Fine-tuning Large Language Models for Automated Diagnostic Screening Summaries

Anonymous ACL submission

Abstract

Improving mental health support in developing countries is a pressing need. One potential solution is the development of scalable, automated systems to conduct diagnostic screenings, which could help alleviate the burden on mental health professionals. In this work, we evaluate several state-of-the-art Large Language Models (LLMs), with and without fine-tuning, on our custom dataset for generating concise summaries from mental state examinations. We rigorously evaluate four different models for summary generation using established ROUGE metrics and input from human evaluators. The results highlight that our top-performing fine-tuned model outperforms existing models, achieving ROUGE-1 and ROUGE-L values of 0.810 and 0.764, respectively. Furthermore, we assessed the fine-tuned model's generalizability on a publicly available D4 dataset, and the outcomes were promising, indicating its potential applicability beyond our custom dataset.

1 Introduction

Mental health disorders are prevalent worldwide. A recent study shows that one in every eight people suffers from some mental health disorder (WHO, 2022). Usually, mental health disorders are diagnosed in clinical settings with Mental State Examination (MSE). An MSE is a structured assessment of the behavioral and cognitive functioning of an individual suffering from a mental health disorder (Martin, 1990; Voss et al., 2019). It aids in comprehending psychological functioning across multiple domains, including mood, thoughts, perception, cognition, etc. Mental health professionals (i.e., psychiatrists and psychologists) utilize MSEs at different treatment stages (prior, during, or after) to grasp the onset of mental health disorders, assess the effectiveness of therapy sessions, and evaluate the progress of treatment.

In developing countries, mental health support is limited, with only a few mental health professionals available for a large number of patients (Majumdar, 2022; Rojas et al., 2019; Saraceno et al., 2007). Resident (junior) doctors, supervised by senior doctors, are commonly employed to manage the demand. The primary responsibility of junior doctors is to conduct initial patient assessments through MSEs and create concise summaries of issues and symptoms for senior doctors. Reviewing these summaries reduces evaluation time for senior doctors, allowing them more time to focus on treatment planning. Unfortunately, junior doctors are typically only accessible in selected hospitals for various reasons. This lack of availability of junior doctors increases the workload for doctors and often leads to longer wait times for patients.

Developing an automated system for initial assessment and summary generation would be pivotal in simulating an AI-driven junior doctor. The system would conduct MSEs and generate concise summaries of the MSE for the attending senior doctor. Implementing such a scalable, automated system would alleviate the demand for junior doctors and lessen the burden on senior doctors. Moreover, such a system would be immensely beneficial in regions with limited mental health professionals, especially in low and middle-income countries.

The automated system for conducting and summarizing MSEs consists of two main parts: (i) a user interface for gathering user responses to MSE questions and (ii) an AI module for summarizing those responses. This study focuses on the latter by evaluating various Large Language Models (LLMs) to determine their effectiveness in generating concise summaries from MSEs. Summarizing accurately and concisely using pre-trained LLMs is challenging due to a lack of relevant mental health conversation datasets and the significant shift in content from non-mental to mental health topics. To tackle these challenges, we first developed a 12-item descriptive MSE and collected data by conducting MSEs with 300 participants. Next, using our dataset, we assessed the performance of four well-known pre-trained LLMs with and without fine-tuning for summarizing MSEs. Our comprehensive evaluation, based on metrics such as ROUGE scores and human judgment, indicates that fine-tuning pre-trained LLMs, even with limited training data, improves the generation of accurate and coherent summaries. Notably, the best fine-tuned models outperform existing baseline LLM models, achieving ROUGE-1 and ROUGE-L scores of 0.810 and 0.764, respectively. Furthermore, we demonstrate the generalizability of the best fine-tuned model by evaluating it on a publicly available dataset using human annotators. The contributions of this work include:

- We evaluate the state-of-the-art LLMs with and without tuning for summary generation.
- We evaluate the generalizability of the best model on two different datasets with two different evaluation metrics (ROUGE & human).
- We collect a real-world dataset for training and testing the LLMs.

2 Related Works

There are two primary methods for summarizing text: extractive and abstractive. Extractive summarization involves directly copying important text from the original text (Kupiec et al., 1995; Filatova and Hatzivassiloglou, 2004). On the other hand, abstractive summarization involves using new words and phrases to create a summary, even if they weren't present in the original text (Rush et al., 2015; Chopra et al., 2016). Both extractive and abstractive summarization methods have their own strengths and weaknesses. However, this paper focuses on abstractive summarization due to its ability to generate more human-friendly summaries.

2.1 Pre-trained model

Large language models (LLMs) like GPT (Radford et al., 2018), BART (Lewis et al., 2020), T5 (Raffel et al., 2020) have gained attention for understanding instructions, generating human-like responses, and adapting to new Natural Language Processing (NLP) tasks such as text generation and summarization. Abstractive summarization methods show promise in utilizing these LLM models for flexible summarization tasks. However, their application in medicine, particularly in the psychological domain, requires exploration to address inaccuracies without domain-specific knowledge.

2.2 Summarization

In abstractive summarization, the advent of sequence-to-sequence (seq-to-seq) models marked a significant advancement (Nenkova and McKeown, 2012). This progress was further enhanced with the introduction of a neural network model incorporating attention mechanisms and a generation algorithm (Rush et al., 2015). Based on this foundation, conditional RNN architecture and a convolutional attention-based encoder significantly improved sentence summarization (Chopra et al., 2016).

Concurrently, alternative architectures emerged to refine seq-to-seq models. A transformer-based encoder-decoder architecture (Enarvi et al., 2020) inspired from (Vaswani et al., 2017) yielded highly accurate summaries. Additionally, a pointing mechanism (See et al., 2017) for word copying from the source document further diversified the summarization process. Recently introduced PEGASUS (Zhang et al., 2020), an innovative summarization framework founded upon a transformer-based encoder-decoder architecture, represents the latest frontier in this evolving landscape.

