

DiagESC: Dialogue Synthesis for Integrating Depression Diagnosis into Emotional Support Conversation

Seungyeon Seo¹, Gary Geunbae Lee^{1,2}

¹Graduate School of Artificial Intelligence, POSTECH, Republic of Korea

²Department of Computer Science and Engineering, POSTECH, Republic of Korea
{ssy319, gblee}@postech.ac.kr

Abstract

Dialogue systems for mental health care aim to provide appropriate support to individuals experiencing mental distress. While extensive research has been conducted to deliver adequate emotional support, existing studies cannot identify individuals who require professional medical intervention and cannot offer suitable guidance. We introduce the Diagnostic Emotional Support Conversation task for an advanced mental health management system. We develop the DESC dataset¹ to assess depression symptoms while maintaining user experience by utilizing task-specific utterance generation prompts and a strict filtering algorithm. Evaluations by professional psychological counselors indicate that DESC has a superior ability to diagnose depression than existing data. Additionally, conversational quality evaluation reveals that DESC maintains fluent, consistent, and coherent dialogues.

1 Introduction

As interest in preventing and treating mental illnesses like depression, anxiety disorders, and panic disorders grows, dialogue system studies on mental health care are gaining attention. Several studies have shown that chatbots can effectively manage the mental health of individuals, particularly in frontline settings, before seeking professional help (Denecke et al., 2021; Lim et al., 2022). These chatbots provide emotional empathy and assist in finding stability for those facing emotional, mental, and psychological distress. Mental health care also involves the early detection of illnesses. Although delayed treatment aggravates symptoms and requires more complex treatment, it is challenging for individuals to self-diagnose (Epstein et al., 2010). Therefore, detecting diseases during conversation is an important factor, and we focus on depression, a representative mental illness.

¹Our dataset DESC is accessible at github.com/seungyeonseo/DiagESC.

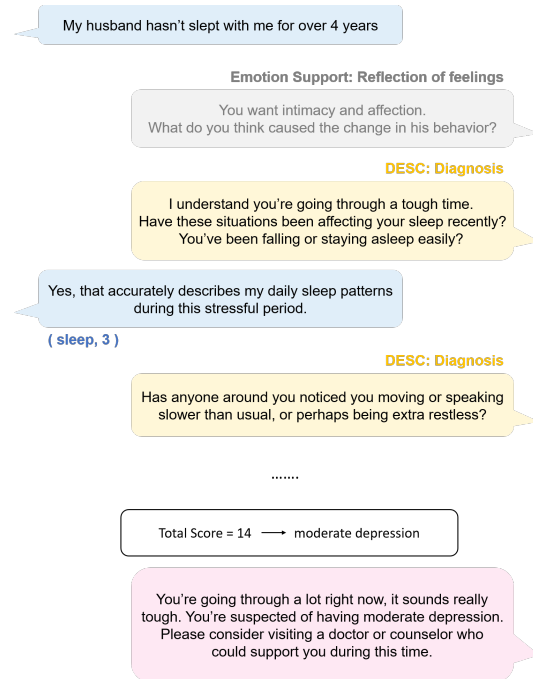


Figure 1: Part of an example conversation sample from DESC. The left is the seeker's, and the right is the supporter's utterance. We initiate a diagnostic conversation by inserting a diagnostic question (yellow) instead of a specific supporting emotion utterance (gray). At the end of the conversation, appropriate assistance (pink) is provided based on the severity of the depression.

Our research aims for an advanced conversation system to facilitate comprehensive mental health management. This system should provide extensive emotional support to individuals while simultaneously employing diagnostic questions to detect early signs of depression proactively.

To achieve this goal, we define a novel task, Diagnostic Emotional Support Conversation (DiagESC), based on Emotional Support Conversation (ESC) (Liu et al., 2021). ESC aims to support by helping reduce the seeker (user)'s mental stress. We also synthesize and release the dataset DESC for this task. We synthesize utterances to

ask questions about depression symptoms while maintaining a positive user experience. Additionally, task-specific strict filtering algorithms ensure data quality. Figure 1 shows part of the dialogue sample in DESC. It includes questions about depression symptoms and labels regarding symptom frequency, enabling assessing the severity of depression. Appropriate advice based on the severity of depression helps the individual receive help. Professional psychological counselors validate the diagnostic ability and conversational quality of DESC.

2 Related Work

2.1 Supportive Dialogue System

Recognizing emotions is essential for dialogue systems to respond appropriately to the user’s feelings. Emotion-tagged dialogue datasets such as DailyDialog (Li et al., 2017), Emotionlines (Hsu et al., 2018), and EmoContext (Chatterjee et al., 2019) have enhanced the conversation quality by enabling emotion-based response generation (Wei et al., 2019; Zandie and Mahoor, 2020; Ide and Kawahara, 2021). In particular, Lubis et al. (2019) shows that integrating emotional context in response generation can elicit positive emotions in users. The dataset EmpatheticDialogues (Rashkin et al., 2019) contains rich emotion labels and high-quality utterances that understand and empathize with users’ emotions, encouraging research on generating empathic responses (Ghosal et al., 2020; Majumder et al., 2020; Li et al., 2022).

To enable more effective emotional support, the ESC task (Liu et al., 2021) is defined by employing response strategies based on the Helping Skills Theory (Hill, 2009). ESC uses more sophisticated strategies, such as questioning and providing suggestions beyond empathy, to improve the users’ emotions and encourage them to overcome difficulties. Cheng et al. (2023) introduced persona generation into ESC and proposed Persona-Augmented Emotional Support (PAL), enabling the creation of responses tailored to an individual’s situation and characteristics.

