ESI - BART	COMET - BART
self-care+health+sleep+nap	15	0.48	0.72	0.75	0.24	0.27
work+struggling	108	0.36	0.59	0.52	0.24	0.16
self-care+health+sleep+good_sleep	11	0.20	0.40	0.43	0.20	0.23
family+partner	55	0.35	0.56	0.42	0.20	0.07
self-care+enjoy_food	45	0.26	0.45	0.48	0.19	0.23
family+child+struggling	103	0.35	0.52	0.46	0.18	0.12
family+care_receiver_(not_child)	24	0.08	0.25	0.27	0.18	0.19
family+child+good_behavior	70	0.31	0.47	0.41	0.16	0.10
family+child+difficult_behavior+tantrums	37	0.39	0.54	0.62	0.15	0.23
self-care+health+sleep+sleep_-_lack_of	67	0.61	0.76	0.72	0.14	0.11
self-care+health+physical_pain	50	0.52	0.65	0.66	0.14	0.14
family+partner+issues	70	0.57	0.71	0.73	0.13	0.16
self-care+not_eating_well	22	0.34	0.47	0.56	0.13	0.22
family+child+health	32	0.33	0.45	0.46	0.12	0.13
general+good	105	0.27	0.39	0.37	0.12	0.10
self-care+health+not_feeling_well	78	0.48	0.59	0.68	0.11	0.20
self-care+exercise	123	0.62	0.71	0.69	0.09	0.07
family+care_receiver_(not_child)+not_well	28	0.12	0.21	0.31	0.09	0.18
self-care+relax	155	0.56	0.65	0.66	0.09	0.10
family+family_conflict	39	0.15	0.23	0.24	0.08	0.09

family+partner+good_communication	12	0.23	0.31	0.00	0.08	-0.23
deny	30	0.68	0.76	0.68	0.08	0.01
work+productive	216	0.66	0.74	0.72	0.08	0.06
work+career_change	23	0.37	0.44	0.43	0.07	0.06
from_others+negative	48	0.19	0.26	0.25	0.07	0.06
from_others+positive_interaction	45	0.27	0.33	0.50	0.06	0.23
family+partner+spend_time	28	0.33	0.39	0.27	0.06	-0.06
managing_time	92	0.32	0.38	0.39	0.06	0.07
weather	18	0.25	0.31	0.36	0.05	0.11
work+generic-good	76	0.33	0.38	0.43	0.05	0.10
self-care+exercise+exercise_-_lack_of	48	0.56	0.59	0.71	0.03	0.14
family+family_time	274	0.67	0.70	0.66	0.03	-0.01
general+negative	117	0.31	0.34	0.41	0.02	0.10
self-care+socialize_with_others	254	0.70	0.72	0.74	0.02	0.04
family+child+spend_time_together	218	0.63	0.65	0.65	0.01	0.01
self-care+help_others	31	0.12	0.13	0.25	0.01	0.13
family+child+difficult_behavior	78	0.32	0.33	0.39	0.01	0.06
family+child	18	0.00	0.00	0.00	0.00	0.00
from_others+others_help	13	0.00	0.00	0.00	0.00	0.00
self-care	12	0.00	0.00	0.00	0.00	0.00
family	50	0.11	0.10	0.14	-0.01	0.03
family+child+joy	20	0.12	0.11	0.24	-0.01	0.11
work+unproductive	91	0.52	0.47	0.57	-0.04	0.05
role_conflict	31	0.20	0.16	0.23	-0.04	0.03
work+generic-bad	56	0.35	0.29	0.31	-0.05	-0.03
work+stressed	27	0.14	0.08	0.10	-0.06	-0.04
self-care+take_break	26	0.28	0.20	0.32	-0.09	0.04
occasion	13	0.44	0.35	0.41	-0.09	-0.03
animals	30	0.35	0.23	0.36	-0.12	0.01
self-care+shopping	22	0.25	0.12	0.37	-0.13	0.12
family+child+not_enough_time	15	0.30	0.14	0.18	-0.17	-0.12
cooking	21	0.37	0.13	0.26	-0.24	-0.11
self-care+health+mental_health	13	0.45	0.10	0.31	-0.35	-0.14

Appendix F: Moderator Script

Introduction & Greetings

Welcome [participant name]. My name is {moderator name}, I am with the COCO team and we would like to thank you for your interest and support in our platform testing.

Overview of the Session

Today, we will conduct a 30 minute testing session with you. You will be interacting with our provider platform and there are specific tasks that we will lay out for you to complete. The GOAL is not to test your ability as a provider, but rather to help us improve the use of the platform.

Training Video

We will begin by watching a brief training video (about 5 mins long) that will introduce you to the platform. After the video I will provide you with a link that you can click to access the platform.

Do you have any questions before we get started?

First, I'd like you to open a browser (Chrome or Firefox), and start sharing your screen.

[Copy & Paste the following Training Video Link below into the Zoom chat]

> https://youtu.be/EoIlBKNwQ7Y

I'm putting a link in the chat, and please go ahead and open the link and start playing.

Thank you for taking the time to watch the video. Do you have any questions before we get started?

[Start First Session]

[Wizard selects the appropriate control or experimental session and script based on randomization, see Study Kit]

Now I am adding a link to the provider platform to the chat. Please open this link, login to the platform using the username and password, and share your screen.

[Copy & paste this into the chat in Zoom]:

> Link to Provider-Facing Platform: https://woz.cocobot.care/
> Participant Username: tester_2
> Participant Password: ococococ

[Moderator or Wizard logs into the client side using incognito mode]:

https://woz.cocobot.care/client.html

Please select the box with the arrow in the box on the top left. This will allow you to view the active messages:

Please select on the Irina Williams "pending message" box. Now please select "Join Session". We will now begin the session and you will use the platform to start a conversation with Irina. You will be initiating the conversation, which is different from the video. Go ahead and choose what you think is most appropriate. And from this time on, please be sure to include both an empathetic response and a therapeutic response.

[End First Session]

At this time, you have completed the first session. Please click on "end the session" on the rightside. You will be prompted to start a new session. Please treat this as a new client.

[Start Second Session]

[Wizard selects appropriate control or experimental session and script, see Study Kit]

Please start the new session and select the appropriate conversation starter.

Again, please choose the responses that make the most sense to you.
If you do not see a response that makes sense to you, you can use "Control + F" to identify a response based on a keyword.

[End Second Session]

*This concludes the session, please click on "**end session"** on the right side of the screen.*

Thank you for completing the testing sessions. I would like to ask you a few questions before we conclude our time together. And if it's ok with you, I'd like to record this part just for our internal use and gather notes.

Appendix G: Virtual Patient Scripts

Sleep

Provider (User): Greetings, {name}! Thank you for taking the time to care for yourself. Tell me about your {day} so far.

Client (Wizard): I am struggling to sleep at night.

Provider (User): I'm sorry to hear you didn't sleep well. It's hard to feel refreshed and ready for the day ahead when you didn't get a good night of sleep. It sounds like you are experiencing sleep disturbance. Is my understanding correct?

Client (Wizard): Yes, that is correct.

Provider (User): Alright. Thank you for letting me know. I would like us to use the Problem Solving Therapy (PST) process that I mentioned to you last time to address your {sleep disturbance}. Before we get started, do you have any questions?

Client (Wizard): not that I can think of.

Provider (User): Great. To start, can you tell me what is causing your sleep disturbance?

Client (Wizard): I'm not sure why I'm having trouble sleeping. My kid is still sick and not sleeping well. I spend a lot of energy taking care of him.

Provider (User): Caregiving can be challenging sometimes, especially when you already have a lot on your plate. In the past week, how often did you experience sleep disturbance?

Client (Wizard): I haven't slept well in days. It has been 5 days out of 7 that I have experienced trouble sleeping.