2.3 Dialogue summarization

Models like BART (Lewis et al., 2020) and GPT-3 (Radford et al., 2018), with their vast number of parameters, demonstrate exceptional performance across various general-purpose tasks. However, their training primarily relies on knowledge-based resources such as books, web documents, and academic papers. Nonetheless, they often require additional domain-specific conversation/dialogue data to understand dialogues better. The lack of publicly available appropriate data sets creates a challenge for generating abstractive summaries. To overcome this challenge, Samsung research team (Gliwa et al., 2019) made their dataset publicly available. Furthermore, (Zhong et al., 2022) introduced a pre-training framework for understanding and summarizing long dialogues.

Similarly, (Yun et al., 2023) enhanced routine functions for customer service representatives by employing a fine-tuning method for dialogue summarization. However, medical dialogues present unique challenges due to the inclusion of critical information such as medical history, the context of the doctor, and the severity of patient responses, necessitating specialized approaches beyond those employed in typical dialogue processing.

2.4 Medical dialogue summarization

Recent advancements in automatic medical dialogue summarization have propelled the field forward significantly. Notably, both LSTM and transformer models have demonstrated the capability to generate concise single-sentence summaries from doctor-patient conversations (Krishna et al., 2021). Furthermore, pre-trained transformer models have been leveraged to summarize such conversations from transcripts directly (Zhang et al., 2021; Michalopoulos et al., 2022; Enarvi et al., 2020).

In addition, the hierarchical encoder-tagger model has emerged as a promising approach, producing summaries by identifying and extracting meaningful utterances, mainly focusing on problem statements and treatment recommendations (Song et al., 2020). However, it is important to note that these models are typically trained on brief, general physician-patient conversations. In contrast, conversations in the psychological domain tend to be longer, with more detailed patient responses. Understanding the nuances of behavior and thinking patterns becomes crucial for accurate disease identification in such contexts. (Yao et al., 2022) addressed this challenge by applying a fine-tuned pre-trained language model to generate abstractive summaries of psychiatrist-patient conversations using a Chinese dataset. However, as of our current understanding, there is a lack of comparable abstractive summarizations of psychiatrist-patient conversations available in English text. This highlights a potential area for further research and development in medical dialogue summarization.

3 Methodology

Figure 1 provides a high level overview of the methodology. Following is a detailed description of the methodology sub-components.

3.1 MSE questionnaire design

We identified the absence of a standardized MSE questionnaire and reviewed existing options online. We aimed to create a preliminary version tailored to students, encompassing key components like socialness, mood, attention, memory, frustration tolerance, and social support. This process yielded an 18-question questionnaire. Subsequently, we

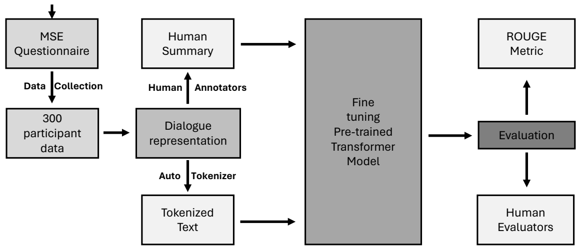


Figure 1: Methodology flowchart

sought the expertise of clinical psychiatrists to refine the questionnaire further. Their valuable insights were instrumental in vetting the relevance and wording of the questions, resulting in a finalized version of the MSE comprising 12 questions. The final MSE questionnaire is provided in the Appendix A.1.

3.2 Data collection

We conducted a data collection study at our institute. Initially, we obtained the study approval from our institute's ethics committee, and subsequently, participants were recruited from the institute. Institute students, regardless of their mental health status, were invited to fill out a Google Form indicating their preferred date and time for the study participation. Subsequently, participants received a separate email from a research assistant (RA) requesting their attendance at the specified venue on their chosen date. Upon arrival, participants were provided with a participant information sheet and an informed consent form. Upon signing the informed consent form, they completed the designed MSE. Participants were not briefed on the MSE questions in advance. On average, participants spent approximately 20 minutes completing the MSE questionnaire. A total of 300 participants, consisting of 202 males and 98 females, participated in the study. The demographic characteristics of the participants are presented in Table 1. Data collection from these 300 participants spanned 80 days.

Each participant's data was assigned a unique anonymized identifier, ensuring that it cannot be traced back to the participant, given the nature of psychological conversations involving personal experiences. After completing the study, participants were provided snacks to acknowledge and accommodate their valuable time.

	#	Age (μ, σ)	Home Residence (urban, rural)
All	300	(21.62, 3.70)	(212, 88)
Male	202	(21.34, 3.69)	(138, 64)
Female	98	(22.19, 3.64)	(74, 24)

Table 1: Participants Demographics

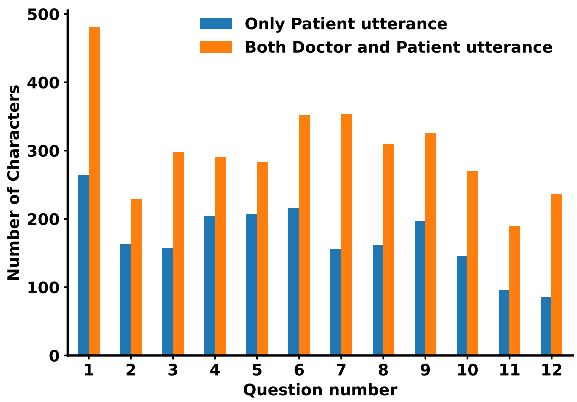


Figure 2: Average lengths of patient (i.e., participant) and doctor utterances for each question, aggregated across all 300 patient-doctor conversations. Note that the length of doctor utterances remains constant for each questionnaire, as the questions were predefined.