However, understanding the situation and providing advice cannot fully help someone suffering from depression. Individuals with depression require professional counseling and medication rather than temporary emotional support. Research on supportive dialogue systems, such as ESC, demonstrates user encouragement capabil-

ities but cannot adequately address the needs of those with depression.

2.2 Depression Detection in Conversation

As with all diseases, early detection of depression is very important for efficient treatment. However, due to difficulties such as a lack of knowledge about the symptoms of depression, it is hard for patients to recognize that they are suffering from depression themselves (Epstein et al., 2010).

Against this background, depression detection research is being conducted to help with early treatment. Ringeval et al. (2019) proposed a classification task for whether a user has depression based on the audio and video features. They released the dataset DAIC-WOZ, which contains video recordings of clinical interviews designed to diagnose psychological disorders. The user participated in the conversation after completing a depression self-diagnosis questionnaire. DAIC-WOZ has significantly advanced depression detection research, contributing to numerous breakthroughs in the field (He and Cao, 2018; Haque et al., 2018; Low et al., 2020). We utilize DAIC-WOZ as a benchmark due to the absence of text-based depression diagnosis conversation datasets.

2.3 Dialogue Data Synthesis

Several methodologies have been proposed for the generation and augmentation of dialogue data to address the constraints associated with the time-intensive and costly data construction process (Lewis et al., 2017; Hou et al., 2018; Tang et al., 2019). With the emergence of the Large Language Model (LLM), the field of data synthesis has transitioned into a novel paradigm (Ding et al., 2024).

Kim et al. (2023); Bao et al. (2023) introduced a novel synthetic dialogue dataset derived from external sources. The data was refined using filtering techniques designed to ensure criteria such as commonsense knowledge, dialogue flow, and coherence. A method for synthesizing Dialogue State Tracking (DST) labeled conversation data from dialogue schemas and templates has shown comparable performance to human-annotated datasets in few-shot DST (Kulkarni et al., 2024). Kim et al. (2024); Li et al. (2024) generated the conversational dataset using task-specific prompting technology, and the test set is certificated through humans.

Building on these advancements, we synthesize the DESC dataset by leveraging the fluent utterance generation capabilities of LLM, thereby con-

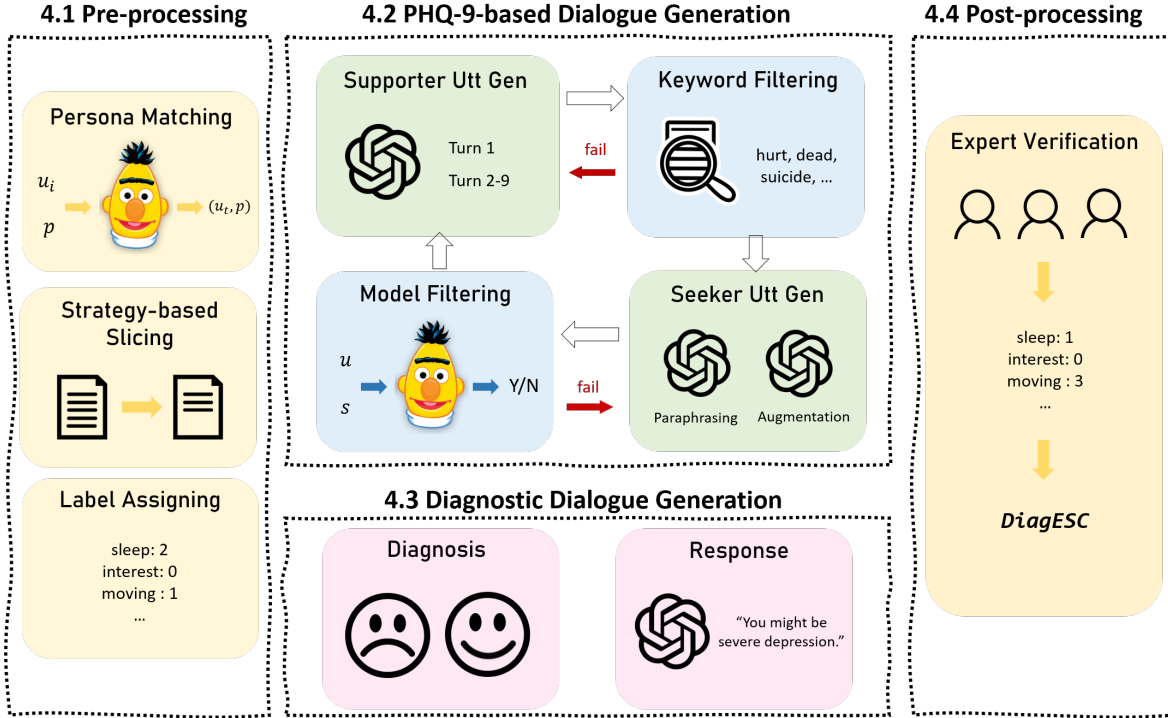


Figure 2: The overview of the DESC synthesis process.

tributing a novel resource for text-based depression detection research.

3 Problem Formulation

DiagESC consists of three sub-tasks—two modes response generation, persona generation, and diagnosis generation. The dual modes of response generation encompass emotional support and diagnostic responses. Persona generation is extracting characteristics based on the seeker’s previous utterance, as suggested by PAL (Cheng et al., 2023). Utilizing the previous persona as input can increase the user experience by facilitating reflection on an individual’s characteristics and serving as a form of memory when the dialogue history cannot include all utterances. Diagnosis generation, introduced for DiagESC, involves generating symptom and corresponding score pairs.