Provider (User): I see. Between 0-5, with 0 being the lowest level of sleep disturbance and 5 being the highest, how would you rate your sleep disturbance today?

Client (Wizard): 4.5

Provider (User): Got it. What have you already tried to help reduce your sleep disturbance?

Client (Wizard): I have not tried anything.

Provider (User): I am here to help. The next step in the PST process is to set a goal to help reduce your {sleep disturbances}. And for our first session today, I would like to suggest a couple for you to choose from: 1. Take some time for yourself, 2. Exercise more

Client (Wizard): 2

Stress

Provider (User): Hello, thanks for taking time to care for yourself. Tell me about your day so far.

Client (Wizard): I had a difficult conversation with my manager and another coworker. So I have been a bit stressed out from work.

Provider (User): Difficult conversations are unavoidable. And they can be pretty stressful sometimes.

It sounds like you are experiencing stress. Is my understanding correct?

Client (Wizard): Yes

Provider (User): Thank you for letting me know. I would like us to use the Problem Solving Therapy (PST) process that I mentioned to you last time to address your {stress}.

Before we get started, do you have any questions?

Client (Wizard): no

Provider (User): Alright. Besides what you have mentioned, is there anything else that contributes to your stress?

Client (Wizard): My son was grumpy all day and I was short with him.

Provider (User): That sounds difficult. Kids bring us joy but sometimes can be hard to deal with. In the past week, how often did you experience stress?

Client (Wizard): 3 times

Provider (User): I see. Between 0-5, with 0 being the lowest level of stress and 5 being the highest, how would you rate your stress today?

Client (Wizard): 2

Provider (User): I understand. What have you already tried to help reduce your stress?

Client (Wizard): Taking a walk

Provider (User): Great! Let's come up with other things that can help you further improve your {stress}. The next step in the PST process is to set a goal to help reduce your stress. And for our first session today I would like to suggest a couple for you to choose from: 1. Take time for yourself; 2. Exercise. Which one would you like to try first?

Client (Wizard): 1

Appendix H: Intake Survey

1. Please provide your current status at UW. If you are at a different institution, please select "other" and share your institution name and your education (e.g., BS, MS)
2. Please share your age in years
3. Please provide your current field of study
4. How many years have you been a nurse working directly with patients and families?
5. Do you have prior mental health training beyond what was included in your undergraduate nursing education?
6. Please explain your mental health training (e.g., through what program and for how long?)
7. Have you worked in a psychiatric/behavioral/mental health setting for more than 6 months in total?
8. What is your current training level?

Appendix I: System Usability Survey Results

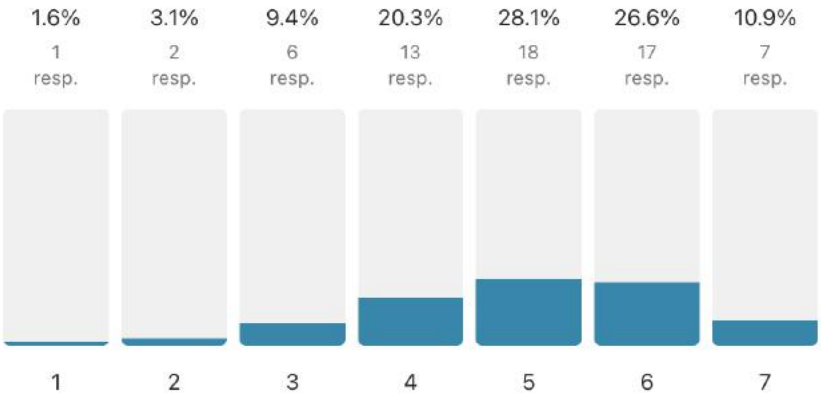


Figure I.1: Using the COCO Provider Platform in my work will help me to accomplish my task more quickly.

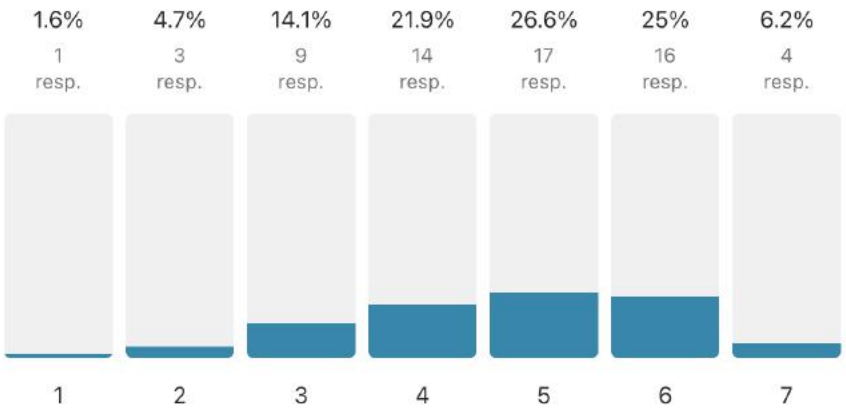


Figure I.2: Using the COCO Provider Platform will improve my work performance.

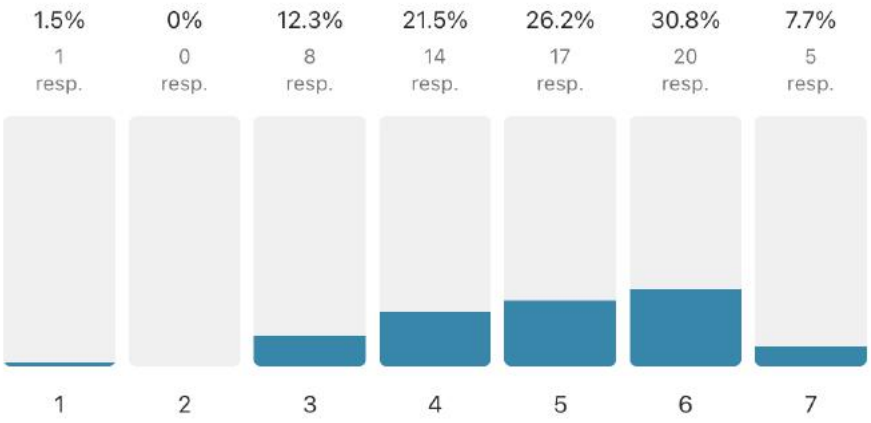


Figure I.3: Using the COCO Provider Platform will increase my work productivity.

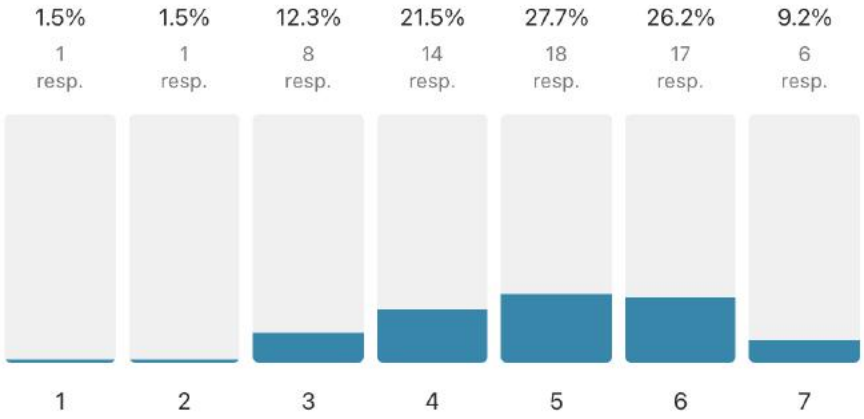


Figure I.4: Using the COCO Provider Platform will enhance my effectiveness at work.