3.3 Dialogue representation

We developed a Python script to transform participants' MSE questionnaire responses into simulated doctor-patient conversations to replicate real-world conversations. This process generated 300 doctor-patient conversation sessions, with 3600 (=12 responses x 300 participants) utterances from participants and an equal number from doctors, totaling 7200 utterances. An anonymized excerpt of such conversation for one participant is presented in Table A.2 in the appendix. Figure 2 shows the average length of utterances for each of the 12 questions. The average length of the dialogue conversation with and without the questionnaire is 3662 and 2054 characters, respectively.

All participants were proficient in English and submitted their responses in English. Despite some participants making spelling errors, these errors were preserved in the dataset to mirror real-world situations where users might misspell words.

3.4 Reference human summaries

To facilitate the training of supervised deep-learning models for summarizing doctor-patient conversations, reference summaries are required. Such summaries should encompass essential information, context, and insights of collected MSEs. Due to the lack of standardized guidelines for creating such summaries and the subjective nature of human-generated summaries influenced by personal perception, we developed a structured summary template approach similar to (Can et al., 2023).

Furthermore, given the structured nature of the MSE questions, the template was well-suited for summarization purposes. The summary template underwent thorough scrutiny through a rigorous review process involving feedback from three independent reviewers (i.e., graduate researchers). Subsequent revisions were made based on their input, ensuring the summary effectively captured key information while maintaining conciseness, clarity, and correctness. After multiple iterations, the final version of the summary template was approved for use by a psychiatrist, leveraging their domain-specific knowledge. The template utilized for creating summaries for each participant can be found in Appendix A.1.

3.5 Training

Our collected dataset contains both doctor-patient conversations and human-generated (reference) summaries. Therefore, we opted for supervised learning approaches. Given the efficiency and widespread use of transformer-based models and considering the limited number of related training datasets available, we chose to fine-tune following existing well-known publicly available pre-trained models.

1. **BART base model** (Lewis et al., 2020): BART is a transformer encoder-decoder model featuring a bidirectional encoder and an autoregressive decoder. It is pre-trained on the English language using two main techniques, i.e., corrupting text with an arbitrary noising function and learning a model to reconstruct the original text. It demonstrates superior efficacy when fine-tuned for text-generation tasks such as summarization and translation (Huang et al., 2020). In our evaluation, we utilized the BART base model from Hugging Face[1], comprising 139 million parameters.

2. **BART-large-CNN model**: BART-large-CNN is a fine-tuned model of BART-base with the CNN Daily Mail dataset (Hermann et al.,

[1] https://huggingface.co/facebook/bart-base

2015). It is tailored for text summarization tasks, leveraging a dataset containing a vast collection of articles from CNN Daily Mail, each accompanied by its summary. Given that the primary objective of BART-large-CNN is text summarization, we used the BART-large-CNN model from Hugging Face[2], which has 406 million parameters.

3. **T5 large**: The T5 Large for Medical Text Summarization model is a tailored version of the T5 transformer model (Raffel et al., 2020), fine-tuned to excel in summarizing medical text. It is fine-tuned on the dataset, encompassing a variety of medical documents, clinical studies, and healthcare research materials supplemented by human-generated summaries. The diverse dataset on medical texts aids the model's capability in accurately and concisely summarizing medical information. Given that the model is designed for medical text summarization tasks, we found it appropriate for fine-tuning on our psychological conversations. We used the model from Hugging Face[3], which encompasses 60.5 million parameters.

4. **BART-large-xsum-samsum model** (Gliwa et al., 2019): The BART-large-xsum-samsum model is trained on the Samsum corpus dataset, comprising 16,369 conversations along with their respective summaries. Given that this model is explicitly trained on conversation data, it was deemed suitable for our task. We utilized the pre-trained model from Hugging Face[4], which contains 406 million parameters. While using this model, we hypothesized that since it has been trained on a dialogue conversation dataset, it would outperform other models while summarizing our collected dataset.

4 Experiments

We adopted the well-known ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric (Lin, 2004) as the primary evaluation criterion, in line with recent literature (Krishna et al., 2021; Zhang et al., 2021; Michalopoulos et al., 2022) on automated summarization. The metric compares the automated summary generated from the trained model with the reference summary. While the metric excels at syntactical textual similarities, it fails to capture semantic similarities between two summaries. However, to address the limitation of the metric in terms of semantic analysis, we have done qualitative analysis using ratings from clinical and non-clinical annotators to check the semantic similarities between reference and model-generated summaries.

The dataset comprising 300 conversations was divided into 200 for training, 50 for validation, and 50 for testing. The Appendix A.2 lists the hyperparameter settings utilized during model training.

4.1 ROUGE evaluation

The average ROUGE values (ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-L-SUM) for the 50 generated test set summaries with different models without and with fine-tuning are shown in Table 2. The values were computed by comparing the model generated and human reference summaries.

The table illustrates that the BART-large-xsum-samsum model, without fine-tuning, attains the highest ROUGE values across all mentioned metrics (ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-L-SUM). This underscores that using the pre-trained weights of these models can yield the highest ROUGE-1 value of 0.290 on our conversation dataset. The superior performance of this model can be attributed to its training on the conversation dataset, distinguishing it from other models. Following fine-tuning with our dataset, the BART-large-CNN model achieves the highest ROUGE-1 and ROUGE-L values of 0.810 and 0.764, respectively.

We conducted experiments varying the number of epochs for each model to compare their relative performance, as depicted in the Figure A.1 in the Appendix. This figure showcases the models' adaptability across different ROUGE metrics as epochs increase. Notably, within just five epochs, the ROUGE-1 score of the BART-large-xsum-samsum model surged from 0.290 to 0.736. Similarly, both the BART-base and BART-large-CNN models demonstrated significant improvements in all ROUGE values within the same timeframe. However, the T5 large model failed to exhibit notable adaptation within five epochs.