Equation 1 refers to DiagESC \mathcal{F} at turn t that generates response $r_t = m \oplus s_t$, persona p_t , and diagnosis pair $d_t = \{\text{Symptom, Score}\}$ given the persona sentences $P_t = \{p_1, p_2, \dots, p_t\}$ and dialogue history $C_t = \{u_1, s_1, u_2, s_2, \dots, u_t\}$. $m \in \{\text{emotional support, diagnostic}\}$ denotes the mode for response. u_t and s_t represent t^{th} seeker (user) and supporter (system) utterance respectively.

$$\mathcal{F}(P_t, C_t) = (r_t, p_t, d_t) \quad (1)$$

x The emotional support response uses existing ESC strategies, and the diagnostic response includes depression symptom questions and the notification diagnosis result.

4 Methodology

We synthesize the dataset for DiagESC named DESC through a four-step process, as illustrated in Figure 2. Initially, the source data undergoes pre-processing to align with the task requirements (Section 4.1). PHQ-9-based Dialogue Generation is for generating conversations that ask and answer about symptoms of depression (Section 4.2). Each task-specific prompt is based on the Patient Health Questionnaire-9 (PHQ-9) symptom item. The process includes filtering for reliability. The severity of depression is then calculated based on the answers obtained by the seeker. Section 4.3 is to inform the seeker of appropriate advice. Finally, to enhance reliability, expert verification is conducted on the validation and test datasets (Section 4.4).

4.1 Pre-processing

Persona Matching We utilize comprehensive annotations and high-quality supporting dialogue from the PESConv dataset of PAL (Cheng et al., 2023), containing persona sentences extracted from previous dialogue history. However, the persona

Symptom Item	Description
Interest	Little interest or pleasure in doing things
Depressed	Feeling down, depressed, or hopeless
Sleep	Trouble falling or staying asleep, or sleeping too much
Tired	Feeling tired or having little energy
Appetite	Poor appetite or overeating
Failure	Feeling you are a failure or have let yourself or your family down
Concentrating	Trouble concentrating on things, such as reading the newspaper or watching television
Moving	Moving or speaking so slowly that other people could have noticed. Or the opposite, being so fidgety or restless that you have been moving around a lot more than usual
Hurting	Thoughts that you would be better off dead, or of hurting yourself

Table 1: The symptoms and descriptions of PHQ-9

sentences do not align exactly with the seeker’s utterance for each turn.

We employ the BERT² (Kenton and Toutanova, 2019) model to obtain the embeddings for all persona sentences p and seeker utterances u . Then, we compute cosine similarities between each persona sentence embedding and every utterance embedding. Each persona sentence has a higher cosine similarity to the utterance from which it is derived than to other utterances. We reassign all of the persona sentences using the following equation.

$$\hat{t}_i = \operatorname{argmax}_{t \in \{1, \dots, T\}} \frac{E(p_i) \cdot E(u_t)}{|E(p_i)| |E(u_t)|} \quad (2)$$

where \hat{t}_i represents the matched turn number for the i -th persona sentence p_i . u_t denotes the utterance at the t -th turn and the function $E(\cdot)$ refers to compute BERT embedding. Equation 2 ensures to align each persona sentence with its derived utterance.

Strategy-based Slicing Determining the appropriate moment to begin diagnostic questions is challenging. It is crucial to consider that abruptly interrupting the flow of conversation may negatively impact the user’s emotional state. Fortunately, ESC dataset has rich annotations, tagging each utterance with its corresponding strategy. The most suitable time for presenting diagnostic questions has been empirically determined to use specific strategies, namely *Restatement or Paraphrasing*, *Reflection of Feeling*, *Self-disclosure*, and *Affirmation and Reassurance*. Figure 1 is an example of using a diagnostic question (yellow) instead of a reflection response (gray). This rule enables a smooth and contextually appropriate transition into diagnostic questioning.

Furthermore, we only utilize truncated data when at least two persona sentences have been

gathered to ensure that diagnostic questions are only posed after comprehensively understanding the seeker’s persona. This criterion helps that sufficient contextual background is considered before diagnostic engagement.

Label Assigning To achieve an even distribution of the final severity level within the generated conversational data, the PHQ-9 labels are pre-assigned. The next step uses predefined labels to generate utterances.

4.2 PHQ-9-based Dialogue Generation

We utilize the Patient Health Questionnaire-9 (PHQ-9), a widely used medical tool for self-assessment of depression (Kroenke et al., 2001), as the basis for the depression diagnostic questions. PHQ-9 aims to quantify the frequency of nine depressive symptoms listed in Table 1 on a scale ranging from 0 to 3, with the options *Not at all*, *Several days*, *More than half the days*, and *Nearly every day*. The aggregated score of all items is used to diagnose depression and assess its severity, categorized as Minimal (0-4), Mild (5-9), Moderate (10-14), Moderately severe (15-19), and Severe (20-27).

4.2.1 Supporter Utterance Generation

We develop two types of prompts to generate supporter utterances for the first and subsequent turns. In the initial turn, it is essential to formulate questions with caution to maintain a positive user experience. For the subsequent turns, which involve further diagnostic questioning, it becomes essential to comprehend and empathize with the seeker’s responses to the preceding questions.

The both prompts involve the three-step Chain-of-Thought technique (Wei et al., 2022; Kim et al., 2024). In the first-turn supporter utterance generation prompt, the steps consist of *Selection*, *Plan-*

²<https://huggingface.co/google-bert/bert-base-uncased>

Prompt Content

You are an emotional supporter. You have to ask about the frequency of depression symptoms without compromising the emotions of the user suspected of having depression.