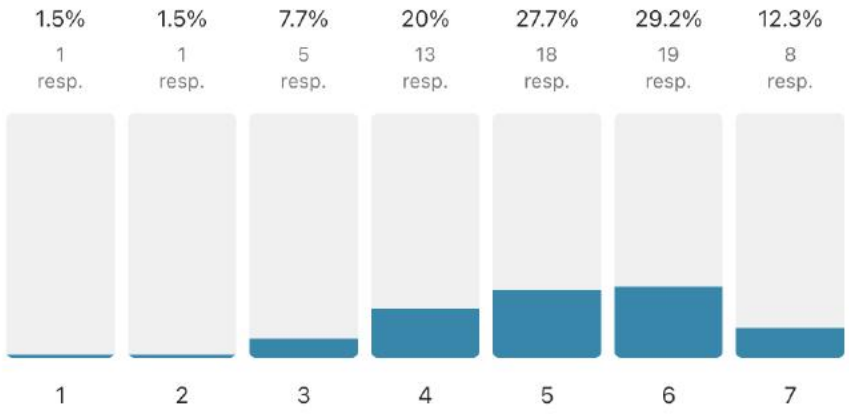


Figure I.5: Using the COCO Provider Platform will make it easier to do my work.

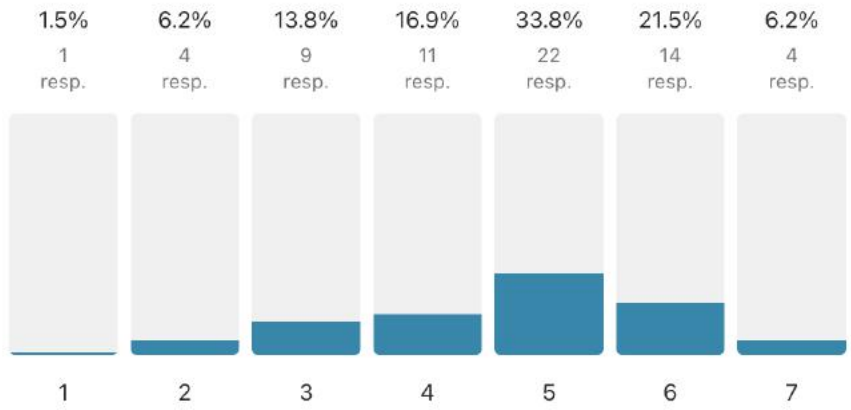


Figure I.6: I find the COCO Provider Platform useful for my work.

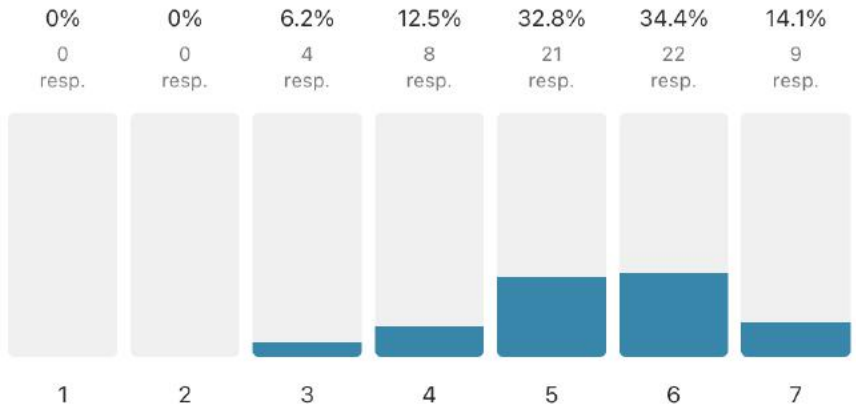


Figure I.7: Learning to operate the COCO Provider Platform appears to be easy.

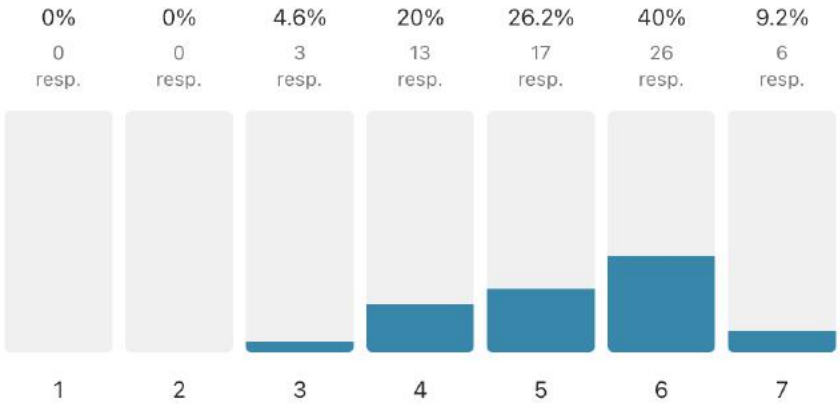


Figure I.8: I observed the COCO Provider Platform easy to navigate.

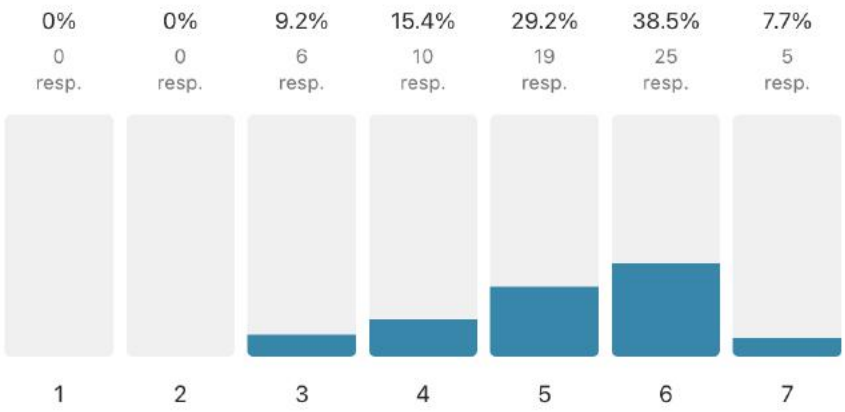


Figure I.9: Interaction with the COCO Provider Platform is clear and understandable.

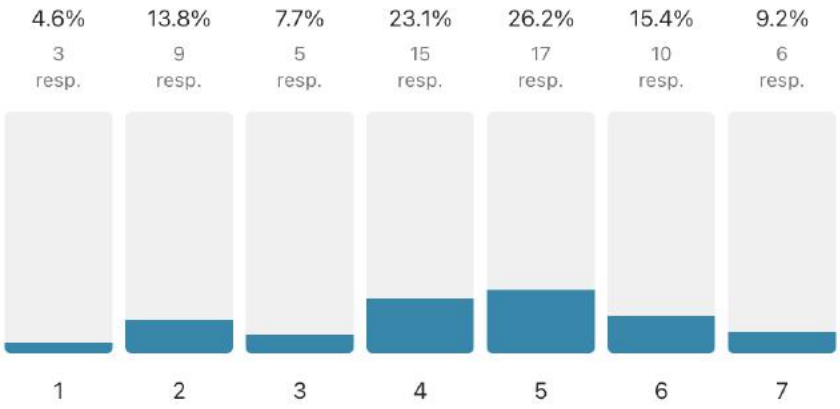


Figure I.10: I would use the COCO Provider Platform frequently (many times per week).