[2] https://huggingface.co/facebook/bart-large-cnn

[3] https://huggingface.co/Falconsai/medical_summarization

[4] https://huggingface.co/lidiya/bart-large-xsum-samsum

	Models	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-L-SUM
Without tuning	BART-base	0.212	0.048	0.106	0.106
	BART-large-CNN	0.186	0.025	0.125	0.125
	T5 large	0.228	0.046	0.140	0.140
	BART-large-xsum-samsum	**0.290**	**0.107**	**0.216**	**0.216**
With tuning	BART-base	0.798	0.671	0.755	0.755
	BART-large-CNN	**0.810**	**0.690**	**0.764**	**0.765**
	T5 large	0.727	0.570	0.662	0.662
	BART-large-xsum-samsum	0.795	0.660	0.749	0.749

Table 2: ROUGE values of the model generated summaries without and with fine-tuning. Reported values represent the average values over the test set summaries of 50 doctor-patient conversations.

With an increase in epochs to 10, we observed improvements in ROUGE values for all three BART-based models, except for the T5-large model. However, beyond 25 epochs, the performance of the BART-based models began to saturate. Remarkably, the T5-large model started to adapt at 25 epochs, with its ROUGE score rising from 0.266 observed at 10 epochs to 0.702. Nevertheless, similar to the BART-based models, it also reached saturation after 50 epochs.

To gain insight into the model-generated summaries, we conducted experiments with all models across different numbers of epochs (epochs = 5, 10, 25, 50, 100). After analyzing the output summaries generated by these models, we randomly selected one of the participant's summaries for further analysis. We found that the pre-trained weights of these models tended to produce incomplete summaries, although they were able to capture smaller contexts of the conversation, as shown in Table A.3 in the Appendix.

Notably, the pre-trained BART-large-xsum-samsum model exhibited greater appropriateness and performance compared to the others. Within just five epochs, both the BART-large-xsum-samsum and BART-large-CNN models demonstrated an ability to capture the broader context of the conversation, albeit missing some important key information. The BART-large-CNN model surpassed all other models within 10 epochs, achieving the highest ROUGE values of 0.81, 0.69, 0.766, and 0.766 in terms of ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-L-SUM, respectively.

Conclusion: Based on the ROUGE results, the fine-tuned BART-large-CNN model emerged as the best-performing model. Consequently, we utilized the summary generated by the BART-large-CNN model for further assessments in the subsequent evaluation sections. The BART-large-CNN model checkpoint at 25^{th} epoch along with a sample conversation from our dataset can be found at this[5] anonymous Google Drive link.

4.2 Human evaluation

To assess the model's semantic effectiveness, we conducted a qualitative analysis with the assistance of two clinicians (psychiatrists) and three non-clinicians (graduate lab researchers not involved in the study). This analysis was performed on ten doctor-patient conversations randomly selected from a test set of 50 participants. The ten participants were chosen randomly using Python's random module, with a fixed random seed 42. We provided the selected conversations and the human-generated and best model-generated (i.e., BART-large-CNN) summaries to the reviewers. Importantly, the reviewers were unaware of whether the summaries were generated by the model or by humans during the evaluation process. Reviewers were instructed to assess summaries on a 5-point scale (1 to 5) based on the following defined evaluation parameters. The evaluation parameters were determined following a brief literature survey (Zhang et al., 2021; Yao et al., 2022):

- *Completeness*: Does the summary cover all relevant aspects of the conversation?
- *Relevance to Medical context*: Does the summary cover sufficient medical information related to mental health disorders as per the conversation?
- *Fluency*: Is the summary well structured, free from awkward phrases, and grammatically correct?
- *Clarity*: Is the summary clear and easy to understand?
- *Missingness*: Does the summary miss any key information?

[5] https://drive.google.com/drive/folders/17afLEsOHOdaRwxbTxn5tMxXG0ePXJtqh

- *Hallucination*: Does the summary contain any extra information that was not presented by the patient?
- *Contradiction*: Does the summary contradict with the information provided by the patient?
- *Repetition*: Does the summary consist of repeated information/sentences?

4.2.1 Qualitative findings

Table 3 presents the average scores of different evaluation parameters for all ten reference and best-model (BART-large-CNN) generated summaries assigned by human evaluators. The discrepancies in quality between the model-generated and human-referenced summaries are minimal in terms of *fluency*, *clarity*, and *repetition*, indicating that the model-generated summaries are as readable as those crafted by humans.

However, the model-generated summaries slightly lack in *completeness* and *relevance* compared to the human-generated summaries. Additionally, the generated summaries contain more *missing* information than the human summary. This difference can be attributed to the summary template used to create a summary for the conversation, which the model did not fully internalize.

Moreover, there was a higher degree of *contradiction* in the generated summary, although a certain level of contradiction was also observed in the human-generated summary. This discrepancy may arise from either the scenario where the domain expert who created the summary may have missed some interpretation or from expert reviewers having different perceptions. Surprisingly, the model did not exhibit hallucination, which is a major problem in large language models. Furthermore, Table A.4 in the Appendix displays the evaluation scores separately from clinicians and non-clinicians. We observed a slight disparity between clinicians and non-clinicians, indicating that clinicians may require a summary with detailed psychological information.

4.2.2 Inter-rater agreement

Inter-rater agreement, also known as inter-rater reliability or inter-observer agreement, refers to the level of agreement between two or more raters or observers when assessing the same data. It is often measured using statistical measures such as Cohen's kappa (ranges 0 to 1) and Pearson correlation coefficients (ranges -1 to 1). The value of 0 indicates no agreement, and 1 indicates complete reliability or agreement.