Depression Symptoms You should ask how ‘often’ a symptom has occurred over the past two weeks. Use one of the following symptoms. Be careful not to distort the medical meaning. (*symptoms*)

Task Description The task proceeds in three stages: Selection, Planning, and Response Generation. The first step, Selection, is to select which of the given persona sentences and dialog history to use in the response generation and what symptoms to ask about. Information that can improve the user experience must be extracted. The second step, Planning, is planning how to use the selected information. You must explain how you will use the information you have selected and why you have selected that information. The final step, Response Generation, uses the selected information to naturally ask the user about depression symptoms. Consistency with persona and history must not be broken. Questions must be asked carefully so that the user does not feel that the question is sudden. Be careful not to ask hard as if you were being interrogated. The generated response must be no more than 25 words.

Example (*examples*)

Table 2: The prompt used to generate the initial supporter utterance of inquiring about PHQ-9 symptoms.

ning, and Response Generation. Table 2 provides detailed instructions for these steps. *Selection* and *Planning* focus on the seeker’s persona and the dialogue history.

Because analyzing the previous answer about the symptom is more critical for subsequent turns, we replaced the *Selection* with the *Analysis*. The steps help to analyze the seeker’s response and generate a response accordingly. Detailed instructions can be found in the Appendix B.

4.2.2 Seeker Utterance Generation

The construction of the seeker’s utterance necessitates including one of the four designated responses from PHQ-9 (Not at all, Several days, More than half the days, Nearly every day). We enhance and utilize the template-based utterance generation method (Kulkarni et al., 2024). After rephrasing

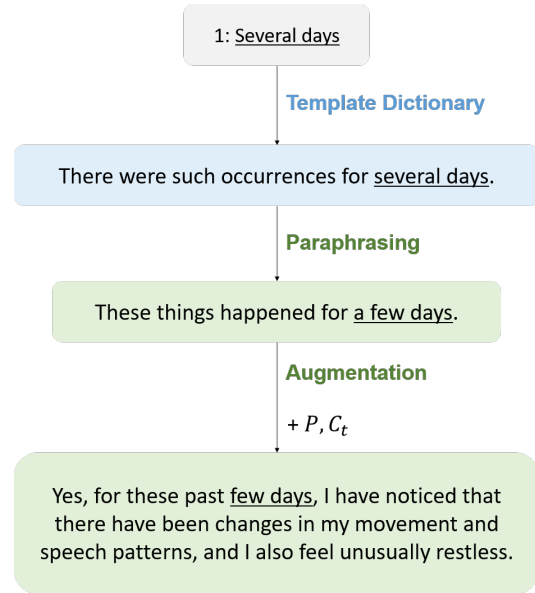


Figure 3: The overview of seeker utterance generation

Score	Type	Content
0	PHQ Temp.	Not at all There is not at all much of that.
1	PHQ Temp.	Several days There were such occurrences for several days.
2	PHQ Temp.	More than half the days That happened for more than half the days.
3	PHQ Temp.	Nearly every day It was like that almost every day.

Table 3: Templates by score used for the Paraphrasing stage in generating seeker utterances.

the templates, we augment the utterances to match the user’s characteristics better, personalizing them to align more closely with specific traits and preferences, providing a more tailored and engaging conversational experience.

Figure 3 shows the process by which the template is transformed into the final utterance via the stages of Paraphrasing and Augmentation using LLM. Initially, we establish a foundational template that directly correlates with a response option from the PHQ-9, as illustrated in Table 3.

Paraphrasing model M_{Para} aims to diversify responses while preserving symptom frequency information. Augmentation model M_{Aug} represents the process of generating the final utterance u_t to align with the seeker’s specific situation and persona, particularly considering the supporter’s last utterance and the overall conversational flow. As can be seen in Equation 3, M_{Para} requires a template of the corresponding score s as input and M_{Aug} demands a paraphrased output, the persona sentences P_t , and

the conversational history C_t .

$$u_t = \mathbf{M}_{\text{Aug}}(P_t, C_t, \mathbf{M}_{\text{Para}}(\text{template}_s)) \quad (3)$$

4.2.3 Filtering

If the LLM misrepresents the medical interpretation of PHQ-9, it will adversely affect the quality of the diagnosis function. We implement strict filtering algorithms to prevent the hallucination of LLM and ensure the reliability of diagnostic conversations.

Keyword Filtering To maintain the integrity of symptom representation in LLM-generated supporter utterances, we pre-define specific keywords associated with each symptom. If each utterance includes no pre-defined keywords, we discard it and regenerate a new one. This process continues until the generated utterance appropriately incorporates the necessary keywords, ensuring accurate and consistent symptom representation in the dialogue.

Model Filtering It is essential to preserve symptom frequency information in the seeker’s utterances. Therefore, we train the classification model M_c , using 256 manually verified utterances. If the predicted label from model M_c differs from the template label, or if the confidence is below a threshold t , the utterance is regenerated.

4.3 Diagnostic Dialogue Generation

The final goal of DiagESC’s diagnostic ability is to estimate the seeker’s mental health status and provide appropriate assistance. The severity level of depression is determined by summing the scores of all nine items obtained from the Diagnosis task and then generating an appropriate response as shown in the final utterance in Figure 1. To achieve this goal, we design the prompt in Table 4 to generate an utterance based on the seeker’s persona and a diagnosed depression severity level. To enhance the naturalness of the conversation, we incorporate a predefined turn expressing gratitude for the honest response before the diagnostic response.