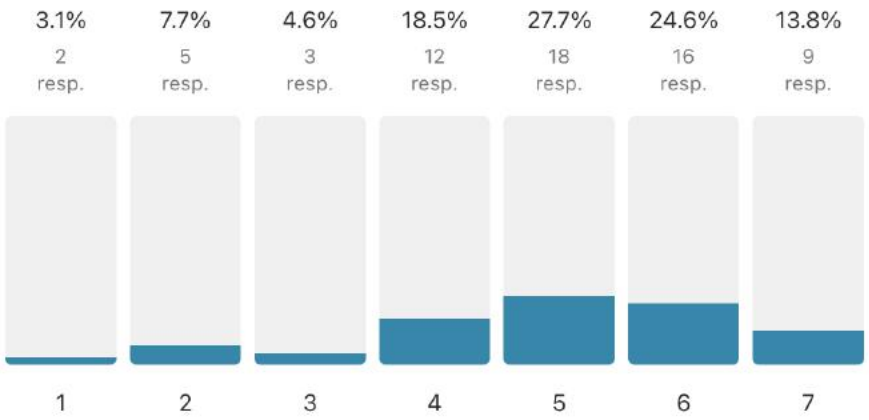


Figure I.11: I would use the COCO Provider Platform to augment clinical care.

Bibliography

[1] Acheampong, F.A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports, 2.

[2] Adams WG, Phillips BD, Bacic JD, Walsh KE, Shanahan CW, Paasche-Orlow MK. Automated conversation system before pediatric primary care visits: a randomized trial. Pediatrics. 2014;134(3):e691-9.

[3] Albert Bandura. 1986. Social foundations of thought and action: a social cognitive theory. Prentice-Hall, Englewood Cliffs, N.J.

[4] Alsaeedi, A., & Zubair, M. (2019). A Study on Sentiment Analysis Techniques of Twitter Data. International Journal of Advanced Computer Science and Applications.

[5] Amini R, Lisetti C, Yasavur U, Rishe N. On-demand virtual health counselor for deliver- ing behavior-change health interventions. IEEE Int Conf on Health Inform. 2013;46-55

[6] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. CoRR, abs/1409.0473.

[7] Bangor, A., Kortum, P.T., & Miller, J.T. (2009). Determining what individual SUS scores mean: adding an adjective rating scale. Journal of Usability Studies archive, 4, 114-123.

[8] Barnett, M.L., Gonzalez, A., Miranda, J. et al. Mobilizing Community Health Workers to Address Mental Health Disparities for Underserved Populations: A Systematic Review. Adm Policy Ment Health 45, 195–211 (2018). https://doi.org/10.1007/s10488-017-0815-0

[9] Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A., & Pollak, S.D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. Psychological Science in the Public Interest, 20, 1 - 68.

[10] Beatty, C.C., Malik, T., Meheli, S., & Sinha, C. (2022). Evaluating the Therapeutic Alliance With a Free-Text CBT Conversational Agent (Wysa): A Mixed-Methods Study. Frontiers in Digital Health, 4.

[11] Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.

[12] Bickmore T, Giorgino T. Health dialog systems for patients and consumers. J Biomed Inform. 2006;39(5), 556–571.

[13] Bickmore TW, Trinh H, Olafsson S, et al. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of siri, alexa, and google assistant. JMIR, 2018.

[14] Black LA, McTear M, Black N, Harper R, Lemon M. Appraisal of a conversational arte- fact and its utility in remote patient monitoring. In: Proceedings 18th IEEE Symposium on CBMS, 2005; 506–508.

[15] Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., & Choi, Y. (2019). COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. ACL.

[16] Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent. COMPSTAT.

[17] Bourgault P, Lavoie S, Paul-Savoie E, Grégoire M, Michaud C, Gosselin E, Johnston CC. Relationship Between Empathy and Well-Being Among Emergency Nurses. J Emerg Nurs. 2015 Jul;41(4):323-8. doi: 10.1016/j.jen.2014.10.001. Epub 2015 Jan 10. PMID: 25583425.

[18] Brown, B. (2020). Atlas of the heart. Random House.

[19] Bunk, T., Varshneya, D., Vlasov, V., & Nichol, A. (2020). DIET: Lightweight Language Understanding for Dialogue Systems. ArXiv, abs/2004.09936.

[20] Busso, C., Bulut, M., Lee, C., Kazemzadeh, E., Provost, E.M., Kim, S., Chang, J.N., Lee, S., & Narayanan, S.S. (2008). IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42, 335-359.

[21] COVID-19 Mental Disorders Collaborators. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. Lancet. 2021 Nov 6;398(10312):1700-1712. doi: 10.1016/S0140-6736(21)02143-7.

[22] Cahn, J.E., & Brennan, S.E. (1999). A Psychological Model of Grounding and Repair in Dialog.

[23] Chen, S., Hsu, C., Kuo, C., Huang, T.'., & Ku, L. (2018). EmotionLines: An Emotion Corpus of Multi-Party Conversations. ArXiv, abs/1802.08379.

[24] Cherry D, Albert M, McCaig LF. Mental health-related physician office visits by adults aged 18 and over: United States, 2012–2014. NCHS Data Brief, no 311. Hyattsville, MD: National Center for Health Statistics. 2018.

[25] Chiu CC, Sainath TN, Wu Y, et al. State-of-the-art Speech Recognition With Sequence- to-Sequence Models.IEEE ICASSP, 2017:3-7.

[26] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-Based Models for Speech Recognition. ArXiv, abs/1506.07503.

[27] Clark, H.H., & Brennan, S.E. (1991). GROUNDING IN COMMUNICATION.

[28] Cocker, F., & Joss, N. (2016). Compassion Fatigue among Healthcare, Emergency and Community Service Workers: A Systematic Review. International journal of environmental research and public health, 13(6), 618.

[29] Colombo, D., Suso-Ribera, C., Fernández-Álvarez, J., Cipresso, P., García-Palacios, A., Riva, G., & Botella, C. (2020). Affect Recall Bias: Being Resilient by Distorting Reality. Cognitive Therapy and Research, 1-13.

[30] Cowen, A. S., & Keltner, D. (2017). 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, 114(38), E7900–E7909. https://doi-org.offcampus.lib.washington.edu/10.1073/pnas.1702247114

[31] Cowen, A. S., & Keltner, D. (2018). Clarifying the Conceptualization, Dimensionality, and Structure of Emotion: Response to Barrett and Colleagues. Trends in cognitive sciences, 22(4), 274–276. https://doi-org.offcampus.lib.washington.edu/10.1016/j.tics.2018.02.003

[32] Creed, T.A., Salama, L., Slevin, R.A., Tanana, M.J., Imel, Z.E., Narayanan, S.S., & Atkins, D. (2022). Enhancing the quality of cognitive behavioral therapy in community mental health through artificial intelligence generated fidelity feedback (Project AFFECT): a study protocol. BMC Health Services Research, 22.

[33] D'zurilla, T.J., & Nezu, A.M. (2007). Problem-solving therapy: a positive approach to clinical intervention.

[34] Darcy, A., Beaudette, A., Chiauzzi, E., Daniels, J., Goodwin, K., Mariano, T.Y., Wicks, P., & Robinson, A. (2022). Anatomy of a Woebot® (WB001): agent guided CBT for women with postpartum depression. Expert Review of Medical Devices, 19, 287 - 301.

[35] De Mori R, Bechet F, Hakkani-Tur D, McTear M. Spoken language understanding. Signal Processing Magazine, IEEE, 2008;50–58.

[36] Deladisma AM, Gupta M, Kotranza A, et al. A pilot study to integrate an immersive virtual patient with a breast complaint and breast examination simulator into a surgery clerkship. Am J Surg. 2009;197(1):102-106.

[37] Deladisma AM, Johnsen K, Raij A, et al. Medical student satisfaction using a virtual patient system to learn history-taking communication skills. Stud Health Technol Inform. 2008;132:101-105.

[38] Demiris, G., Oliver, D.P., Washington, K.T., Fruehling, L.T., Haggarty-Robbins, D., Doorenbos, A.Z., Wechkin, H.A., & Berry, D.L. (2010). A Problem Solving Intervention for hospice caregivers: a pilot study. Journal of palliative medicine, 13 8, 1005-11 .