We calculated Cohen's Kappa and Pearson's correlation coefficient separately for two clinical and three non-clinical annotators (or reviewers). Our clinical annotators achieved a Cohen's kappa coefficient of 0.45 and a correlation coefficient of 0.87, indicating moderate agreement and strong correlation, respectively. Among non-clinical annotators, annotators 1 and 2 achieved a higher Cohen's kappa coefficient of 0.7 and a correlation coefficient of 0.97, demonstrating good reliability in their assessments. Table A.5a displays the Cohen's Kappa coefficient, while Table A.5b shows the correlation coefficient among clinical annotators. Similarly, Table A.6a presents the Cohen's Kappa coefficient, and Table A.6b displays the correlation coefficient among non-clinical annotators.

5 Generalization

To assess the generalizability of our best fine-tuned model (BART-large CNN), we utilized the publicly available D4[6] dataset released by (Yao et al., 2022). We used three independent non-clinical reviewers to rate the generated summaries by our best fine-tuned model of ten randomly selected conversations from the D4 dataset. The parameters utilized for evaluating the generated summaries included *completeness, relevance to the medical context, fluency, clarity, missingness, hallucination, contradiction*, and *repetitions* discussed in Section 4.2. It is important to note that the D4 dataset was in Chinese language. Therefore, we utilized Google Translate to translate the conversations from Chinese to English. We extracted ten doctor-patient conversations and assigned a dummy participant identifier to these files. Further, we shared the translated English conversation and their corresponding fine-tuned BART-large-CNN model-generated summaries.

Upon reviewing the reviewers' ratings, we found that the best fine-tuned model's summary scored well in *relevance, fluency, clarity*, and *repetition* as shown in Table 4. However, the generated summary was slightly lacking in terms of *missing information, hallucination,* and *contradiction*. Tables A.7 and A.8 in the appendix present dialogue conversations taken from (Yao et al., 2022) alongside the corresponding summaries generated by the fine-tuned BART-large-CNN model.

[6] https://x-lance.github.io/D4/

	Completeness (μ, σ)	Relevance (μ, σ)	Fluency (μ, σ)	Clarity (μ, σ)	Missingness (μ, σ)	Hallucination (μ, σ)	Contradiction (μ, σ)	Repetition (μ, σ)
Reference summary	(4.66,0.51)	(4.70,0.50)	(4.36,0.74)	(4.52,0.67)	(1.34,0.62)	(1.10,0.36)	(1.60,0.85)	(1.10,0.36)
Best model summary	(4.10,0.93)	(4.12,0.93)	(4.44,0.64)	(4.54,0.61)	(2.08,1.17)	(1.02,0.14)	(2.04,1.22)	(1.10,0.46)

Table 3: Average human evaluation scores on ten reference and best-model (i.e., BART-large-CNN) generated summaries on eight evaluation parameters. For *Completeness*, *Relevance*, *Fluency*, and *Clarity*, a rating closer to 5 indicates the best, whereas for *Missingness*, *Hallucination*, *Contradiction*, and *Repetition*, a rating closer to 1 is preferable.

	Completeness (μ, σ)	Relevance (μ, σ)	Fluency (μ, σ)	Clarity (μ, σ)	Missingness (μ, σ)	Hallucination (μ, σ)	Contradiction (μ, σ)	Repetition (μ, σ)
Generated Summary	(4.43, 0367)	(4.43, 0.73)	(4.37, 0.62)	(4.80, 0.48)	(1.57, 0.73)	(1.5, 0.63)	(1.67, 0.76)	(1, 0)

Table 4: Average human evaluation scores of best model (BART-large-CNN) generated summaries from ten conversations of the D4 dataset. For *Completeness*, *Relevance*, *Fluency*, and *Clarity*, a rating closer to 5 indicates the best, whereas for *Missingness*, *Hallucination*, *Contradiction*, and *Repetition*, a rating closer to 1 is preferable.

6 Comparison with the previous work

Our work represents the first attempt to summarize psychological conversation data, which differs from traditional text summarization. However, it shares similarities with dialogue summarization, such as summarizing conversations between individuals or medical dialogues between doctors and patients. On comparing (see Table 5) our accuracy to the only work done in psychological conversation summary by (Yao et al., 2022), our model trained on our dataset achieved a ROUGE-L score of 0.764, whereas they achieved only 0.26. Moreover, our fine-tuned model produced fluent and comprehensive summaries even when applied to the dataset used by (Yao et al., 2022).

Table 5 presents a comparative report of our work with existing research in doctor-patient conversation analysis. The table shows that our fine-tuned model outperforms the existing work (Yao et al., 2022; Krishna et al., 2021; Michalopoulos et al., 2022; Zhang et al., 2021) in terms of the ROUGE metric. However, it is essential to note that (Yao et al., 2022; Krishna et al., 2021; Zhang et al., 2021) fine-tuned existing state-of-the-art models while (Michalopoulos et al., 2022) developed the model from scratch. It is essential to recognize that all of these works utilized different datasets, whereas we have demonstrated the effectiveness of our model on our and the D4 dataset shared by (Yao et al., 2022). However, it is important to note that existing studies have their own specific objectives beyond solely summarizing entire conversations. While our work primarily aims at generating summaries of psychological conversations, it encounters its own challenges, such as dealing with lengthy conversation data, resulting in longer utterances.

7 Conclusion

The automatic generation of medical summaries from psychological patient conversations faces several challenges, including limited availability of publicly available data, significant domain shift from the typical pre-training text for transformer models, and unstructured lengthy dialogues. This paper investigates the potential of using pre-trained transformer models to summarize psychological patient conversations. We demonstrate that we can generate fluent and adequate summaries even with limited training data by fine-tuning transformer models on a specific dataset. Our resulting models outperform the performance of pre-trained models and surpass the quality of previously published work on this task. We evaluate transformer models for handling psychological conversations, compare pre-trained models with fine-tuned ones, and conduct extensive and intensive evaluations.