4.4 Post-processing

Despite applying strict task-specific filtering protocols, the potential for inaccuracies remains. To ensure the reliability of the PHQ-9 labels, Expert Filtering is conducted on the validation and test sets of DESC. Three psychologists, who are native English speakers or bilingual and have over four years of professional experience³, assessed scores

³We hired psychologists through <https://www.upwork.com>

Prompt Content

You are emotional support. You have provided counseling to the user about the concerns and even completed questions about depression symptoms. Generate an utterance that concludes the counseling by referring to the depression diagnosis results and the user’s persona. If the severity of depression is high, you should be recommended to see a hospital or counselor. Please generate the utterance friendly conversational style and generated utterance must be no more than 30 words.

Example (*examples*)

Table 4: The prompt used to generate utterance for notifying diagnosis result.

for each symptom. The seeker utterances are then re-labeled to the mode value of the three scores.

5 Experiments

5.1 Diagnostic Ability Evaluation

The DAIC-WOZ dataset, used as a baseline, comprises clinical dialogues in video and audio features with PHQ-8 (Kroenke et al., 2009) labels. The PHQ-8 is a modified version of the PHQ-9, excluding the items related to suicide, and performs just as well as the PHQ-9 in diagnosing depression. Although the modality is different from ours, due to the absence of conversation data explicitly labeled for depression, we use transcripts of DAIC-WOZ. We randomly sample four dialogues from DAIC-WOZ for each severity level. Employing the same methodology with expert labeling described in Section 4.4, three psychological counselors evaluate scores for the PHQ-8 items.

The Quadratic Weighted Kappa (QWK) score is a metric that evaluates the agreement between two predictions, offering advantages by acknowledging both exact and partial alignment in assessments (Cohen, 1968). The QWK score ranges from -1 to 1. A score closer to -1 indicates that the predictions are nearly opposite. A score near 0 reflects randomness, implying no consistent agreement between the predictions. If a score approaches 1, the predictions are almost identical and have a high level of agreement. As it is suitable for medical fields where symptoms can be interpreted slightly differently depending on the individual doctor (Yoshida et al., 2015; Nirthika et al., 2020; Chivinge et al., 2022), the QWK score is widely used in disease diagnosis. Therefore, we adopt the QWK score as

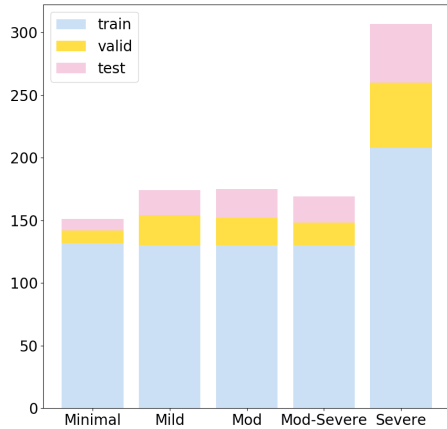


Figure 4: Distribution of depression severity labels in DESC. Minimal (0-4), Mild (5-9), Moderate (10-14), Moderately severe (15-19), and Severe (20-27).

the principal metric to evaluate diagnostic ability.

5.2 Conversational Quality Evaluation

To evaluate the quality of the conversational data, we sample 10 dialogues for each severity level from the DESC validation and test sets. We then requested the same evaluators with diagnostic ability evaluation to rate the following three items on a scale from 1 (Very Poor) to 5 (Excellent).

- **Fluency** evaluates the grammatical correctness, naturalness, and smoothness of the dialog.
- **Consistency** assesses how well the dialogue maintains a consistent user persona throughout the interaction. This involves the user’s interests and personality traits.
- **Coherence** measures how contextually appropriate the responses are, considering the previous dialogue turns and the overall context of the conversation.

5.3 Automatic Validation in Baseline

We establish the baseline models by tuning a small LLM utilizing Low-Rank Adaptation (LoRA) (Hu et al., 2022) to evaluate the operation of DESC in each model. We train the LoRA adapters on DESC for the three sub-tasks: response, persona, and diagnosis generation. Additionally, performance in a multi-adapter setting is measured to evaluate the multi-tasking capabilities of the system. In a single-task setting, the inputs for the next turn are the true labels of other tasks. However, the inferred results

from the previous turn are used as input for all tasks in a multi-task setting. We systematically provide symptom item sequences to ensure consistency and effectiveness in the diagnostic process.

5.4 Implementation Details

We use GPT-4 as LLM to generate utterances of DESC. For Model Filtering of seeker utterance, we adopt RoBERTa⁴ (Liu et al., 2019) and train the classification model M_c for 5 epochs. The labels predicted by the fine-tuned model are utilized for filtering purposes, with the threshold $t = 0.7$.

Llama2⁵ (Touvron et al., 2023) is used as a baseline model, and the adapters are trained with the train set for 5 epochs on 4 NVIDIA A6000 GPUs, and the final model with the lowest validation loss was selected. We employ AdamW with a learning rate of $5e-5$ and a linear scheduler.

6 Results and Analysis

6.1 Basic Statistics of DESC

The DESC comprises 976 dialogues, including 730 train, 126 validation, and 120 test samples. Each dialogue has an average of 42 turns, with the maximum number of turns per dialogue being 111 and the minimum being 24. Figure 4 illustrates the distribution of dialogue samples across five levels of depression severity. The Severe level has more samples than the other levels because it covers a wider range of scores.