[39] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A.S., Nemade, G., & Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. ArXiv, abs/2005.00547.

[40] Devault D, Artstein R, Benn G, et al. SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. AAMS, 2014.

[41] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. 2019.

[42] Dickerson R, Johnsen K, Raij A, et al. Virtual patients: assessment of syn- thesized versus recorded speech. Stud Health Technol Inform. 2006;119:114-119. http://www.ncbi.nlm.nih.gov/pubmed/16404028.

[43] Duan J, Zhao H, Zhou Q, Qiu M, and Liu M. (2020) A Study of Pre-trained Language Models in Natural Language Processing. 2020 IEEE International Conference on Smart Cloud (SmartCloud); 116-121.

[44] Dumais ST, Furnas GW, Landauer TK, Deerwester S, Harshman R. Using latent semantic analysis to improve access to textual information. InProceedings of the SIGCHI conference on Human factors in computing systems 1988 May 1 (pp. 281-285).

[45] D'Zurilla T. and Nezu A.M. 1999. Problem-Solving Therapy (2nd. ed.). Springer Publishing, New York, NY.

[46] Ekman, P., & Friesen, W.V. (1971). Constants across cultures in the face and emotion. Journal of personality and social psychology, 17 2, 124-9 .

[47] Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. Journal of advanced nursing, 62(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

[48] Fagnano M, Emily Berkman, Elise Wiesenthal, Arlene Butz, and Jill S. Halterman. 2012. Depression among caregivers of children with asthma and its impact on communication with health care providers. Public Health, 126, 12: 1051-1057. DOI:https://doi.org/10.1016/j.puhe.2012.08.007

[49] Fan, R., Varol, O., Varamesh, A. et al. The minute-scale dynamics of online emotions reveal the effects of affect labeling. Nat Hum Behav 3, 92–100 (2019). https://doi.org/10.1038/s41562-018-0490-5

[50] Farzanfar R, Frishkopf S, Friedman R, Ludena K. Evaluating an automated mental health care system: making meaning of human-computer interaction. Comput Human Behav. 2007;23(3):1167-1182.

[51] Farzanfar R, Frishkopf S, Migneault J, Friedman R. Telephone-linked care for physi- cal activity: a qualitative evaluation of the use patterns of an information technology program for patients. J Biomed Inform. 2005; 38(3), 220–228.

[52] Feller CP, Cottone RR. The importance of empathy in the therapeutic alliance. The Journal of Humanistic Counseling, Education and Development. 2003 Mar;42(1):53-61.

[53] Ferrer, R.A., & Ellis, E.M. (2019). Moving beyond categorization to understand affective influences on real world health decisions. Social and personality psychology compass, 13.

[54] Firth, J.R. (1957). A Synopsis of Linguistic Theory, 1930-1955.

[55] Fisher, A.P., Gies, L.M., Narad, M.E., Austin, C.A., Yeates, K.O., Taylor, H.G., Zhang, N., & Wade, S.L. (2021). Parent- and Adolescent-reported Executive Functioning in the Context of Randomized Controlled Trials of Online Family Problem-Solving Therapy. Journal of the International Neuropsychological Society, 28, 123 - 129.

[56] Fitzpatrick KK, Darcy A, Vierhile M. 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 Ment Heal. 2017;4(2):e19.

[57] Fleming M, Olsen D, Stathes H, et al. Virtual reality skills training for health care profes- sionals in alcohol screening and brief intervention. J Am Board Fam Med. 2009;22(4):387- 398.

[58] Forster AJ, LaBranche R, McKim R, Faught JW, Feasby TE, Janes-Kelley S, Shojania KG, van Walraven C. Automated patient assessments after outpatient surgery using an interactive voice response system. Am J Manag Care, 2008;14(7):429-36.

[59] Fox L, Schlesinger J. Giving Patients an Active Role in Their Health Care. Harvard Business Review. HBR 2017; Accessed at https://hbr.org/2016/11/ giving-patients-an-active-role-in-their-health-care

[60] Freeman M. (2022). The World Mental Health Report: transforming mental health for all. World psychiatry : official journal of the World Psychiatric Association (WPA), 21(3), 391–392. https://doi.org/10.1002/wps.21018

[61] Friedman SA, Goldschmidt K. Let Me Introduce You to Your First Virtual Patient. J Pediatr Nurs. 2014;29(3):281-283.

[62] Gao J, Galley M, Li L. Neural approaches to conversational AI. SIGIR, 2018.

[63] Ghosal, D., Majumder, N., Gelbukh, A., Mihalcea, R., & Poria, S. (2020). COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. FINDINGS.

[64] Gratch J, Lucas G, King A, Morency L-P. It's Only a Computer: The Impact of Human- agent Interaction in Clinical Interviews. In: Proceedings of the 2014 International Con- ference on Autonomous Agents and Multi-Agent Systems. 2014; 85-92.

[65] Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing Machines. ArXiv, abs/1410.5401.

[66] Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S.R., & Smith, N.A. (2018). Annotation Artifacts in Natural Language Inference Data. NAACL.

[67] Gutierrez-Maldonado J, Ferrer-Garcia M. Are virtual patients effective to train diagnos- tic skills? In: Proc 19th ACM Symp Virtual Real Softw Technol. 2013:267.

[68] Harless WG, Zier MA, Duncan RC, Hudak JL, McGarvey MD, McLeod DG. An assess- ment of the virtual conversations method for prostate cancer patient education. Urol Nurs. 2007;27(6):499-506.

[69] Harless WG, Zier MA, Harless MG, et al. Evaluation of a virtual dialogue method for breast cancer patient education. Patient Educ Couns. 2009;76(2):189-195.

[70] Henderson M, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2019. ConveRT: efficient and accurate conversational representations from transformers. arXiv:1911.03688.

[71] Hermann, K.M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). Teaching Machines to Read and Comprehend. NIPS.

[72] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9, 1735-1780.

[73] Hojat M. Empathy and Patient Outcomes. (2016) Empathy in Health Professions Education and Patient Care. Springer International Publishing. p. 189-201. http://www.link.springer.com/10.1007/978-3-319-27625-0_11.

[74] Hong, R.H., Murphy, J.K., Michalak, E.E., Chakrabarty, T., Wang, Z., Parikh, S.V., Culpepper, L., Yatham, L.N., Lam, R.W., & Chen, J. (2021). Implementing Measurement-Based Care for Depression: Practical Solutions for Psychiatrists and Primary Care Physicians. Neuropsychiatric Disease and Treatment, 17, 79 - 90.

[75] Houser SH, Ray MN, Maisiak, R, Panjamapirom, A, Willing, J, Schiff, G. D, & Berner, E. S. (2013). Telephone follow-up in primary care: Can interactive voice re- sponse calls work? Studies in Health Technology and Informatics, 192(1–2), 112–116. https://doi.org/10.3233/978-1-61499-289-9-112

[76] HRSA (2019) Health Workforce Shortage Areas. Retrieved January 2019, from https://data.hrsa.gov/topics/health-workforce/shortage-areas

[77] Hubal RC, Day RS. Informed consent procedures: An experimental test using a virtual character in a dialog systems training application. J Biomed Inform. 2006;39(5):532-540.

[78] Huberty J, Green J, Puzia M, Stecher C. (2021). Evaluation of Mood Check-in Feature for Participation in Meditation Mobile App Users: Retrospective Longitudinal Analysis. JMIR Mhealth Uhealth, 9(4):e27106

[79] Hwang, J.D., Bhagavatula, C., Le Bras, R., Da, J., Sakaguchi, K., Bosselut, A., & Choi, Y. (2021). COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs. AAAI.

[80] Inkster, B., Sarda, S., & Subramanian, V. (2018). An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. JMIR mHealth and uHealth, 6.

