MedNgage: A Dataset for Understanding Engagement in Patient-Nurse Conversations

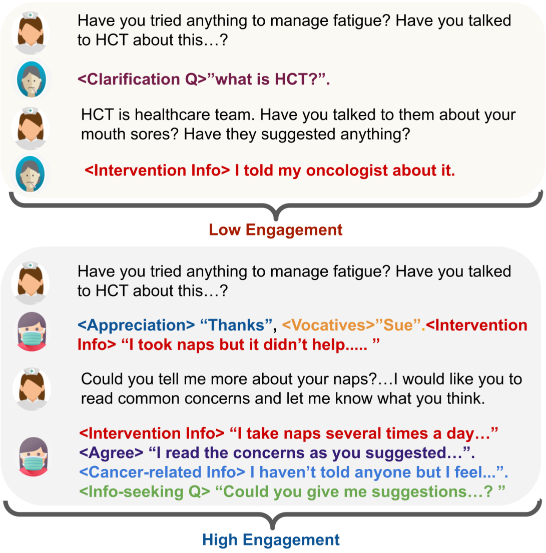
Yan Wang[1], **Heidi Ann Scharf Donovan**[1], **Sabit Hassan**[2], **Malihe Alikhani**[2]
[1] School of Nursing, [2] School of Computing and Information
University of Pittsburgh, Pittsburgh, PA
{yaw75,donovanh,sabit.hassan,malihe}@pitt.edu

Abstract

Patients who effectively manage their symptoms often demonstrate higher levels of *engagement* in conversations and interventions with healthcare practitioners. This engagement is multifaceted, encompassing cognitive and socio-affective dimensions. Consequently, it is crucial for AI systems to understand the engagement in natural conversations between patients and practitioners to better contribute toward patient care. In this paper, we present a novel dataset (MedNgage), which consists of patient-nurse conversations about cancer symptom management. We manually annotate the dataset with a novel framework of categories of patient engagement from two different angles, namely: i) socio-affective engagement (**3.1K** spans), and ii) cognitive engagement (**1.8K** spans). Through statistical analysis of the data that is annotated using our framework, we show a positive correlation between patient symptom management outcomes and their engagement in conversations. Additionally, we demonstrate that pre-trained transformer models fine-tuned on our dataset can reliably predict engagement categories in patient-nurse conversations. Lastly, we use LIME (Ribeiro et al., 2016) to analyze the underlying challenges of the tasks that state-of-the-art transformer models encounter. The de-identified data is available for research purposes upon request [1].

1 Introduction

Due to the ease of use and efficiency of digital health interventions (DHIs) (Greaves et al., 2018), we are witnessing a surge in online conversations between patients and healthcare providers. Literature suggests that actively engaged patients are more likely to obtain the full benefits of an intervention and exhibit better outcomes (Yardley et al., 2016). Therefore, it is critical to understand patients' *engagement* in online conversations and extract insights to aid healthcare professionals in providing in-time support to the patients.


Have you tried anything to manage fatigue? Have you talked to HCT about this…?

<Clarification Q>"what is HCT?".

HCT is healthcare team. Have you talked to them about your mouth sores? Have they suggested anything?

<Intervention Info> I told my oncologist about it.

Have you tried anything to manage fatigue? Have you talked to HCT about this…?

<Appreciation> "Thanks", <Vocatives>"Sue".<Intervention Info> "I took naps but it didn't help..... "

Could you tell me more about your naps?…I would like you to read common concerns and let me know what you think.

<Intervention Info> "I take naps several times a day…"
<Agree> "I read the concerns as you suggested…".
<Cancer-related Info> I haven't told anyone but I feel…".
<Info-seeking Q> "Could you give me suggestions…? "

Figure 1: Our dataset contains patient-nurse conversations annotated with cognitive and socio-affective engagement. We hypothesize that patients who have high engagement tend to have better symptom control.

While previous research has made progress in modeling engagement in human-human and human-agent conversations (Reddy et al., 2021; Sano et al., 2016; Xu et al., 2020), these works do not translate well to modeling engagement in patient-provider conversations. To bridge this gap, we introduce a novel resource called **MedNgage**, which consists of patient-private asynchronous message boards from an online intervention led by study nurses for symptom management. The dataset encompasses 2.1K turns between 68 patients and the nurses. We hypothesize that patient socio-affective engagement (e.g., sharing cancer-related experiences) and cognitive engagement (e.g., information-seeking question) can predict their symptom management outcomes. It is important to note that datasets containing patient-provider conversations are rarely found in the ex-

[1] https://github.com/YanRayray/MedNgage

Findings of the Association for Computational Linguistics: ACL 2023, pages 4613–4630
July 9-14, 2023 ©2023 Association for Computational Linguistics

isting literature.

To effectively model engagement in our proposed dataset, we create a novel framework called Socio-Affective Cognitive Engagement (SACe), drawing inspiration from linguistic theories of discourse (Asher et al., 2003), cognitive science of grounding in communication (Clark and Brennan, 1991), and the model of social presence (Swan et al., 2009). With SACe, we manually annotate the dataset, classifying patient engagement into eight categories of socio-affective engagement (e.g., personal information (Swan et al., 2009)) across 3.1K spans and seven categories of cognitive engagement (e.g., information-seeking question, clarification question (Asher et al., 2003)) across 1.8K spans within conversations. Additionally, we investigate the extent to which this framework can predict patient symptom management outcomes. Figure 1 showcases excerpts from our dataset, which have been annotated using the SACe framework.

We conduct Spearman's rank correlation and Kruskal-Wallis tests to analyze the relationship between engagement categories and changes in patients' perceived symptom control. The tests provide empirical evidence that patients who have a higher level of engagement in certain categories gain more control over their symptoms. Thus, automated prediction of engagement from conversations can help practitioners to detect low patient engagement levels and identify effective intervention strategies so that they could tailor content for improved outcomes. To facilitate this, we conduct a range of experiments evaluating the efficacy of both traditional machine learning (e.g., SVMs) and state-of-the-art transformer models in predicting engagement, followed by an analysis using LIME (Ribeiro et al., 2016) to identify the challenges of the tasks. Thus, our contributions to this paper are:

- We present a novel framework for measuring engagement in patient-nurse conversations, as well as a novel dataset annotated with (i) **socio-affective** and (ii) **cognitive** engagement. The framework can be generalized to interactions in other healthcare scenarios that involve adopting healthy behaviors and improving clients' mental and physical well-being.

- With our statistical tests, we show a **positive correlation** between patient engagement level in conversations and perceived symptom control.

- We show transformer models can reliably predict engagement (F1 score > **78**) when trained on our data to help practitioners identify patients with low engagement and intervene accordingly. We analyze classifier errors using LIME to provide insights into the challenges of the tasks.

2 Framework

Inspired by theories of discourse coherence, we explore linguistic and conceptual models of engagement across various disciplines. Through iterative collaboration with expert nurses and linguists during the initial content analysis phase, we have generated a set of categories that effectively capture the nuances of our data. Therefore, our categorization of engagement combines existing literature with data-driven adaptation to match the distinctive features of our dataset, MedNgage. We will discuss the theoretical underpinnings behind our proposed SACe framework and the categories of engagement for annotating our dataset.