6.2 Diagnosis Ability

According to the results presented in Table 5, DESC achieves a notably high average QWK score of 0.70 compared to baseline. In contrast, DAIC-WOZ obtains low scores, with a certain item showing negative values. The result indicates a substantial challenge in predicting the frequency of a seeker’s depression symptoms solely from conversational history with an agent in the dataset. This comparison may be considered unfair because the DAIC-WOZ does not include questions about all the symptoms of the PHQ-8.

The most important result is the final depression diagnostic capability of each dataset, as presented in Table 6. The depression severity is classified into five levels—minimal, mild, moderate, moderately severe, and severe—based on the cumulative scores of the assessed items. Scores exceeding 10 points,

⁴FacebookAI/roberta-base

⁵meta-llama/Llama-2-7b-chat-hf

Dataset	Interest	Depressed	Sleep	Tired	Appetite	Failure	Concentrating	Moving	Hurting	Avg
DAIC-WOZ	0.16	0.39	0.18	0.15	0.03	0.03	-0.13	0.04	-	0.11
DESC	0.44	0.69	0.64	0.70	0.80	0.80	0.64	0.78	0.81	0.70

Table 5: Average QWK Scores of each dataset against expert annotations for each symptom in PHQ-8 and PHQ-9.

Dataset	Level Acc	Depression			
		Acc	Precision	Recall	F1
DAIC-WOZ	0.45	0.70	0.50	1.00	0.67
DESC	0.71	0.89	1.00	0.86	0.92

Table 6: Accuracy of predicting depression severity level and accuracy, precision, recall, and f1 score of estimating depression diagnosis.

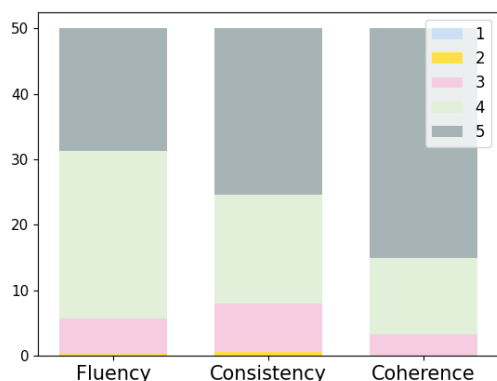


Figure 5: Distribution of evaluated scores for DESC's fluency, consistency, and coherence.

classified as moderate or higher, are considered depression.

The Level Acc indicates the accuracy of predicting severity levels, with DESC showing 0.26 higher performance than DAIC-WOZ. In depression diagnosis, the superior accuracy and F1 score of DESC, compared to the baseline, demonstrate its robustness and effectiveness. The results suggest that our PHQ-9-based data generation process ensures reliable diagnostic capabilities.

6.3 Conversation Quality

Figure 5 shows the distribution of obtained scores in the human evaluation performed to evaluate conversation quality. The average scores are 4.25 for fluency, 4.33 for consistency, and 4.63 for coherence. Most samples received scores of 3 or higher across all items, indicating that the DESC is consistent and comprises high-quality conversations without contextual awkwardness. Notably, the high coherence score suggests that the diagnostic questions generated through strategy-based slicing and first-turn supporter prompt configuration help the seeker perceive them as natural and non-abrupt.

	Response		Persona	Diagnosis
	Mode	BLEU	BLEU	Acc
Single	0.83	31.08	34.03	0.78
Multi	0.83	30.78	34.65	0.77

Table 7: The performance on baseline models for single-task and multi-task settings.

6.4 Automatic Validation

Table 7 shows the baseline performance of DESC. Mode indicates the prediction accuracy of the response mode, divided into emotional support and diagnosis. It shows equal performance of 0.83 in both single-task and multi-task settings. Generating response and persona sentences achieve high BLEU scores, all above 30. Diagnosis accuracy measures the prediction of each symptom and its corresponding score. Across all metrics, single-task and multi-task settings demonstrate similar performance.

7 Conclusion

This work proposes the DiagESC task for a comprehensive mental health care dialogue system that goes beyond the limitations of supportive dialogue systems that do not detect mental risk. DiagESC contributes to emotional support and early detection of depression, an important part of mental health. We have released the novel dataset DESC by synthesizing diagnostic conversations based on a depression self-diagnosis questionnaire with emotional support data. Task-specific prompts and strict filtering protocols facilitate questions about depression symptoms while ensuring continued user engagement. Evaluation by a psychological counseling expert proves that DESC has superior diagnostic performance and conversational quality. We hope that DiagESC will contribute significantly to developing more effective and supportive dialogue systems in mental health care. Moreover, the release of the DESC dataset provides a valuable resource for the research community, encouraging further advancements and innovations in this critical area.

Limitation

The research aims to identify depression signs during conversations with the user, subsequently notifying them of potential risks. It is imperative to note that the diagnostic outcomes derived from the proposed dataset and model are intended solely for guidance. An accurate and definitive diagnosis should be ascertained through consultation with a medical professional.