[81] Jeong S, Breazeal C. Toward Robotic Companions that Enhance Psychological Wellbeing with Smartphone Technology. In: Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, 2017:345-346.

[82] Jetty A, Petterson S, Westfall JM, Jabbarpour Y. Assessing Primary Care Contributions to Behavioral Health: A Cross-sectional Study Using Medical Expenditure Panel Survey. Journal of Primary Care & Community Health. 2021;12. doi:10.1177/21501327211023871

[83] Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022). MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 7184–7190, Marseille, France. European Language Resources Association.

[84] Jim, H.S., Hoogland, A.I., Brownstein, N.C., Barata, A., Dicker, A.P., Knoop, H., Gonzalez, B.D., Perkins, R.M., Rollison, D.E., Gilbert, S.M., Nanda, R.H., Berglund, A.E., Mitchell, R., & Johnstone, P.A. (2020). Innovations in research and clinical care using patient-generated health data. CA: A Cancer Journal for Clinicians, 70.

[85] Johnson K, Poon A, Shiffman S, Lin R, Fagan L. A history-taking system that uses continuous speech recognition. Proceedings Symp Comput Appl Med Care. 1992;757-761.

[86] Johnson TR, Lyons R, Chuah JH, Kopper R, Lok BC, Cendan JC. Optimal learning in a virtual patient simulation of cranial nerve palsies: The interaction between social learning context and student aptitude. Med Teach. 2013;35(1):1-17.

[87] Johnson TR, Lyons R, Kopper R, Johnsen KJ, Lok BC, Cendan JC. Virtual patient simulations and optimal social learning context: A replication of an aptitude-treatment interaction effect. Med Teach. 2014;36(6):486-494.

[88] Jordan MA. The Role of the Health Coach in a Global Pandemic. Glob Adv Health Med. 2021 Aug 9;10:21649561211039456. doi: 10.1177/21649561211039456. PMID: 34395059; PMCID: PMC8361512.

[89] Kang Y, Tan A, Miao C. An Adaptive Computational Model for Personalized Persuasion. IJCAI, 2015.

[90] Kaplan B, Farzanfar R, Friedman RH. Personal relationships with an intelligent interac- tive telephone health behavior advisor system: a multimethod study using surveys and ethnographic interviews. Int J Med Inform. 2003;71(1), 33–41.

[91] Kearns, W.R., Chi, N., Choi, Y.K., Lin, S., Thompson, H.J., & Demiris, G. (2019). A Systematic Review of Health Dialog Systems. Methods of information in medicine, 58 6, 179-193 .

[92] Kenny P, Parsons TD, Gratch J, Rizzo AA. Evaluation of Justina: A virtual patient with PTSD. Lect Notes Comput Sci. 2008;5208 LNAI:394-408.

[93] Kenny PG, Parsons TD, Rizzo A. A comparative analysis between experts and novices interacting with a virtual patient with PTSD. Annu Rev CyberTherapy Telemed. 2009;7(1):122-124. doi:10.3233/978-1-60750-017-9-122

[94] Klingenbjerg, P.M. (2016). Smartphone-Based Conversational Agents and Responses to Questions about Mental Health, Interpersonal Violence, and Physical Health. Journal of Emergency Medicine, 51, 340.

[95] Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. NPJ Digital Medicine. 2021 Jun 3;4(1):1-3.

[96] Krishnan V, Foster A, Kopper R, Lok B. Virtual human personality masks: A human computation approach to modeling verbal personalities in virtual humans. Lect Notes Comput Sci. 2012;7502 LNAI:146-152.

[97] Kroeger, Paul (2005). Analyzing Grammar: An Introduction. Cambridge: Cambridge University Press. p. 54.

[98] Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc. 2018;25(9):1248–1258.

[99] Laster N, Chanda N. Holsey, Derek G. Shendell, Frances A. Mccarty, and Marianne Celano. 2009. Barriers to asthma management among urban families: Caregiver and child perspectives. Journal of Asthma 46, 7: 731-739. DOI: https://doi.org/10.1080/02770900903082571

[100] Leo A, Schuelke M, Hunt D, Miller J, Areán P, Cheng A Digital Mental Health Intervention Plus Usual Care Compared With Usual Care Only and Usual Care Plus In-Person Psychological Counseling for Orthopedic Patients With Symptoms of Depression or Anxiety: Cohort Study. JMIR Form Res 2022;6(5):e36203

[101] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. ACL.

[102] Li J, Monroe W, Ritter A, Jurafsky D, Galley M, Gao J. Deep Reinforcement Learning for Dialogue Generation. In Proceedings of EMNLP. 2016.

[103] Li J, Monroe W, Shi T, Jean S, Ritter A, Jurafsky D. Adversarial Learning for Neural Dialogue Generation. In Proceedings of EMNLP. 2017.

[104] Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. IJCNLP.

[105] Lieu, T. A., Altschuler, A., Weiner, J. Z., East, J. A., Moeller, M. F., Prausnitz, S., Reed, M. E., Warton, E. M., Goler, N., & Awsare, S. (2019). Primary Care Physicians' Experiences With and Strategies for Managing

Electronic Messages. JAMA network open, 2(12), e1918287. https://doi-org.offcampus.lib.washington.edu/10.1001/jamanetworkopen.2019.18287

[106] Lin CJ, Pao CW, Chen YH, Liu CT, Hsu HH. Ellipsis and Coreference Resolution in a Computerized Virtual Patient Dialogue System. J Med Syst. 2016;40(9).

[107] Lindquist, K.A., Wager, T.D., Bliss-Moreau, E., Kober, H., & Barret, L.F. (2012). What are emotions and how are they created in the brain? Behavioral and Brain Sciences, 35, 172 - 202.

[108] Liu, N.F., Gardner, M., Belinkov, Y., Peters, M.E., & Smith, N.A. (2019). Linguistic Knowledge and Transferability of Contextual Representations. ArXiv, abs/1903.08855.

[109] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2022). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Computing Surveys (CSUR).

[110] Louie AK, Coverdale J, Roberts LW. Virtual patients as novel teaching tools in psychiatry. Acad Psychiatry. 2007;31(1):64.

[111] Lucas GM, Gratch J, King A, Morency LP. It's only a computer: Virtual humans increase willingness to disclose. Comput Human Behav. 2014;37:94-100.

[112] L´opez V, Eisman EM, Castro JL. A tool for training primary health care medical stu- dents: The virtual simulated patient. In: Proc ICTAI. 2008;2:194-201.

[113] Majumder, N., Hong, P., Peng, S., Lu, J., Ghosal, D., Gelbukh, A., Mihalcea, R., & Poria, S. (2020). MIME: MIMicking Emotions for Empathetic Response Generation. ArXiv, abs/2010.01454.

[114] Malcarne VL, Ko CM, Roesch SC, Banthia R, Sadler GR. Efficacy of problem-solving therapy for spouses of men with prostate cancer: A randomized controlled trial. Psychooncology. 2019;28(3):497-504. doi:10.1002/pon.4964

[115] Maslow, A.H. (1954). Motivation and Personality.

[116] Mathias JR, Dodd ME, Walters KB, Yoo SK, Erik A, Huttenlocher A. Closing the feedback loop: an interactive voice response system to provide follow-up and feedback in primary care settings. J Med Syst. 2013;37(2):9905.

[117] Mazza M, Mariano M, Peretti S, Masedu F, Pino MC, Valenti M. The Role of Theory of Mind on Social Information Processing in Children With Autism Spectrum Disorders: A Mediation Analysis. J Autism Dev Disord. 2017 May;47(5):1369-1379. doi: 10.1007/s10803-017-3069-5. PMID: 28213839.

[118] Mehta, P., & Pandya, D. (2020). A Review On Sentiment Analysis Methodologies, Practices And Applications. International Journal of Scientific & Technology Research, 9, 601-609.