Socio-Affective Engagement Interventions are social: in our online nurse-led symptom management intervention, the nurses worked to create a relationship with the patient based on trust, respect, and closeness and encouraged patients to talk about their feelings and concerns (Phillips, 2016). Thus, it is important for patients to feel affectively connected to the nurses and intervention. Adapted from social presence in the Community of Inquiry (CoI) framework (Swan et al., 2009), for socio-affective engagement, we look for positive evidence of developing an affective connection (a trusting, close, and respectful therapeutic relationship) with the nurses. Appreciation, emotional expression, self-disclosure, greetings and salutations, and vocatives are examples of CoI categories that signal patient engagement in conversations. The mapping of CoI framework (Swan et al., 2009) to categories of socio-affective engagement is in Appendix A. However, CoI focused on online learning in higher education. Engagement indicators such as patients sharing their cancer experiences and interest in further communication are not included in this theory. Moreover, we excluded classes such as "humor" that do not appear commonly in our context. The eight final socio-affective categories are listed in Table 1.

Cognitive Engagement The primary goal of conversations between a patient and nurse in our dataset is to complete the intervention following

Socio-affective engagement	**Examples**	**Count**
Appreciation	"I really like the way you edited my goals.Thank you!"	343
Positive sentiment	"I feel that all the strategies are good ones and THEY WORK!!"	357
Negative sentiment	"I have now lost 2 replies! This will be my last try for tonight."	25
Cancer-related experience	"I have caught myself cycling through emotional fatigue times where I'm completely frustrated with myself and life and want to give up and then have some good news or an outing and everything is back to fine again.	262
Personal information	"Yea. My daughter and son-in-law came to visit. We had a great " Thanksgiving dinner.	385
Interest in communication	"Please let me know if you need any other information."	161
Vocatives	The patient addresses the study nurse by the name.	612
Greetings	"Sincerely,", "Good morning", "With God's grace on us both,"	956
Total Socio-affective spans		*3101*
Cognitive engagement	**Examples**	**Count**
Intervention information	"It's hard to get up. I try to sleep late (until 8 or 8:30) and get in to work by 10, although I'd rather sleep until 10. I force myself to get up."	1224
Information-seeking question	"Are there other foods which make sores worse?"	177
Clarification question	"Are you talking about a based on a 1 to 10 type of thing?"	14
Acknowledgment	"I have received and read your email."	50
Agreement	"The care plan looks good. I don't have anything to add."	298
Disagreement	"He won't read the material, his mind is made up."	13
Initiative taking	"I think we can start with the sleep disturbance first. I am working on my eating now by myself, so that will take a while to see the affects of that, so let's move on to the next one, shall we."	17
Total Cognitive engagement spans		*1793*

Table 1: Examples and total counts of patient engagement categories from the asynchronous message boards.

the intervention protocols with some degree of individualization. To do so, it takes the nurse and the patient together to coordinate the content and plan of action on the message boards. They do it by building common ground (mutual understanding, assumption, and knowledge) and updating their common ground moment by moment (Clark and Brennan, 1991). Therefore, for cognitive engagement, we look for the positive evidence of patients coordinating the content and process of the intervention with the nurse within collaborative goals—acknowledgment and initiation of the relevant next turn (e.g., answering protocolized questions, clarification questions), based on Clark and Brennan (1991). The mapping of grounding in communication framework (Clark and Brennan, 1991) to categories of cognitive engagement is in Appendix A. Moreover, we found *taking initiative* to be an important patient engagement marker in our corpus that is not defined by Clark and Brennan (1991). This category demonstrates patients taking the initiative on managing their symptoms without being prompted by the nurses.

We have also studied discourse frameworks such as Segmented Discourse Representation Theory (SDRT) (Asher et al., 2003). Discourse relation classes from SDRT such as information-seeking question, and clarification question overlap with (Clark and Brennan, 1991) and belong to our framework. The background class is closely related to Cancer-related experience in our data. Although there is an overlap between frameworks of engagement and discourse, engagement is a broader construct that captures a speaker's involvement in a conversation. While the discourse frameworks focus on logical and semantic relationships at Elementary Discourse Unit (EDU) level, engagement is observed across multiple EDUs in a conversation. As such, the remaining categories of SDRT and other discourse frameworks are not applicable directly. The final seven categories of cognitive engagement are listed in Table 1.

3 Dataset

In this section, we describe our data source, followed by our annotation protocol and a brief discussion about the patient characteristics in our dataset.

3.1 Data Source

The data source for our work is asynchronous message boards between patients[2] with recurrent ovarian cancer and study nurses. All interactions be-

[2] Patient informed consent and approval of our institutional review board were obtained.

# Patients	# Patient Turns	# Nurse Turns	# Unique Tokens
68	1K	1.1K	29K

# Socio-affective spans	# Cognitive spans
3101	1793

Table 2: Overview of the annotated dataset.

tween the nurses and patients are captured verbatim. Patients interacted 1:1 with a nurse to go through the theory-guided intervention elements to develop individualized Symptom Care Plans for 3 target symptoms that they hoped to gain better control over (e.g., fatigue, pain, and nausea). A description of the intervention process is listed in Appendix B.

One of the primary outcomes of the intervention is patient self-reported symptom controllability (i.e., an individual's confidence in one's ability to control symptoms with medications and behaviors), which was assessed by a validated and reliable measure —Symptom Representation Questionnaire (Donovan et al., 2008). First, the patients completed a 28-item symptom inventory (e.g., pain, fatigue, depression, nausea) and reported symptom severity at its worst in the past week from 0 (did not experience the symptom) to 10 (as bad as I can imagine). The patient then identified three target symptoms they would like to control better. The Symptom Controllability Scale (e.g., "I can do a lot to control this symptom"; five items) was used for each targeted symptom on a 0 (strongly disagree) to 4 (strongly agree) scale at the beginning and the end of the intervention.

3.2 Manual Annotation

Based on the theoretical framework discussed earlier, two trained raters (one graduate and one undergraduate nursing student) independently begin by abstracting the sentences in the patients' posts that reflect cognitive and socio-affective engagement as meaning units. To determine the minimal meaningful units and their respective categories, the raters initiate the analysis by examining the first word of a patient's post. We then gradually expand the analysis, observing for points where the category of engagement changed. When a category shift is identified, the section analyzed up to that point is marked as a "span." If the current span does not fall under either the cognitive or socio-affective engagement category, it is skipped, and the analysis continues until the next category shift is detected. This iterative process is repeated until the entire post is classified into smaller sections or spans, each corresponding to a distinct engagement category. Examples from our annotation are listed in Table 1. An overview of the annotated dataset is provided in Table 2.

Inter-rater Agreement Code differences between the two raters were discussed and decided by the principal investigator of the intervention. A coding scheme with examples (Appendix B) is developed in an iterative manner to ensure inter-rater reliability. Inter-rater reliability is evaluated by Cohen κ statistic across 131 turns by two annotators. Cohen κ for cognitive and socio-affective engagement are 0.87 and 0.86, respectively.

Patient Characteristics In the dataset we annotated, the mean age of the patients was 59.7 (SD = 9.5), ranging from 24 to 83. The majority (75%) were married or living with a partner, and 51.5% had a bachelor's degree and above. Based on the Charlson Comorbidity Index (CCI), 45.6% of the patients did not have any comorbidity, and 54.4% had at least one comorbidity.

4 Analysis

In this section, we first report the distribution of different engagement categories in our data. Then we illustrate the relationship between engagement and symptom controllability. The significance level for statistical analysis is $p < 0.05$.