Acknowledgements

This work was supported by Smart Health-Care Program(www.kipot.or.kr) funded by the Korean National Police Agency(KNPA, Korea) [Project Name: Development of an Intelligent Big Data Integrated Platform for Police Officers' Personalized Healthcare / Project Number: 220222M01] This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2024-RS-2024-00437866) supervised by the IITP(Institute for Information Communications Technology Planning Evaluation)

References

- Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. [A synthetic data generation framework for grounded dialogues](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10866–10882, Toronto, Canada. Association for Computational Linguistics.
- Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 39–48.
- Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023. Pal: Persona-augmented emotional support conversation generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 535–554.
- Lincoln Chivinge, Leslie Kudzai Nyandoro, and Kudakwashe Zvarevashe. 2022. [Quadratic weighted kappa score exploration in diabetic retinopathy severity classification using efficientnet](#). In *2022 1st Zimbabwe Conference of Information and Communication Technologies (ZCICT)*, pages 1–9.
- Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.
- Kerstin Denecke, Alaa Abd-Alrazaq, and Mowafa Househ. 2021. Artificial intelligence for chatbots in mental health: opportunities and challenges. *Multiple perspectives on artificial intelligence in health-care: Opportunities and challenges*, pages 115–128.
- Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. *arXiv preprint arXiv:2403.02990*.
- Ronald M Epstein, Paul R Duberstein, Mitchell D Feldman, Aaron B Rochlen, Robert A Bell, Richard L Kravitz, Camille Cipri, Jennifer D Becker, Patricia M Bamonti, and Debora A Paterniti. 2010. “i didn’t know what was wrong:” how people with undiagnosed depression recognize, name and explain their distress. *Journal of general internal medicine*, 25:954–961.
- Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481.
- Albert Haque, Michelle Guo, Adam S Miner, and Li Fei-Fei. 2018. Measuring depression symptom severity from spoken language and 3d facial expressions. *arXiv preprint arXiv:1811.08592*.
- Lang He and Cui Cao. 2018. Automated depression analysis using convolutional neural networks from speech. *Journal of biomedical informatics*, 83:103–111.
- Clara E. Hill. 2009. *Helping Skills: Facilitating, Exploration, Insight, and Action*, 3 edition. American Psychological Association.
- Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245.
- Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. Emotion-lines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Tatsuya Ide and Daisuke Kawahara. 2021. Multi-task learning of generation and classification for emotion-aware dialogue response generation. In *Proceedings*

- of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 119–125.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, et al. 2023. Soda: Million-scale dialogue distillation with social commonsense contextualization. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Yumin Kim, Heejae Suh, Mingi Kim, Dongyeon Won, and Hwanhee Lee. 2024. Kocosa: Korean context-aware sarcasm detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9890–9904.
- Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.
- Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173.
- Atharva Kulkarni, Bo-Hsiang Tseng, Joel Moniz, Dhivya Piraviperumal, Hong Yu, and Shruti Bhargava. 2024. SynthDST: Synthetic data is all you need for few-shot dialog state tracking. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1988–2001, St. Julian’s, Malta. Association for Computational Linguistics.
- Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453.
- Jinpeng Li, Zekai Zhang, Quan Tu, Xin Cheng, Dongyan Zhao, and Rui Yan. 2024. Stylechat: Learning recitation-augmented memory in llms for stylized dialogue generation. *arXiv preprint arXiv:2403.11439*.
- Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10993–11001.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.
- Shi Min Lim, Chyi Wey Claudine Shiau, Ling Jie Cheng, and Ying Lau. 2022. Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: A systematic review and meta-regression. *Behavior Therapy*, 53(2):334–347.
- Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.
- 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*.
- Daniel M Low, Kate H Bentley, and Satrajit S Ghosh. 2020. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope investigative otolaryngology*, 5(1):96–116.
- Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2019. Positive emotion elicitation in chat-based dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4):866–877.
- Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979.
- Rajendran Nirthika, Siyamalan Manivannan, and Amirthalingam Ramanan. 2020. Loss functions for optimizing kappa as the evaluation measure for classifying diabetic retinopathy and prostate cancer images. In *2020 IEEE 15th international conference on industrial and information systems (ICIIS)*, pages 144–149. IEEE.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.
- Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. 2019. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, pages 3–12.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, and Yuchong Hu. 2019. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1401–1410.

Kanako Yoshida, Rebecca J Barr, Sandro Galea-Soler, Richard M Aspden, David M Reid, and Jennifer S Gregory. 2015. Reproducibility and diagnostic accuracy of kellgren-lawrence grading for osteoarthritis using radiographs and dual-energy x-ray absorptiometry images. *Journal of Clinical Densitometry*, 18(2):239–244.

Rohola Zandie and Mohammad H Mahoor. 2020. Empransfo: A multi-head transformer architecture for creating empathetic dialog systems. In *The thirty-third international flairs conference*.

A Dialogue Example

Figure 6 illustrates the utterance configuration of the DESC data sample. During emotional support conversation, the supporter initiates the PHQ-9-based dialogue to inquire about symptoms. After identifying all symptoms, the supporter provides the user with appropriate advice, such as a recommendation to visit a hospital based on the diagnostic results. To enhance the natural flow of the conversation, we insert the predefined turns between symptom inquiries and result notifications. We select these turns from 23 supporter and 16 seeker utterance candidates.

B Detailed Prompt Instructions

Table 8 is the prompt for supporter utterance generation. In contrast to the utterance generation of turn 1, this stage focuses on analyzing the prior

Emotional Support	
Supporter	Hi, how are you? What would you like to discuss today?
Seeker	hi, I am not feeling too well.
Supporter	How long have you felt this way?
Seeker	I have been feeling this low for close to a month now
...	
Supporter	You had an affair because you were dissatisfied with your marriage. Why do you think that was?
Seeker	my husband hasn't slept with me for over 4 years
PHQ-9-based Dialogue	
Supporter	I understand you're going through a tough time. Have these situations been affecting your sleep recently? You've been falling or staying asleep easily?
Seeker	Yes, that accurately describes my daily sleep patterns during this stressful period.
Supporter	Has anyone around you noticed you moving or speaking slower than usual, or perhaps being extra restless?
Seeker	Indeed, those around me have noticed my unusual behavior for quite a few days now.
Supporter	Is it common for you to think about no longer being around or inflicting harm upon yourself?
Seeker	In reality, I don't regularly have thoughts of that nature.
...	
Supporter	I understand. Have you been having trouble focusing on things like reading or watching TV often?
Seeker	Yes, over the past few days, I've indeed found it increasingly difficult to focus on things, like reading or watching television.
Pre-defined Turn	
Supporter	Thank you for being so direct in your answer.
Seeker	Thank you for your help; it means a lot in these tough times.
Notification	
Supporter	You're going through a lot right now, it sounds really tough. You're suspected of having moderate depression. Please consider visiting a doctor or counselor who could support you during this time.