[119] Melissa H. Bellin, Cassie Land, Angelica Newsome, Joan Kub, Shawna S. Mudd, Mary Elizabeth Bollinger, and Arlene M. Butz. 2017. Caregiver perception of asthma management of children in the context of poverty. Journal of Asthma 54, 2: 162-172. DOI:https://doi.org/10.1080/02770903.2016.119 8375

[120] Mercer SW, Reynolds WJ. Empathy and quality of care. Br J Gen Pract. 2002 Oct;52 Suppl(Suppl):S9-12. PMID: 12389763; PMCID: PMC1316134.

[121] Mikolov, T., Chen, K., Corrado, G.S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. International Conference on Learning Representations.

[122] Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6(7).

[123] Montenegro, J.L., Costa, C.A., & Righi, R.D. (2019). Survey of conversational agents in health. Expert Syst. Appl., 129, 56-67.

[124] Morey DJ. Development and Evaluation of Web-Based Animated Pedagogical Agents for Facilitating Critical Thinking in Nursing. Nurs Educ Perspect. 2012;33(2):116-120.

[125] Mottaghi, S., Poursheikhali, H., & Shameli, L. (2020). Empathy, compassion fatigue, guilt and secondary traumatic stress in nurses. Nursing ethics, 27(2), 494–504.

[126] Mouza AM. IVR and administrative operations in healthcare and hospitals. J Healthc Inf Manag. 2003;17(1):68-71. http://www.ncbi.nlm.nih.gov/pubmed/12553225.

[127] Nienhuis, J. B., Owen, J., Valentine, J. C., Winkeljohn Black, S., Halford, T. C., Parazak, S. E., Budge, S., & Hilsenroth, M. (2018). Therapeutic alliance, empathy, and genuineness in individual adult psychotherapy: A meta-analytic review. Psychotherapy research : journal of the Society for Psychotherapy Research, 28(4), 593–605.

[128] Novielli N, Mazzotta I, De Carolis B, Pizzutilo S. Analysing user's reactions in advice- giving dialogues with a socially intelligent ECA. Cogn Process. 2012;13 Suppl 2:487-497.

[129] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L.E., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., & Lowe, R.J. (2022). Training language models to follow instructions with human feedback. ArXiv, abs/2203.02155.

[130] Parsons TD, Kenny P, Ntuen CA, et al. Objective structured clinical interview training using a virtual human patient. Stud Health Technol Inform. 2008;132:357-362.

[131] Pennington, J., Socher, R., & Manning, C.D. (2014). GloVe: Global Vectors for Word Representation. Conference on Empirical Methods in Natural Language Processing.

[132] Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-Year-Olds' Difficulty with False Belief: The Case for a Conceptual Deficit. British Journal of Developmental Psychology, 5, 125-137. http://dx.doi.org/10.1111/j.2044-835X.1987.tb01048.x

[133] Picard, R.W. (1997). Affective Computing. MIT Press.

[134] Plutchik, R. (1980). A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION.

[135] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Inf. Fusion, 37, 98-125.

[136] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. ArXiv, abs/1810.02508.

[137] Posner, J., Russell, J.A., & Peterson, B.S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and Psychopathology, 17, 715 - 734.

[138] Premack, D., & Woodruff, G. (1978). Does the Chim,panzee Have a Theory of Mind? Behavioral and Brain Sciences, 4, 515-526. http://dx.doi.org/10.1017/S0140525X00076512

[139] Prochaska, J.J., Vogel, E.A., Chieng, A., Kendra, M.S., Baiocchi, M.T., Pajarito, S., & Robinson, A. (2021). A Therapeutic Relational Agent for Reducing Problematic Substance Use (Woebot): Development and Usability Study. Journal of Medical Internet Research, 23.

[140] Provoost S, Lau HM, Ruwaard J, Riper H. Embodied Conversational Agents in Clinical Psychology: A Scoping Review. J Med Internet Res. 2017;19(5):e151.

[141] Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.

[142] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

[143] Rashkin, H., Smith, E.M., Li, M., & Boureau, Y. (2018). I Know the Feeling: Learning to Converse with Empathy. ArXiv, abs/1811.00207.

[144] Reed C, Boswell B, Neville R. Multi-agent Patient Representation in Primary Care.

[145] Reid RD, Pipe AL, Quinlan B, Oda J. Interactive voice response telephony to promote smoking cessation in patients with heart disease: a pilot study. Patient Educ Couns. 2007; 66(3), 319–326.

[146] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. ArXiv, abs/1908.10084.

[147] Rivera-Gutierrez D, Welch G, Lincoln P, et al. Shader Lamps Virtual Patients: The physical manifestation of virtual patients. Stud Health Technol Inform. 2012;173:372- 378.

[148] Safer, M.A., & Keuler, D.J. (2002). Individual differences in misremembering pre-psychotherapy distress: personality and memory distortion. Emotion, 2 2, 162-78 .

[149] Sahler OJZ, Varni JW, Fairclough DL, et al. Problem-solving skills training for mothers of children with newly diagnosed cancer: a randomized trial. J Dev Behav Pediatr JDBP. 2002;23(2):77-86. doi:10.1097/00004703-200204000-00003

[150] Salari, N., Hosseinian-Far, A., Jalali, R. et al. Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: a systematic review and meta-analysis. Global Health 16, 57 (2020). https://doi.org/10.1186/s12992-020-00589-w

[151] Samson, L. W., Tarazi, W., Turrini, G., & Sheingold, S. (2021). Medicare beneficiaries' use of telehealth in 2020: Trends by beneficiary characteristics and location. ASPE. Retrieved November 16, 2022, from https://aspe.hhs.gov/reports/medicare-beneficiaries-use-telehealth-2020

[152] Sanjo Y, Yokoyama T, Sato S, Ikeda K, Nakajima R. Ergonomic automated anesthesia recordkeeper using a mobile touch screen with voice navigation. J Clin Monit Comput. 1999;15(6):347-356.

[153] Sansen H, Boudy J, Chollet G, Milhorat P. vAssist: Building the Personal Assistant for Dependent People - Helping Dependent People to Cope with Technology through Speech Interaction. In: Proc Int Conf Heal Informatics. 2014:490-495.

[154] Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., Roof, B., Smith, N.A., & Choi, Y. (2019). ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. ArXiv, abs/1811.00146.

[155] Sarikaya R, Crook PA, Marin A, Jeong M, Robichaud JP, Celikyilmaz A, Radostev V. An overview of end-to-end language understanding and dialog management for personal digital assistants. In: Proceedings of IEEE Workshop on SLT, 2016;391–397.

[156] Schueller S, Neary M, Lai J, Epstein D. Understanding People's Use of and Perspectives on Mood-Tracking Apps: Interview Study JMIR Ment Health 2021;8(8):e29368 https://mental.jmir.org/2021/8/e29368 doi: 10.2196/29368

[157] Schwartz, R., Dodge, J., Smith, N., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63, 54 - 63.

[158] Serban I, Sordoni A, Bengio Y, Courville AC, Pineau J. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In Proceedings of AAAI. 2015.

[159] Sharma, A., Lin, I.W., Miner, A.S., Atkins, D., & Althoff, T. (2022). Human-AI Collaboration Enables More Empathic Conversations in Text-based Peer-to-Peer Mental Health Support. ArXiv, abs/2203.15144.

[160] Sherrard H, Duchesne L, Wells G, Kearns SA, Struthers C. Using interactive voice response to improve disease management and compliance with acute coronary syn- drome best practice guidelines: A randomized controlled trial. Can J Cardiovasc Nurs. 2015;25(1):10-15.