	Mean of *f* (SD)	**Median of *f* (IQR)**
Socio-affective engagement categories		
Positive sentiment	3.63 (5.12)	2 (4.25)
Negative sentiment	0.35 (0.99)	0 (0)
Appreciation	3.96 (4.27)	3 (5.25)
Cancer-related experience	2.76 (2.97)	2 (3.25)
Personal information	4.38 (3.84)	3 (4)
Vocatives	8.6 (9.11)	6.5(10.5)
Interest in communication	2.24 (2.73)	1 (3)
Greetings	8.81(7.48)	7 (8)
Cognitive engagement categories		
Intervention information	12.41 (9.36)	10 (12.25)
Acknowledgement	0.74 (1.19)	0 (1)
Information-seeking question	2.26 (3)	1 (3)
Clarification question	0.21 (0.59)	0 (0)
Agreement	3.38 (3.9)	2 (6)
Disagreement	0.19 (0.4)	0 (0)
Initiative taking	0.22 (0.59)	0 (0)

Table 3: The mean and median of the frequency (*f*) of each engagement category across 68 patients.

Engagement category	Freq-ρ (p)
Socio-affective engagement category	
Positive sentiment	.35 (.008)
Appreciation	.25 (.007)
Cancer-related experience	.34 (.01)
Personal information	.34 (.01)
Vocatives	.33 (.01)
Interest in communication	.32 (.01)
Greetings	.27 (.04)
Cognitive engagement category	
Intervention information	.38 (.003)
Information-seeking question	.26 (.05)
Agreement	.34 (.009)
Initiative taking	.26 (.047)

Table 4: Correlations between controllability score changes and the frequency of each engagement category. Insignificant results are not reported (p > 0.05).

4.1 Distribution of Engagement

The most common socio-affective engagement categories are Greetings (mean = 8.81, SD = 7.48) and Vocatives (mean = 8.6, SD = 9.11). The least common one is expressing Negative sentiment (mean = 0.35, SD = 0.99).

The most commonly used cognitive category is Intervention information (mean = 12.41, SD = 9.36). The least common ones are Disagreement (mean = 0.19, SD = 0.4), Clarification question (mean = 0.21, SD = 0.59), and Initiative taking (mean = 0.22, SD = 0.59). Table 3 lists the mean and median of the frequency of each engagement category.

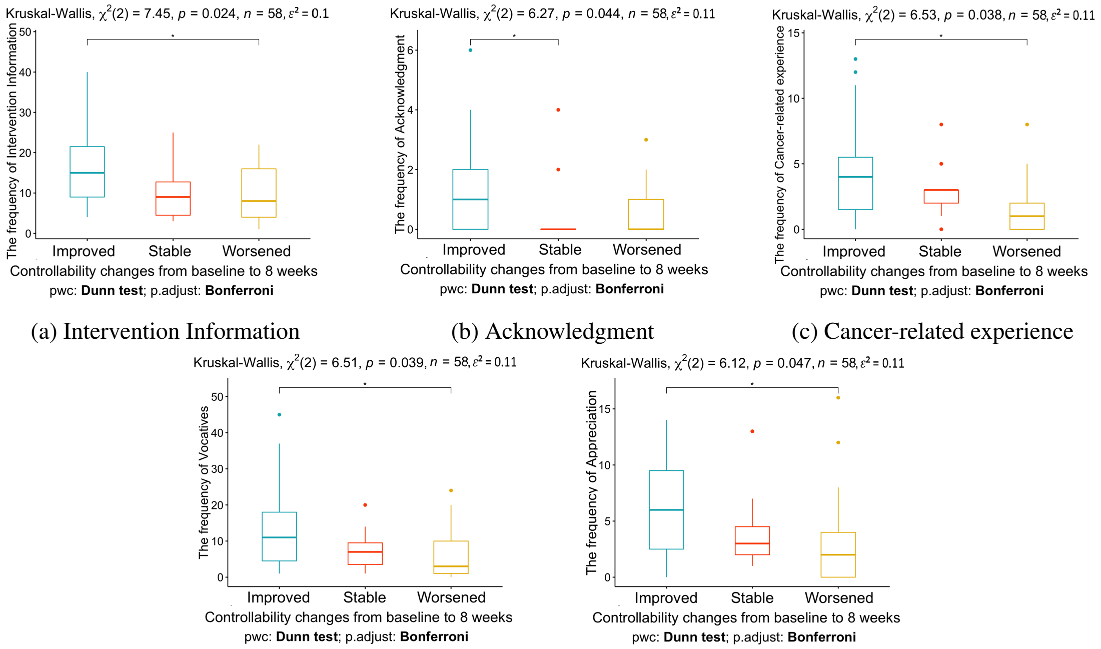
4.2 Symptom Controllability and Engagement

Of 68 patients, 58 reported symptom controllability scores at baseline and at the end of the intervention (8 weeks). The average controllability score change is 0.21 (SD = 0.61), ranging from -1.85 to 1.73. Based on symptom controllability score change, we divided patients into three groups: (1) improved (n = 23) with changes greater than +0.3 SD (> 0.389), (2) stable (n = 14) with changes within +/-0.3 SD (0.029-0.389), and (3) worsened (n = 21) with changes greater than -0.3 SD (0.029).

Spearman's rank correlation was computed to assess the correlations between controllability and engagement. Most categories' frequencies have a significant positive relationship with changes in patient controllability scores from baseline to 8 weeks ($0.2<\rho< 0.4$). Table 4 shows the correlations between patient controllability scores and the frequency of each category. Kruskal-Wallis tests (Kruskal and Wallis, 1952) were used to examine the differences in the frequency of engagement categories among patients with an improved vs. stable vs. worsened sense of control over their symptoms. When a significant difference was detected, a post-hoc Dunn's test (Dunn, 1961) with Bonferroni adjustment (Napierala, 2014) for multiple pairwise comparisons was applied to distinguish the significant and non-significant pairs. We used ϵ^2 to calculate the effect sizes for the Kruskal-Wallis tests and biserial correlation r for post-hoc Dunn's tests. Although the sample size is smaller, we were able to detect a moderate to large effect size.

We discovered variations in the frequency of specific engagement categories between patients who experienced improved control and those with stable or worsening control. Regarding cognitive engagement, patients who reported improved perceived control over symptoms provided information toward intervention (i.e., content, process, and technical issues) more often (Mdn = 17) than those who reported worsened control (Mdn = 8), with an effect size of 0.4 (Figure 2a). Individuals who reported improved control acknowledged the nurse's contributions more frequently (Mdn = 1) than those who had a stable sense of control (Mdn = 0), with an effect size of 0.41 (Figure 2b). Although marginally significant (P = 0.053), patients who reported improved control agreed with the nurse or agreed to do the intervention activities more frequently (Mdn = 4) than those who reported worsened control (Mdn = 1), with an effect size of 0.36.

In terms of socio-affective engagement, individuals who reported improved controllability shared cancer-related experience (e.g., cancer story, vulnerability) more often (Mdn = 4) than those who reported worsened symptom control (Mdn = 1), with an effect size of 0.38 (Figure 2c). Patients who reported an improved sense of control addressed nurses by their names more frequently (Mdn = 11) than those who reported worsened symptom control (Mdn = 3), with an effect size of 0.39 (Figure 2d). Moreover, patients who reported improved controllability appreciated and recognized the nurses' contributions significantly more often (Mdn = 6) than those who reported worsened symptom control (Mdn = 2) (Figure 2e). Although marginally significant (P = 0.053), patients who appeared to report an improved sense of control expressed positive sentiment toward intervention more frequently (Mdn = 4) than those



(a) Intervention Information

(b) Acknowledgment

(c) Cancer-related experience

(d) Vocatives

(e) Appreciation

Figure 2: Kruskal-Wallis tests and post-hoc Dunn's tests on the frequencies of engagement categories between patients with an improved vs stable vs worsened sense of control over their symptoms. * indicates statistical significance ($p \leq 0.05$).

who reported worsened symptom control (Mdn = 1), with an effect size of 0.37.