8 Ethical Consideration

Indeed, our psychological conversation data contained sensitive personal information about the participants and their past and present experiences. Therefore, we utilized anonymized numerical identifiers to store the participants' data for storage and further use. We ensured that the personal participants' information, such as name, age, and email address, could not be traced back using the anonymized numerical identifiers. Additionally, this study was approved by the ethics committee of the host institute.

Reference	Model (own/ fine-tuned)	Dataset	ROUGE-1	ROUGE-2	ROUGE-L
(Krishna et al., 2021)	fine-tuned	Medical (Own prepared)	0.57	0.29	0.38
	fine-tuned	AMI medical corpus	0.45	0.17	0.24
(Michalopoulos et al., 2022)	own	MEDIQA 2021 - history of present illness	0.48	-	0.35
	own	MEDIQA 2021 - physical examination	0.68	-	0.64
	own	MEDIQA 2021 - assessment and plan	0.44	-	0.37
	own	MEDIQA 2021 - diagnostic imaging results	0.27	-	0.26
(Song et al., 2020)	fine-tuned	Medical problem Description	0.91	0.87	0.91
	fine-tuned	Medical diagnosis or treatment	0.80	0.72	0.80
	fine-tuned	Medical problem Description	0.91	0.87	0.91
	fine-tuned	Medical diagnosis or treatment	0.81	0.73	0.81
(Zhang et al., 2021)	fine-tuned	Doctor patient conversation	0.46	0.19	0.44
(Yao et al., 2022)	fine-tuned	Chinese psychological conversation	-	-	0.26
Our Paper	**fine-tuned**	**Psychological conversation (own)**	**0.81**	**0.69**	**0.76**

Table 5: Comparison of our best model results in terms of ROUGE with existing works.

8.1 Implications

The pre-trained model demonstrated its effectiveness on our dataset. The models used in this paper were able to learn from just 250 conversations in a fewer number of epochs. This indicates that in the future, rather than developing models from scratch, leveraging pre-trained models may yield better results. Since developing models from scratch would require large datasets and more time for training and fine-tuning the model, thus utilizing already trained large models tailored to specific tasks could be a more efficient strategy.

While selecting the models for fine-tuning, we hypothesized that the BART-large-xsum-samsum model trained on dialogue summarization data would yield better results than other summarization models. Initially, our hypothesis held for a smaller number of epochs. However, we observed that the BART-large-CNN model outperformed in terms of all ROUGE metrics, indicating that our hypothesis was incorrect. Nevertheless, further exploration is warranted.

In this work, we presented the best fine-tuned summarization models for generating accurate and concise summaries from MSEs for the attending doctor. The goal was to leverage state-of-the-art technologies to reduce the workload of already overburdened psychiatrists. The primary intention of this technology is not to replace doctors but to serve as an assistant to attending doctors by offering concise summaries of patients' mental health. This approach holds particular promise for implementation in low-income countries with a shortage of mental health professionals. However, further research is necessary to address privacy concerns and ensure the accuracy of the data utilized.

9 Limitations

In this work, we achieved a better ROUGE score by comparing the generated and human reference summaries. However, our work does have several limitations, as outlined below:

1. When conducting MSE, it is important to note that MSE also encompasses the physical behavior and appearance of the participants, which, unfortunately, we were unable to incorporate in this work. However, this could be addressed by implementing a module where the front camera or webcam of participants' phones is activated while recording their responses.

2. There were several instances where the participants' utterances were unclear to the reviewers. In real-world scenarios, when a patient's utterance is unclear, a doctor typically asks them to repeat and explain. However, in our case, this poses a major challenge. This issue could potentially be mitigated by testing the user's response for fluency and completeness after each utterance. If the model detects an issue, a new prompt could be sent to the user to encourage them to elaborate on their answers.

References

Duy-Cat Can, Quoc-An Nguyen, Binh-Nguyen Nguyen, Minh-Quang Nguyen, Khanh-Vinh Nguyen, Trung-Hieu Do, and Hoang-Quynh Le. 2023. Uetcorn at mediqa-sum 2023: Template-based summarization for clinical note generation from doctor-patient conversation. In *CLEF*.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 93–98.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, et al. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the first workshop on natural language processing for medical conversations*, pages 22–30.

Elena Filatova and Vasileios Hatzivassiloglou. 2004. Event-based extractive summarization.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *EMNLP-IJCNLP 2019*, page 70.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469.

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2021. Generating soap notes from doctor-patient conversations using modular summarization techniques. 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 4958–4972.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Promita Majumdar. 2022. Covid-19, unforeseen crises and the launch of national tele-mental health program in india. *Journal of Mental Health*, 31(4):451–452.

David C Martin. 1990. The mental status examination. *Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition.*

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. *Mining text data*, pages 43–76.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Graciela Rojas, Vania Martínez, Pablo Martínez, Pamela Franco, and Álvaro Jiménez-Molina. 2019. Improving mental health care in developing countries through digital technologies: a mini narrative review of the chilean case. *Frontiers in public health*, 7:391.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Benedetto Saraceno, Mark van Ommeren, Rajaie Batniji, Alex Cohen, Oye Gureje, John Mahoney, Devi Sridhar, and Chris Underhill. 2007. Barriers to improvement of mental health services in low-income and middle-income countries. *The Lancet*, 370(9593):1164–1174.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Rachel M Voss et al. 2019. Mental status examination.

World Health Organization WHO. 2022. World mental health report: transforming mental health for all.

Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. 2022. D4: a chinese dialogue dataset for depression-diagnosis-oriented chat. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2438–2459.