[161] Shi, W., Shang, Z., Bao, S., & Li, G. (2019). Generation Based on Empathetic Dialogues between Nurses and Patients. Atlantic Press: Advances in Computer Science Research vol. 88.

[162] Shiffman, S., Stone, A.A., & Hufford, M.R. (2008). Ecological momentary assessment. Annual review of clinical psychology, 4, 1-32.

[163] Simonyan K, Dieleman S, Senior A, Graves A. WaveNet: a generative model for raw audio. SSW, 2016.

[164] Sinclair, S., Raffin-Bouchal, S., Venturato, L., Mijovic-Kondejewski, J., & Smith-MacDonald, L. (2017). Compassion fatigue: A meta-narrative review of the healthcare literature. International journal of nursing studies, 69, 9-24 .

[165] Stevens A, Hernandez J, Johnsen K, et al. The use of virtual patients to teach medical students history taking and communication skills. Am J Surg. 2006;191(6):806-811.

[166] Stinson, L., Liu, Y. & Dallery, J. Ecological Momentary Assessment: A Systematic Review of Validity Research. Perspect Behav Sci 45, 469–493 (2022). https://doi.org/10.1007/s40614-022-00339-w

[167] Stratou G, Morency LP, Devault D, et al. A demonstration of the perception system in SimSensei, a virtual human application for healthcare interviews. In: Proceedings of the Int Conf Affect Comput Intell Interact ACII. 2015. 2015:787-789.

[168] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360, Online. Association for Computational Linguistics.

[169] Sukhbaatar, S., Szlam, A.D., Weston, J., & Fergus, R. (2015). End-To-End Memory Networks. NIPS.

[170] Talbot TB, Kalisch N, Christoffersen K, Lucas G, Forbell E. Natural Language Under- standing Performance & Use Considerations in Virtual Medical Encounters. Stud Health Technol Inform. 2016;220:407-13.

[171] Teasdale A, Limbers CA. Online assessment of problem-solving skills among fathers of young and school-age children with type 1 diabetes: Associations with parent and child outcomes. J Child Health Care Prof Work Child Hosp Community. 2021;25(3):379-392. doi:10.1177/1367493520942711

[172] Thomas, D.L., & Diener, E. (1990). Memory accuracy in the recall of emotions. Journal of Personality and Social Psychology, 59, 291-297.

[173] Torre JB, Lieberman MD. Putting Feelings Into Words: Affect Labeling as Implicit Emotion Regulation. Emotion Review. 2018;10(2):116-124. doi:10.1177/1754073917742706

[174] Toseland RW, Blanchard CG, McCallion P. A problem solving intervention for caregivers of cancer patients. Soc Sci Med 1982. 1995;40(4):517-528. doi:10.1016/0277-9536(94)e0093-8

[175] Turian, J.P., Ratinov, L., & Bengio, Y. (2010). Word Representations: A Simple and General Method for Semi-Supervised Learning. Annual Meeting of the Association for Computational Linguistics.

[176] Turunen M, Hakulinen J, St˚ahl O, et al. Multimodal and mobile conversational Health and Fitness Companions. Comput Speech Lang. 2011;25(2):192-209.

[177] Vlasov, V., Mosig, J.E., & Nichol, A. (2019). Dialogue Transformers. ArXiv, abs/1910.00486.

[178] Wang, W., Cai, X., Huang, C., Wang, H., Lu, H., Liu, X., & Peng, W. (2021). Emily: Developing An Emotion-affective Open-Domain Chatbot with Knowledge Graph-based Persona. ArXiv, abs/2109.08875.

[179] Washington, K.T., Demiris, G., Parker Oliver, D., Albright, D.L., Craig, K.W., & Tatum, P.E. (2018). Delivering problem-solving therapy to family caregivers of people with cancer: A feasibility study in outpatient palliative care. Psycho-Oncology, 27, 2494 - 2499.

[180] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. ArXiv, abs/2201.11903.

[181] Wenze, S. J., Gunthert, K. C., & German, R. E. (2012). Biases in affective forecasting and recall in individuals with depression and anxiety symptoms. Personality & social psychology bulletin, 38(7), 895–906. https://doi-org.offcampus.lib.washington.edu/10.1177/0146167212447242

[182] Weston, J., Chopra, S., & Bordes, A. (2014). Memory Networks. CoRR, abs/1410.3916.

[183] Weston, J., Ratle, F., Mobahi, H., & Collobert, R. (2008). Deep learning via semi-supervised embedding. International Conference on Machine Learning.

[184] Willcox, G.F. (1982). The Feeling Wheel A Tool for Expanding Awareness of Emotions and Increasing Spontaneity and Intimacy. Transactional Analysis Journal, 12, 274-276.

[185] Williams, M.T., Lewthwaite, H., Fraysse, F., Gajewska, A., Ignatavicius, J., & Ferrar, K.E. (2021). Compliance With Mobile Ecological Momentary Assessment of Self-Reported Health-Related Behaviors and Psychological Constructs in Adults: Systematic Review and Meta-analysis. Journal of Medical Internet Research, 23.

[186] Wu, L.Y., Fisch, A., Chopra, S., Adams, K., Bordes, A., & Weston, J. (2018). StarSpace: Embed All The Things! ArXiv, abs/1709.03856.

[187] Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J.R., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G.S., Hughes, M., & Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. ArXiv, abs/1609.08144.

[188] Yang, K., Zhang, T., & Ananiadou, S. (2022). A mental state Knowledge–aware and Contrastive Network for early stress and depression detection on social media. Information Processing & Management.

[189] Ye, J. (2021). The impact of electronic health record-integrated patient-generated health data on clinician burnout. Journal of the American Medical Informatics Association : JAMIA.

[190] Yi, J., Kim, M.A., Choi, K., Droubay, B.A., & Kim, S. (2019). Compassion satisfaction and compassion fatigue among medical social workers in Korea: the role of empathy. Social Work in Health Care, 58, 970 - 987.

[191] Young S, Gaˇsi M, Thomson B, Williams JD. POMDP-based statistical spoken dialogue systems: a Review. Proc IEEE, 2013;101(5), 1160–1179.

[192] Yuwen W, Frances M. Lewis, Amy J. Walker, and Teresa M. Ward. 2017. Struggling in the dark to help my child: Parents' experience in caring for a young child with juvenile idiopathic arthritis. Journal of Pediatric Nursing 37: e23-e29.

[193] Yuwen W, Maida Lynn Chen, Kevin C. Cain, Sarah Ringold, Carol A. Wallace, and Teresa M. Ward. 2016. Daily sleep patterns, sleep quality, and sleep hygiene among parent-child dyads of young children newly diagnosed with juvenile idiopathic arthritis and typically developing children. Journal of Pediatric Psychology 41, 651- 660. DOI:10.1093/jpepsy/jsw007

[194] Zahiri, S.M., & Choi, J.D. (2018). Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks. ArXiv, abs/1708.04299.

[195] Zhang, A., Park, S., Sullivan, J.E., & Jing, S. (2018). The Effectiveness of Problem-Solving Therapy for Primary Care Patients' Depressive and/or Anxiety Disorders: A Systematic Review and Meta-Analysis. The Journal of the American Board of Family Medicine, 31, 139 - 150.

[196] Zhang, L., Ren, Z., Jiang, G., Hazer-Rau, D., Zhao, C., Shi, C., Lai, L., & Yan, Y. (2021). Self-Oriented Empathy and Compassion Fatigue: The Serial Mediation of Dispositional Mindfulness and Counselor's Self-Efficacy. Frontiers in psychology, 11, 613908.

[197] Zhou, H., Huang, M., Zhang, T., Zhu, X., & Liu, B. (2018). Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. ArXiv, abs/1704.01074.

[198] Zhu, L., Pergola, G., Gui, L., Zhou, D., & He, Y. (2021). Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. ACL.