5 Experiments

In order to assess the feasibility of predicting cognitive and socio-affective engagement, we train traditional SVMs and fine-tuned pre-trained transformer models on our dataset.

5.1 Dataset for experiments

For our experiments, we isolate the cognitive and socio-affective spans and use 80% of the data for training, 10% for validation, and 10% for testing. All numbers are reported on the test set. Due to the relatively small number of instances, we leave out the low-frequency classes ("Disagreement" and "Clarification questions"), yielding eight socio-affective and five cognitive engagement categories.

5.2 Classification

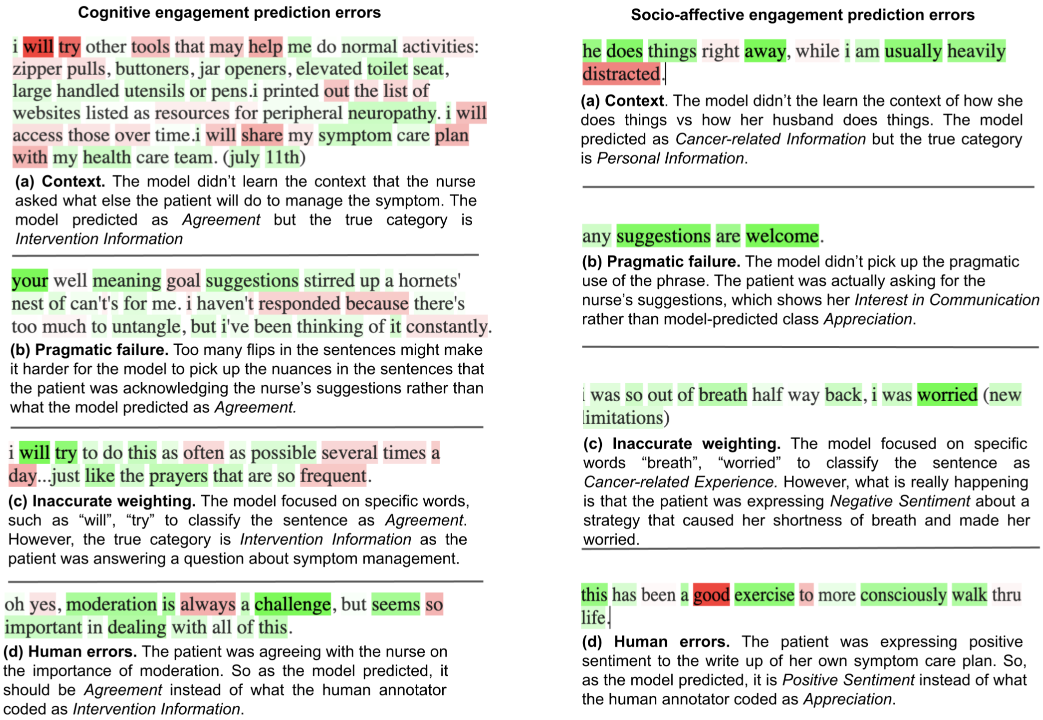
SVMs As baselines, we train SVMs on word bigrams and character n-grams (2-5), weighted by term frequency-inverse document frequency (tf-idf) using scikit-learn's[3] default parameters.

	Socio-Affective		
Model	**P**	**R**	**F1**
SVM-W	81.1 (0.0)	39.8 (0.0)	40.7 (0.0)
SVM-C	**84.7 (0.0)**	59.7 (0.0)	65.6 (0.0)
Bio-BERT	75.4 (2.1)	78.0 (1.1)	76.4 (1.7)
BERT	75.1 (1.1)	76.6 (0.5)	75.6 (0.6)
XLNet	77.4 (1.9)	78.7 (1.6)	77.7 (1.4)
RoBerta	78.6 (1.3)	**79.6 (1.2)**	**78.8 (1.0)**
	Cognitive		
Model	**P**	**R**	**F1**
SVM-W	62.1 (0.0)	58.9 (0.0)	57.6 (0.0)
SVM-C	72.1 (0.0)	73.5 (0.0)	72.5 (0.0)
Bio-BERT	**78.0 (2.1)**	**79.3 (0.4)**	**78.5 (1.3)**
BERT	77.5 (0.9)	78.3 (0.6)	77.6 (0.6)
XLNet	61.9 (0.8)	63.9 (2.3)	62.8 (0.7)
RoBerta	61.5 (2.3)	66.9 (0.2)	63.8 (1.3)

Table 5: Results of predicting socio-affective and cognitive engagement. Means across three runs are reported along with standard deviation in parenthesis. SVM-W and SVM-C refer to SVMs trained with word and character n-grams respectively. Bio-BERT achieves the best performance for Cognitive and RoBerta achieves the best performance for Socio-affective engagement.

Pre-trained transformers We fine-tune four pre-trained transformers: i) bert-base-cased (Devlin et al., 2019), ii) BioClinicalBERT (Lee et al., 2019), iii) RoBERTa-base (Liu et al., 2019) and iv) XLNet-base-cased (Yang et al., 2019). All the transformer

[3] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC



Cognitive engagement prediction errors

i will try other tools that may help me do normal activities: zipper pulls, buttoners, jar openers, elevated toilet seat, large handled utensils or pens.i printed out the list of websites listed as resources for peripheral neuropathy. i will access those over time.i will share my symptom care plan with my health care team. (july 11th)

(a) Context. The model didn't learn the context that the nurse asked what else the patient will do to manage the symptom. The model predicted as *Agreement* but the true category is *Intervention Information*

your well meaning goal suggestions stirred up a hornets' nest of can't's for me. i haven't responded because there's too much to untangle, but i've been thinking of it constantly.

(b) Pragmatic failure. Too many flips in the sentences might make it harder for the model to pick up the nuances in the sentences that the patient was acknowledging the nurse's suggestions rather than what the model predicted as *Agreement.*

i will try to do this as often as possible several times a day...just like the prayers that are so frequent.

(c) Inaccurate weighting. The model focused on specific words, such as "will", "try" to classify the sentence as *Agreement*. However, the true category is *Intervention Information* as the patient was answering a question about symptom management.

oh yes, moderation is always a challenge, but seems so important in dealing with all of this.

(d) Human errors. The patient was agreeing with the nurse on the importance of moderation. So as the model predicted, it should be *Agreement* instead of what the human annotator coded as *Intervention Information*.

Socio-affective engagement prediction errors

he does things right away, while i am usually heavily distracted.

(a) Context. The model didn't the learn the context of how she does things vs how her husband does things. The model predicted as *Cancer-related Information* but the true category is *Personal Information*.

any suggestions are welcome.

(b) Pragmatic failure. The model didn't pick up the pragmatic use of the phrase. The patient was actually asking for the nurse's suggestions, which shows her *Interest in Communication* rather than model-predicted class *Appreciation*.

i was so out of breath half way back, i was worried (new limitations)

(c) Inaccurate weighting. The model focused on specific words "breath", "worried" to classify the sentence as *Cancer-related Experience.* However, what is really happening is that the patient was expressing *Negative Sentiment* about a strategy that caused her shortness of breath and made her worried.

this has been a good exercise to more consciously walk thru life.