Jiseon Yun, Jae Eui Sohn, and Sunghyon Kyeong. 2023. Fine-tuning pretrained language models to enhance dialogue summarization in customer service centers. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 365–373.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

A Appendix

A.1 Summary Template

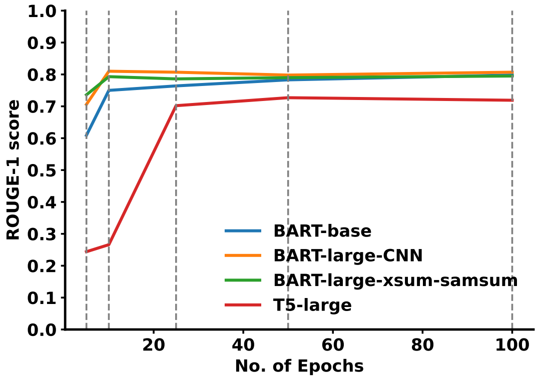
Patient is a ___ year old [girl/boy/lady/man]. [His/Her] mood is generally____________ [and remains study/but goes up and down] throughout the day. [He/She] [takes/does not take] part in extracurricular activities and _________ [socializes/does not socialize] socialize with others. For daily frustration [He/She] (*activities*). [He/She] [feels/does not feel] academic pressure and for this [He/She] (*activities*). [His/Her]concentration and task atten- ding ability is [good/bad]. [He/She] [feels/does not feel] difficulty with memory. [He/She] feels better by (*activities*). [He/She] [feels /does not feel] supported by his family and friends. On a bad day, [he/she] prefers ______________. [He/She] is [experiencing/ not experiencing] _______ [stress/anxiety/depression] symptoms such as _____________.

A.2 Hyperparameters

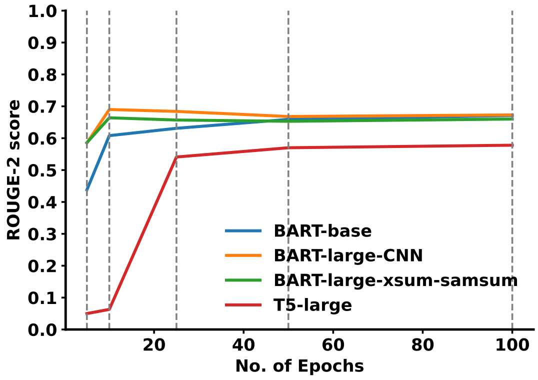
These are the hyperparameters we used across four models - BART base, BART-large-CNN, T5 large, and BART-large-xsum-samsum, using the Pytorch module: { *max token length*: 1024 tokens, *warmup steps*: 500, *weight decay*: 0.01, *evaluation strategy*: ‘steps’, *evaluation step*s: 500, *save steps*: 1e6, *gradient accumulation steps*: 16 }. The models were trained on an *NVIDIA A100-PCIE-40GB GPU*, with an average training time of 45 minutes.

MSE Questionnaires
Q1. Please describe your social life at the *anonymized* campus. Are you actively participating in extracurricular activities, interacting with others, or taking initiative to socialize with others?
Q2. Describe your typical daily Mood?
Q3. Does your Mood remain steady or goes up and down throughout the day without any reason or on trivial matters?
Q4. How do you handle day-to-day irritations or frustrations?
Q5. How do you handle pressure related to academics?
Q6. Describe your ability to attend to the task at hand or concentrate on daily tasks (academic, non-academic)?
Q7. Have you noticed any difficulties with memory, such as unable to register new information, forgetting recent events, or not able to recall older personal/factual events?
Q8. What do you do to feel better? For example, some people take caffeine, talk with people, or watch movies to feel better.
Q9. Describe how supported you feel by others (e.g., friends, family) around you and how they help you?
Q10. What do you usually do when you have a bad day or when you are not able to concentrate on work?
Q11. Are you experiencing symptoms of stress, anxiety, or depression? If yes, describe the symptoms?
Q12. Are you doing anything (by self or help seeking) for the ongoing stress, anxiety, or depression, if any? If yes, what?

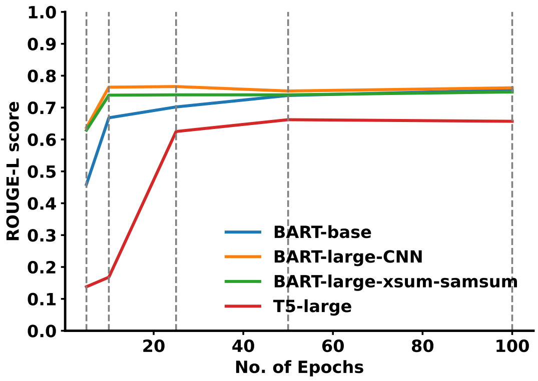
Table A.1: Finalized MSE Questionnaire



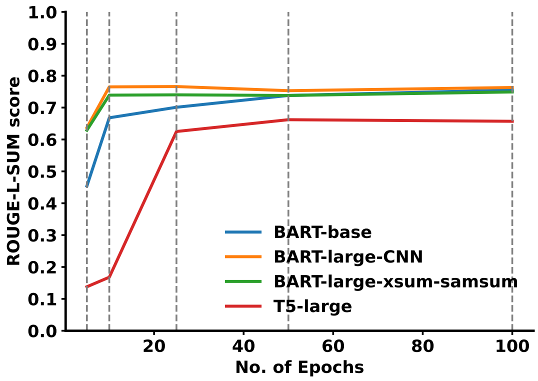
(a) ROUGE-1 score



(b) ROUGE-2 score



(c) ROUGE-L score



(d) ROUGE-L-SUM score

Figure A.1: ROUGE metric obtained after fine-tuning on BART-base, BART-large-CNN, T5 large, and BART-large-xsum-samsum on # epochs= [5,10,25,50,100]