Figure 6: Part of an example conversation sample of DESC.

seeker response. Table 9 and 10 are the prompts for seeker utterance generation.

C Keyword Filtering

The keyword filtering process ensures that the PHQ-9 maintains its medical meaning. Table 11 shows detailed keywords for each symptom item. The generated utterance must contain at least one of the keywords.

D Distribution of DESC

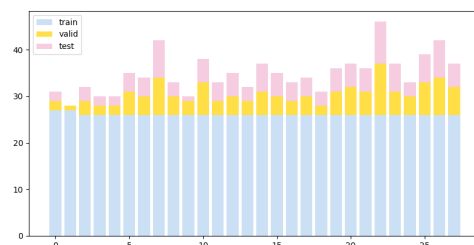


Figure 7: Distribution of aggregated score in DESC.

Each dialogue sample has score labels for the PHQ-9 items. The sum of all nine scores determines the severity of depression. Figure 7 is the distribution of aggregated scores.

Prompt Content

(same with turn 1)

Depression Symptoms You should ask how ‘often’ a symptom has occurred over the past two weeks. Symptoms are given and the frequency of the symptoms should be naturally asked of the user. The meaning of a given symptom should never be changed.

Task Description The task proceeds in three stages: Analysis, Planning, and Response Generation. The first step, Analysis, is to determine the user’s status through the user’s previous responses. The second step, Planning, is planning how to use the status information to support the user’s emotions and ask about the frequency of a different given symptom. The final step, Response Generation, is to ask the user about the symptom according to plan. Question must be asked carefully so that the user does not feel that the question is sudden. Be ... (same with turn 1)

Example (examples)

Table 8: The prompt used to generate the subsequent turn supporter utterance of inquiring about PHQ-9 symptoms.

E Human Evaluation

We conduct two types of human evaluations. In the diagnostic ability evaluation, three evaluators, all psychologists, read the conversations and scored each item of the PHQ-9. We use the mode value as the final label to reduce individual subjectivity. If the evaluators’ scores differ, we use the mean score as the final label.

The conversation quality evaluation assesses performance based on fluency, consistency, and coherence scores. The evaluators read the dialogues and assign a score between 1 and 5 for each criterion, following the descriptions provided for each item.

Fluency evaluates the grammatical correctness, naturalness, and smoothness of the dialogue.

- **Very poor** numerous errors, hard to understand.
- **Poor** lacks smoothness but can be followed with some effort.
- **Normal** a natural rhythm to the conversation despite occasional awkwardness.
- **Good** natural, easy to follow.

Prompt Content

Rephrase the sentence while retaining the original meaning. The sentences are conversation with counseling diagnosis chatbot system and the user. In particular, do not change the frequency-related meaning of the user’s words. Use synonyms or related words to express the sentences with the same meaning. Use conversational language and paraphrase the following sentences. Generate a crisp and to the point single sentence from the given sentences using conversational language.

Table 9: The Paraphrasing prompt used in seeker utterance generation

Prompt Content

Please augment the user utterance to fit the dialog history while maintaining its original meaning. The sentence is the user’s utterance in a conversation between the counseling diagnosis chatbot system and the user. In particular, do not change the frequency-related meaning of user’s words. Please augment and modify the given user utterance to match the system’s last words and the flow of the conversation, especially user’s situation and persona.

Table 10: The Augmentation prompt used in seeker utterance generation

- **Excellent** natural, grammatically sound and logically structured.

Consistency assesses how well the dialogue maintains a consistent user persona throughout the interaction. This involves the user’s interests, and personality traits.

- **Very poor** frequent contradictory utterances; feels like by a completely different person.
- **Poor** regular contradictory utterances; a general sense of the original character remains perceivable but persona seem inconsistent.
- **Normal** some contradictory utterances; occasional contradictory utterances that mildly affect the coherence of the user persona but do not substantially alter the overall character impression.
- **Good** few errors; it’s pretty much the same person speaking.

Item	Keywords
Interest	interest, pleasure, enjoy
Depressed	depressed, hopeless, down
Sleep	sleep
Tired	tired, energy
Appetite	appetite, eat
Failure	fail, down
Concentrating	concentrate, concentrating, TV, television, read
Moving	move, moving, slow, restless, figety
Hurting	hurt, dead, suicide, self, harm height

Table 11: The keyword list of each symptom item.

- **Excellent** no or negligible errors; user personas are fully maintained.

Coherency measures how contextually appropriate the responses are, considering the previous dialogue turns and the overall context of the conversation.

- **Very poor** conversations frequently veer off-topic without a clear reason.
- **Poor** related to main topic but may include irrelevant details.
- **Normal** related topic with occasional lapses in focus or clarity.
- **Good** related topic and minor deviations are quickly corrected.
- **Excellent** every response directly contributes to a coherent, logical, and engaging.