(d) Human errors. The patient was expressing positive sentiment to the write up of her own symptom care plan. So, as the model predicted, it is *Positive Sentiment* instead of what the human annotator coded as *Appreciation*.

Figure 3: Examples of LIME output on classifier errors, along with the error type. The socio-affective model emphasizes words that achieve pragmatic usage while the cognitive model emphasizes words with the patient's goal. Green highlighted words have positive weights and contribute to the classifier predicting a certain class, while red highlights have negative weights and reduce the likelihood of that class. The darker the color, the greater the impact of the word on the prediction.

models are trained for 3 epochs under the same settings: learning rate of 8e-8, batch size of 16, and a maximum length of 200 tokens.

Experiment results The experiment results are presented in Table 5. We report macro-averaged precision, recall, and F1 scores. Each experiment is repeated three times and the mean results, along with the standard deviation are reported. Since SVMs are deterministic, the standard deviation is 0. We observe that fine-tuned transformer models can reliably predict engagement in our dataset with mean F1 scores of **78.5** and **78.8** respectively for socio-affective and cognitive categories.

5.3 Error Analysis

Despite the promising results achieved by transformer models, it is important to assess the limitations of these models due to the sensitivity of patient conversations. Thus, we manually annotate all instances where the best model (BioClinicalBERT for cognitive, and RoBerta for socio-affective engagement) make errors. To aid our analysis, LIME (Ribeiro et al., 2016) is used to identify the words and phrases that support the models' selection of a particular class. LIME determines the contribution score of specific words on the prediction of a classifier by generating variations of an input sentence by randomly removing a word and observing changes in the prediction. The "contribution score" represents how much weight the word had in the original prediction by the model.

Observing the output of LIME, we identify four main categories of errors: i) missing context, ii) inaccurate weighting of words; iii) pragmatic failures; and iv) human error in the annotation. Figure 3 shows the examples of errors and Figure 4 shows the weighted percentages of errors based on the count of categories for the two tasks.

Cognitive Errors: Context mistakes and inaccurate weighting account for around 63% of the errors in the prediction of cognitive engagement after being adjusted for the count of the cognitive engagement categories. Due to these two types of errors, the model commits a high rate of errors in the categories of Intervention information and Agreement (15% of the errors examined after the

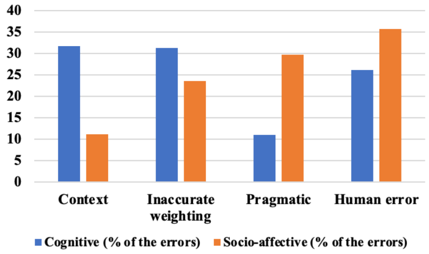


Figure 4: Weighted percentages of errors based on LIME. The socio-affective engagement model tends to make pragmatic errors, while the cognitive model tends to make context errors.

adjustment for each type of error). Of these two categories, human mistake rates also account for 15% of the errors. Due to inaccurate weighting, the model made errors in differentiating the categories of Information-seeking question and Agreement (16% of the errors after adjustment).

Socio-affective Errors: The most frequent error types for socio-affective engagement, apart from human error, are pragmatic failures and inaccurate weighting, which account for 53% of the errors after being adjusted for the count of the socio-affective engagement categories. Due to pragmatic failure (10% of the errors evaluated after the adjustment), the model tends to make mistakes between the categories of appreciation and positive sentiment. Similarly, the human error rate is also the highest between these two categories (10% of the errors analyzed after the adjustment).

We find a large difference in pragmatic errors and context errors between cognitive and socio-affective tasks. We think this is because when predicting socio-affective engagement, the model needs to consider more pragmatic factors (e.g., situational context, the individuals' mental states) than cognitive engagement. For example, in Figure 3b which shows LIME output, we observe that the model focuses on the word "welcome", which is an indicator of "Appreciation", but in practical usage, it was used as interest in communication. On the other hand, detecting cognitive engagement relies more on tone and structure, leading to context errors in Figure 3a.

6 Implications on DHIs

By examining the relationship between patients' engagement and their symptom controllability, our study provides empirical evidence of the potential pathways of patient learning and behavior change to gain better control over symptoms through meaningful engagement. The classification models presented in this paper can be used to track patient engagement based on patient-provider interactions on asynchronous message boards. The models could be used as complementary tools to augment clinicians' capabilities to identify patients with a low engagement level so that they can focus their energy and time (i.e., tailoring content and communication accordingly) to enhance engagement and help those who struggle the most to obtain the maximum benefits of a given DHI. Based on predicted engagement markers, or lack thereof, the provider can tailor the intervention content (e.g., suggest more strategies based on professional experience, segment a series of questions into a few posts, provide timely emotional support) to achieve common ground and cultivate an affective connection with the patient. For example, nurses can encourage the patients to i) provide relevant information to complete intervention activities and tasks (e.g., describing their symptom beliefs), and ii) share their feelings, cancer treatment, and symptom experiences. For patients participating in DHIs to achieve desired outcomes, it is also important to feel close to the provider, for example, addressing the provider by name, acknowledging the provider's message, and recognizing their contributions to the care and support.

Our dataset, MedNgage contains rich narratives of various symptom experiences and management processes from women who were fighting advanced cancer in an online cognitive behavior intervention. The Representational Approach (Donovan et al., 2007) underlying the intervention is disease-agnostic and designed to help clients understand how their disease representations relate to behaviors. Therefore, the annotation system and model developed in this study can be applied to various healthcare scenarios involving the promotion of healthy behaviors and the enhancement of clients' mental and physical well-being. The corpus includes comprehensive representations of patient symptoms, concerns, and the obstacles they faced while seeking the best symptom management. This presents an excellent opportunity for various clinical NLP tasks, such as developing a dialog system, optimizing patient self-report/expression review, generating text summaries of patient conversations, and customizing automated relevant responses to improve patient engagement, as suggested in recent works on summarization (Xu et al., 2020) and style transfer (Atwell et al., 2022). For example, the system can use patient-specific language or generate text that encourages patients to ask questions or express their concerns.

7 Related Work

Engagement in human-human conversation has been studied from both socio-affective and cognitive aspects, resulting in numerous frameworks and theories for modeling engagement. One of the inspirations behind our work, the Community of Inquiry framework (Swan et al., 2009), which builds on the educational philosophy and practice of Dewey, models conversations in the online learning environment in terms of social presence, cognitive presence, and teaching presence. Rourke et al. (1999), proposes a community of inquiry framework that synthesizes pedagogical principles with the dynamics of computer conferencing, focusing on social presence. Clark and Brennan (1991), the inspiration behind the cognitive aspect of engagement in our work, studies how grounding in conversation is shaped by *purpose* and *medium* of the conversation. The authors highlight different grounding references that we adapt to our work.

Previous research has investigated engagement in medical interactions, including conversations. Studies have specifically examined the cognitive engagement of individuals with schizophrenia in conversations, with a focus on medication adherence (McCabe et al., 2013; Howes et al., 2012). In contrast, our study employs various methods to identify and predict patient engagement in a complex intervention aimed at modifying multiple health behaviors, including diet, physical activity, relaxation, and medication adherence. Further, we investigate the impact of engagement on patient symptom outcomes, providing practical strategies for nurses and other practitioners to effectively engage patients and deliver optimal care.

Other studies have explored engagement in different contexts such as political argument settings (Shugars and Beauchamp, 2019), conversations around terrorist attacks (Chiluwa and Odebunmi, 2016), socio-affective aspects of conversations such as emotion (Yu et al., 2004), student engagement in online discussion forums (Liu et al., 2018), cognitive engagement in MOOC forums (Wen et al., 2014), real-time engagement in reducing binge drinking through intervention text messages (Irvine et al., 2017), user engagement in online health communities (Wang et al., 2020). However, none of these studies specifically consider the dynamics of socio-affective and cognitive engagement in online conversations between patients and healthcare providers. Our work is the first to develop a dataset with aggregated annotations and models for computationally modeling engagement in patient-provider scenarios.

8 Conclusions

We have developed a framework **(SACe)** that effectively captures patient engagement in patient-nurse conversations. Through the analysis of a unique dataset **(MedNgage)** consisting of online patient-nurse conversations, we have identified eight categories for socio-affective engagement and seven categories for cognitive engagement. These findings provide valuable insights for behavioral scientists to understand and monitor patients' engagement in healthcare interactions. The approach could be applied in other fields, such as online education, where engagement plays a crucial role. Our analysis confirms that higher levels of engagement, including the increased coordination and emotional connections between patients and healthcare practitioners during the intervention, result in improved symptom control. Additionally, we have demonstrated that fine-tuned transformer models can reliably predict fine-grained engagement in conversations so that practitioners can adjust their communication style and tailor strategies promptly, to promote patient engagement for improved outcomes. Our analysis of model output using LIME has shown the challenges that state-of-the-art transformer models encounter for the two tasks, which can be used to improve these models for similar tasks in the future. We expect our system could also aid subsequent text generation tasks, such as summarization of patient conversations, and tailoring automated responses in clinical settings.

Limitations

While this dataset is unique and pioneering, its size is limited, and it involves specific patients. To enhance the generalizability of the findings, a larger dataset may be required. Similarly, although our framework is innovative, we anticipate the development of more comprehensive and informative annotation protocols in the future. For instance, we observed a higher frequency of the "Intervention information" category within cognitive engagement, likely because the intervention predominantly follows a Q (nurse) & A (patient) format. We hope that the coding scheme established in this study can aid future research in refining this category with finer granularity, based on specific intervention theories.

Ethical Considerations

We used the NLM Scrubber offered by NIH to produce HIPAA-compliant deidentified health information for scientific use, including dates, and places. Two independent annotators evaluated the NLM-Scrubber on the dataset to make sure no events or other people in patients' posts can allow patients to be traceable. The de-identified version of our data will be shared with researchers upon request who have completed an ethical review from their institution and a data request application form from us.

Since the domain of our dataset is specific, the models trained on our dataset may exhibit subtle biases on out-of-domain data. Further, pre-trained models that we use in our work have been shown to exhibit biases (Li et al., 2021). We hope future researchers could use these models with caution regarding the biases that these pre-trained models have.

The long-term goal of our work is to *aid* healthcare providers to quickly identify poorly engaged patients to allocate their energy and resources to provide in-time support. Models trained on our data should not be deployed in the real world without human supervision because, despite the potential of transformer models, they cannot be relied on completely in sensitive medical scenarios.

Acknowledgements

The randomized clinical trial (WRITE Symptoms intervention) was supported by a grant from the National Institutes of Health, National Institute of Nursing Research NIH-NINR R01NR010735-NRG Oncology GOG-259 (H.S.D., PI) as well as NIH-NCI grants to NRG Oncology (U10CA180822), NRG Operations (U10CA180868), and UG1CA189867 (NCORP). This study is supported by the Eta Chapter Sigma Theta Tau International Honor Society of Nursing Research Award (Y.W., PI) and the Margaret E. Wilkes Scholarship Fund Award (Y.W., PI). We extend our gratitude to Hannah Bergquist for her invaluable contributions to the annotation of the asynchronous message boards. Furthermore, we are immensely grateful to all the women who participated in this trial, as their involvement has been instrumental to the success of this study.

References

N. Asher, N.M. Asher, A. Lascarides, S. Bird, Cambridge University Press, B. Boguraev, D. Hindle, M. Kay, D. McDonald, and H. Uszkoreit. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press.

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. Appdia: A discourse-aware transformer-based style transfer model for offensive social media conversations. In *International Conference on Computational Linguistics*.

Innocent Chiluwa and Akin Odebunmi. 2016. On terrorist attacks in nigeria: Stance and engagement in conversations on nairaland. *Communication and the Public*, 1:109 – 91.

Herbert H. Clark and Susan Brennan. 1991. Grounding in communication. In *Perspectives on socially shared cognition*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

John Dewey. My pedagogic creed. *Journal of Education*, 104:542 – 542.

Heidi S. Donovan, Sandra E. Ward, Paula R Sherwood, and Ronald C. Serlin. 2008. Evaluation of the symptom representation questionnaire (srq) for assessing cancer-related symptoms. *Journal of pain and symptom management*, 35 3:242–57.

Heidi S. Donovan, Sandra E. Ward, Mi-Kyung Song, Susan M. Heidrich, Sigridur Gunnarsdottir, and Christopher M. Phillips. 2007. An update on the representational approach to patient education. *Journal of nursing scholarship : an official publication of Sigma Theta Tau International Honor Society of Nursing*, 39 3:259–65.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56:52–64.

Felix Greaves, Indra Joshi, Mark Campbell, Samantha Roberts, Neelam Patel, and John Powell. 2018. What is an appropriate level of evidence for a digital health intervention? *The Lancet*, 392(10165):2665–2667.

C. Howes, Matthew Purver, Rosemarie McCabe, Patrick G. T. Healey, and Mary Lavelle. 2012. Predicting adherence to treatment for schizophrenia from dialogue transcripts. In *SIGDIAL Conference*.

Linda Irvine, Ambrose J Melson, Brian Williams, Falko F Sniehotta, Andrew McKenzie, Claire Jones, and Iain K Crombie. 2017. Real time monitoring of engagement with a text message intervention to reduce binge drinking among men living in socially disadvantaged areas of scotland. *International journal of behavioral medicine*, 24(5):713–721.

William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *CoRR*, abs/1901.08746.

Luoqiu Li, Xiang Chen, Hongbin Ye, Zhen Bi, Shumin Deng, Ningyu Zhang, and Huajun Chen. 2021. On robustness and bias analysis of bert-based relation extraction. In *CCKS*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhi Liu, Wenjing Zhang, Hercy N. H. Cheng, Jianwen Sun, and Sannyuya Liu. 2018. Investigating relationship between discourse behavioral patterns and academic achievements of students in spoc discussion forum. *Int. J. Distance Educ. Technol.*, 16:37–50.

Rosemarie McCabe, Patrick G. T. Healey, Stefan Priebe, Mary Lavelle, David Dodwell, Richard Laugharne, Amelia Snell, and Stephen Bremner. 2013. Shared understanding in psychiatrist-patient communication: association with treatment adherence in schizophrenia. *Patient education and counseling*, 93 1:73–9.

Matthew A. Napierala. 2014. What is the bonferroni correction ?

Gareth Phillips. 2016. Nurses are best placed to ensure the ethical application of dnrs. *Nursing Standard (2014+)*, 30(37):31.

Sravana Reddy, Mariya Lazarova, Yongze Yu, and Rosie Jones. 2021. Modeling language usage and listener engagement in podcasts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 632–643, Online. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Liam Rourke, Terry Anderson, D. Randy Garrison, and Walter Archer. 1999. Assessing social presence in asynchronous text-based computer conferencing.

Shumpei Sano, Nobuhiro Kaji, and Manabu Sassano. 2016. Prediction of prospective user engagement with intelligent assistants. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1203–1212, Berlin, Germany. Association for Computational Linguistics.

Sarah Shugars and Nick Beauchamp. 2019. Why keep arguing? predicting engagement in political conversations online. *SAGE Open*, 9.

Karen Swan, D. Randy Garrison, and Jennifer C. Richardson. 2009. A constructivist approach to online learning: The community of inquiry framework.

Xiangyu Wang, Andrew C. High, Xi Wang, and Kang Zhao. 2020. Predicting users’ continued engagement in online health communities from the quantity and quality of received support. *Journal of the Association for Information Science and Technology*, 72:710 – 722.

Miaomiao Wen, Diyi Yang, and Carolyn Penstein Rosé. 2014. Linguistic reflections of student engagement in massive open online courses. In *ICWSM*.

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5021–5031, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Lucy Yardley, Bonnie J Spring, Heleen Riper, Leanne G Morrison, David H Crane, Kristina Curtis, Gina C Merchant, Felix Naughton, and Ann Blandford. 2016. Understanding and promoting effective engagement with digital behavior change interventions. *American journal of preventive medicine*, 51(5):833–842.

Chen Yu, Paul M. Aoki, and Allison Woodruff. 2004. Detecting user engagement in everyday conversations. *ArXiv*, cs.SD/0410027.

A Appendix: Mapping of Engagement Categories

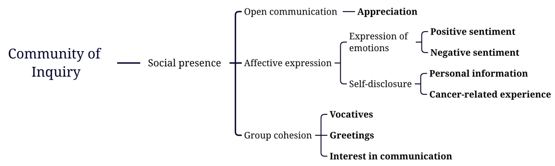


Figure 5: Mapping Community of Inquiry framework (Swan et al., 2009) to socio-affective engagement categories.

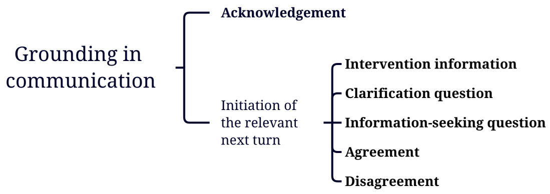

Grounding in communication
Acknowledgement
Initiation of the relevant next turn
Intervention information
Clarification question
Information-seeking question
Agreement
Disagreement

Figure 6: Mapping grounding in communication framework (Clark and Brennan, 1991) to cognitive engagement categories.

B Appendix: Description of Intervention Process and Coding Scheme

Description of Intervention Process

Step	Intervention element	Description
1		The nurse initiates the message board by introducing herself and explaining how the intervention works on the message boards. The nurse then invited the patient to talk about herself and her experience with ovarian cancer and symptoms in general.
2	Representational assessment	The nurse proceeds to ask a series of protocolized questions in the first few messages. This is for both the patient and the nurse to get a full understanding of the patient's experience with each symptom. As the patient thinks about the questions and writes responses, she may find that she sees some things more clearly or perhaps differently. A few examples of the protocolized questions include: • What does your fatigue feel like and how severe is it? • How has this symptom affected your life? Are you unable to do anything because of it? • How does your fatigue affect you emotionally?
3	Identifying and exploring Gaps, Errors, and Confusions	Throughout interactions with the participant, the nurse keeps attending to evidence of any confusion, concerns, or misconceptions about symptom management that the patient states or suggests during her message board posts. Any of these constitutes a barrier to effective symptom management for participants.
4	Creating conditions for conceptual change	If any concerns/ misconceptions/gaps are identified, the nurse will (1) discuss the relationship between identified concerns/confusion/misconceptions and consequences of poor symptoms that the patient talked about during representational assessment and (2) provide information to address/counter patient's concerns, gaps, or misconceptions.
5	Introducing replacement information	The nurse will then guide the patient to read the relevant clinical practice guidelines and provide relevant information to the patient to choose and use different strategies that fit well into their life and needs for managing symptoms.
6	Summary	The nurse will summarize information and the ways she believes it can benefit the patient. For example, increase comfort and less interference with life.
7	Goal Setting and Planning	The nurse will encourage the patient to work together on developing symptom goals and specific strategies to reach those goals.
8	Goal and Strategy Review	After 2 weeks, the nurse will ask protocolized questions to work with the patient to evaluate strategies, reinforce success, and outline modifications as needed. The nurse will also discuss any barriers that occur in the implementation of strategies and work with the patient to identify new or different strategies that could be integrated into her life. A few examples of the protocolized questions include: • Were you able to use the above strategies? • If not, what things prevented you from doing so? • If yes, how well did the strategies (name/list strategies) work in helping you reach your goal?
The back-and-forth discussion introduces the patient to a problem-solving approach to symptom management that could be applied to any symptom, not just the three they were working on.		

Coding Scheme

To ensure the accuracy and reliability of the engagement coding for the message board posts, two independent raters conducted a thorough review of various engagement models from different disciplines, including education, linguistics, cognitive science, behavior science, and computer-human interactions. Using the Taguette tagging software, they iteratively coded each post, extracting semantic information and identifying discourse behaviors based on linguistic cues. The engagement categories were then summarized based on the context of the nurse's queries or suggestions and the participant responses. To ensure inter-rater reliability, the raters discussed coding for every 100 posts coded with the linguist, the developer and principal investigator of the clinical trial. The coding scheme was developed and updated iteratively, with clear guidelines and examples for each code. If engagement was present, a value of 1 was assigned, and if absent, a value of 0 was assigned. The resulting engagement categories can aid in identifying patients who require additional attention or support in online communication or telehealth settings and can also help practitioners evaluate the effectiveness of intervention strategies.

Socio-affective engagement	**Description**	**Example**
Positive sentiment	1. Participant explicitly expresses positive emotions or feedback on the intervention design, symptom management progress (e.g., the effectiveness of strategies). 2. *Very emotional greetings and goodbyes can fall under this code provided there is a clearly expressed strong emotional connection (intensity), for example "Oh. It's so nice to meet you!"	1. **Participant**: "Other than that, I feel that all the strategies are good ones and THEY WORK!!" 2. **Participant**: "I feel relieved now that I know many other people have similar problems. That makes me feel normal ☺" 3. **Participant**: "I will look forward to your advice on the issues!"
Negative sentiment	Participant explicitly expresses negative emotions or feedback on the intervention design or content, the symptom management progress (e.g., the effectiveness of strategies).	**Participant**: "We have been having computer problems. After writing part of a draft yesterday and half of one today, we contacted Jim from your group for tech support. Both messages to you were spontaneously deleted!!!!!!!!!!"
Appreciation	Participant explicitly appreciate and recognize the nurse's contributions and support.	**Participant**: "I really appreciate the time you take to discuss on my symptoms. You have much insight and your recommendations are very helpful."
Cancer-related experience	1. Participant shows her vulnerability (e.g., fear, guilt, traumatic cancer experience). 2. Participant opens to the nurse about her feelings and beliefs of cancer trajectory and symptom management. 3. Participant discusses with nurse about the cancer-related issues that she finds hard to discuss with others.	1. **Participant**: "I have caught myself cycling through emotional fatigue times where I'm completely frustrated with myself and life and want to give up and then have some good news or an outing and everything is back to fine again." 2. **Participant**: "Honestly I feel like a prisoner to it. I have to plan for it weekly, bug the nurse on my HCT to get me appointments, I feel it constantly everyday, I can't wear some of my clothes anymore or I fear I can't get an appointment and hit a weekend and have to go 3 extra days is misery." 3. **Participant**: "It's really hard for me to tell my family about my pain. My husband doesn't understand how much pain I have."
Personal information	Participant shares personal information with the nurse (e.g., family, friends, pets, health care team, religion)	**Participant**: "Yea. My daughter and son-in-law came to visit. We had a great Thanksgiving dinner. Thanks for asking. "
Interest in communication	1. Participant invites nurse to ask for more information or clarification if needed. 2. Participant explicitly expresses interest in further communication in general.	1. **Participant**: "Please let me know if you need any other information." 2. **Participant**: "I will update you on the treatment next Wednesday. "
Vocatives	Participant addresses the study name by the name.	"Names of study nurses"

Greetings	1. Participant starts posts with a greeting (e.g., Dear, Hello) 2. Participant signs off the email (e.g., yours, …, sincerely, …., “thanks, …”) 3. Participant shows benevolence, kindness, and social respect to the nurse.	1. **Participant**: “Good morning.” 2. **Participant**: “Take care, [name]” 3. **Participant**: “May He bless your service in helping women to survive with less pain and life to gain.”
Cognitive engagement	**Description**	**Example**
Intervention information	1. Participant answers nurse questions seeking information to complete intervention activities. 2. Participant answers nurse questions seeking clarification about something participant said. The goal is to reduce misunderstanding. 3. Participant reports completing intervention tasks assigned/suggested by the nurse. 4. Participant provides information or updates unprompted (not asked by the nurse). For example, proposing new strategies, technical issues of the website. 5. Participant updates the nurse about her cancer treatment that impact her symptom management.	1. **Nurse**: “How does it affect you emotionally?” **Participant**: “I probably get cranky when I am tired, but my dear husband doesn’t complain.”. 2. **Nurse**: “What do you mean you don’t like to take pills?” (Participant mentioned that she doesn’t like taking pills for symptoms). **Participant**: “I have to take many prescribed medications. I don’t want to take more that I have to”. 3. **Participant**: “I did read the guide you suggested. I guess I can start walking around the block”. 4. **Participant**: “I just went to see the dentist and walk with Jim after that.” (Context: seeing a dentist is a strategy participant chose to manage her mouth sore). 5. **Participant**: “I was completely drained after chemo last Wed. I didn’t have energy to do anything.”
Clarification question	Participant asks nurse questions seeking clarification about something the nurse said. The goal is to reduce misunderstanding.	1. **Participant**: “my nutritionist told me I should eat more fruits and vegetable. But the guide says I should not have raw fruits/vegetables if I have diarrhea. Can I still have salads sometimes?” 2. **Participant**: “taking pain meds won’t cause addiction, even matter if I take it every day?”
Information-seeking question	1. Participant proposes a question to seek for answers. The specific indicators contain “?”, “who”, “what”, “where”, “when”, “why”, “how”, “how-much/many” or other key question symbols. 2. Participant seeks for advice/suggestions. 4. Participant asks the nurse for a favor.	1. **Participant**: “I checked out the diet website. Do you know what’s the regimen?” 2. **Participant**: “I look forward to your answer to my questions about your list.” 3. **Participant**: “It would be wonderful if you could add the goals and strategies as I'm getting ready to go out of town and am really busy preparing for that and trying to pace myself.”
Acknowledgement	Participant acknowledges/affirms the nurse suggestions, ideas, summary, questions.	**Participant**: “I did read your e-mail last night before going to bed. I didn’t answer because I was tired, but I did keep thinking about your questions.”
Agreement	1. Participant explicitly shares the same opinion with the nurse on the topics (e.g., suggestions on symptom management strategies, barriers to good symptom management, reading resources), including restating or paraphrasing or summarizing what the nurse has said. The specific indicators contain “I agree”, “I think so”, etc. 2. Participant agrees to the plans or do what the nurse suggests doing, including stating that she will try what the nurse has said (being open to different/new experience). The specific indicators contain “I will…”, “Okay”, etc. 3. Participant approves what the nurse does or suggests.	1. **Participant**: “In just the few days of doing yoga I have become aware that my posture is not good. I think maybe that is affecting the way I walk as you suggested when you asked how I walk.” 2. **Nurse**: “I was thinking what might make the most sense is to do your 2-week follow-up for Sleep before moving on to headaches. How’s that sound to you? I will have the Sleep Follow-Up posted up there (on web page) under the Sleep topic heading.” 3. **Participant**: “okay, I will take a look at the Sleep Follow-up. ” 4. **Participant**: “The plan looks fine. I don’t have anything to add.”

Disagreement	1. Participant explicitly disagrees with nurse on a certain topic. The specific indicators contain “I disagree”, “I don’t think so”, etc. 2. Participant disagrees to do what the nurse suggests doing, including explaining why they don’t want to or can’t try what the nurse suggested. The specific indicators contain “I don’t…”, “doesn’t work”, etc.	1. **Participant**: “What do you mean? I don’t think this is helping me. I was hoping that the program can help me solve the problem.” 2. **Nurse**: “here are other similar activities that you might also like, such as Tai Chi. Many community centers offer both of these. Look for classes for people over 50... they often start out a little slower and easier. Don’t push too hard or fast when you start out.” **Participant**: “I tried Tai Chi two years ago at the senior center. My sister does it and recommended it. I did not get the feeling of renewal from it that she did. Actually, I did not get much from it at all. I went twice a week for two months.”
Initiative taking	1. Participant anticipates the later steps of the intervention. For example, when the nurse asks questions about patient’s symptom representation (element 1 Representation Assessment), the patient both describes the symptoms AND proceeds to describe the strategies she wants to try during the intervention (element 6 Goal setting and Planning). 2. Participant changes the course of the intervention by taking initiative in changing symptoms to work with the nurse before the nurse’s prompts. 3. Participant offers suggestions to future study design	1. **Nurse**: “Can you tell me about your hot flashes? You can describe a typical day.” **Participant**: “ It usually happen at night. I have to get up several times because it is too hot. In looking at the Symptom Care Guides, there are four strategies I would like to work on: 1. Keep a symptom diary. (I have printed a template. I will keep it on the end table where I sit to watch TV, crochet, read, etc.)2. Yoga. I will look on the TV an see if I can find a yoga class.3. Vitamin E. I will call and get approval from my HCT.4. Paced Respirations (abdominal breathing).” **Nurse**: “Wow, you’re jumping right into this process! That’s wonderful. We are in the process of updating our [Resource Library] for Hot Flashes and will add some Yoga web sites, soon. Before we go too much further, though, I do have a few more questions for you..... Don’t I always ☺ This is so we can both get the full picture. Can you tell me a little more about how your hot flashes feel?” 2. **Participant**: “I think we can start with the sleep disturbance first. I am working on my eating now by myself, so that will take a while to see the affects of that, so let's move on to the next one, shall we.” 3. **Participant**: “Oh just remembered an idea: I wonder if a message box would be helpful to add to the questionnaires? The patient could type out any comments/explanations in answering the questions. Just wondering in the event of any chemo and/or life changes that occurred during this research time that impacted the current questionnaire.”

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation section at the end of the paper

☑ A2. Did you discuss any potential risks of your work?
Section 6 and Ethical Considerations section at the end of the paper

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Sections 2-5

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We are presenting a new dataset and we are adhering to the rules of our institutions ethics review board.

□ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Researchers who wish to gain access to our dataset, will have to complete appropriate ethics review protocols

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 6

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 1

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3.2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
5.2

C ☑ Did you run computational experiments?

Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We do not do hyper-parameter search.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5.3

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 3

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We describe key points of annotation protocol. We don't provide full text because of space limits.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
3.1

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
3.1

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
3.1

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
3.2