Doctor: *What is your patient ID?*
Patient: *#3A4C307*
Doctor: *What is your age?*
Patient: *19*
Doctor: *What is your gender?*
Patient: *Female*
Doctor: *Please describe your social life at the *anonymized* campus. Are you actively participating in extracurricular activities, interacting with others, or taking initiative to socialize with others?*
Patient: *I am part of football team and a core member in Physics club and Singularity working team. I also make contacts with my seniors and other staffs for both personal and official discussions. I volunteer for community fests and other initiatives.*
Doctor: *Describe your typical daily Mood?*
Patient: *I always try to find happiness in every single moment of my life. But at times I turnout t be moody.*
Doctor: *Does your Mood remain steady or goes up and down throughout the day without any reason or on trivial matters?*
Patient: *My mood is dynamic. It goes up and down for both valid and unknown reasons. I get upset on simple jokes and responses from my close circle.*
Doctor: *How do you handle day-to-day irritations or frustrations?*
Patient: *I try to connect more with the Almighty through daily prayers. But mostly I prefer sleeping with no disturbance for hours. Nowadays I try to engage myself with a busy schedule and locations.*
Doctor: *How do you handle pressure related to academics?*
Patient: *lately I started purposeful ignorance of academic pressure. I will engage my times studying or with close friend. I also try to phone my parents when I feel so exhausted.*
Doctor: *Describe your ability to attend to the task at hand or concentrate on daily tasks (academic, non-academic)?*
Patient: *I am mostly able to focus on my task and complete on time. But when I am in a bad mood I will distract myself from the task with social media and resume when I feel fine.*
Doctor: *Have you noticed any difficulties with memory, such as unable to register new information, forgetting recent events, or not able to recall older personal/factual events?*
Patient: *Yes I do, and only very lately. I find it very difficult to comprehend what I see and try reading. I also noticed forgetting recent events which where not very important but still to be considered. I also have difficulty in recalling but the least.*
Doctor: *What do you do to feel better? For example, some people take caffeine, talk with people, or watch movies to feel better.*
Patient: *Sleep mostly. But if it is with communication gap, I only settle after conveying my last note. I also sing a song or try dancing in my room but I prefer privacy for this*
Doctor: *Describe how supported you feel by others (e.g., friends, family) around you and how they help you?*
Patient: *I feel supported very less even from family. And so I don't expect any support from anyone and try to figure out all alone.*
Doctor: *What do you usually do when you have a bad day or when you are not able to concentrate on work?*
Patient: *I sleep for hours or the entire day. I also get some ease after crying or talking about it. I used talk to myself which helped me figure out the situation and motivated to push through.*
Doctor: Are you experiencing symptoms of stress, anxiety, or depression? If yes, describe the symptoms?
Patient: *Yes, all stress, anxiety and depression*
Doctor: *Are you doing anything (by self or help seeking) for the ongoing stress, anxiety, or depression, if any? If yes, what?*
Patient: *Yes, I'm reading books on self-development and self-improvement.*

Table A.2: Doctor-patient conversation dialogue of an anonymized participant.

No. of Epochs	Generated Summary
0 (Without tuning)	Patient: I always try to find happiness in every single moment of my life. I get upset on simple jokes and responses from my close circle. I try to connect more with the Almighty through daily prayers. I sleep for hours or the entire day. I also get some ease after crying or talking about it.
5	Patient is a 19 year old girl. Her mood is generally happy but goes up and down throughout the day. She takes part in extracurricular activities and socializes with others. For daily frustration she tries to connect more with the Almighty. She feels academic pressure. Her concentration and task attending ability is good. She does not feel difficulty with memory. On a bad day, she sleeps for hours, talks to herself, and
10	Patient is a 19 year old girl. Her mood is generally happy but goes up and down throughout the day. She takes part in extracurricular activities and socializes with others. For daily frustration she connects with the Almighty and sleeps. She feels academic pressure. Her concentration and task attending ability is good. She does feel difficulty with memory such as forgetting recent events and older personal/factual events.She feels better by sleeping, singing songs, dancing or dancing in her room. She does not feel supported by her family and friends. On a bad day, she sleeps or talks to herself. She is feeling stress, anxiety and depression symptoms. To overcome this she reads self improvement books.
25	Patient is a 19 year old girl. Her mood is generally happy but goes up and down throughout the day. She takes part in extracurricular activities and socializes with others. For daily frustration she connects with the Almighty and sleeps. She feels academic pressure. Her concentration and task attending ability is okay, she feels difficulty with memory. She does not feel supported by her family and friends. On a bad day, she sleeps or cries and talks to herself. She is feeling stress, anxiety and depression symptoms. To overcome this she reads self development and self improvement books.
50	Patient is a 19 year old girl. Her mood is generally happy but goes up and down throughout the day. She takes part in extracurricular activities and socializes with others. For daily frustration she tries to connect with the lord through daily prayers. She does not feel academic pressure. Her concentration and task attending ability is good. She feels difficulty with memory. She feel better by sleeping, singing songs, dancing, and hiding in her room. She doesn't feel supported by her family and friends. On a bad day, she sleeps or talks to herself. She is feeling all stress, anxiety and depression symptoms.
100	Patient is a 19 year old girl. Her mood is generally happy but goes up and down throughout the day. She takes part in extracurricular activities and socializes with others. For daily frustration she connects with the Almighty and sleeps. She does not feel academic pressure. Her concentration and task attending ability is good. She feels difficulty with memory such as forgetting recent events and personal facts and recalling the least.. She feels better by sleeping, singing songs, dancing in room, and asking for privacy. She did not feel supported by her family and friends. On a bad day, she sleeps or talks to herself. She is feeling all stress, anxiety and depression symptoms

Table A.3: BART-large-CNN generated summaries at different epochs tested on conversation given in Table A.2 in the Appendix

	Completeness (μ, σ)	Relevance (μ, σ)	Fluency (μ, σ)	Clarity (μ, σ)	Missingness (μ, σ)	Hallucination (μ, σ)	Contradiction (μ, σ)	Repetition (μ, σ)
Reference summary	(4.15, 0.48)	(4.25, 0.55)	(4.15, 0.48)	(3.90, 0.64)	(1.80, 0.76)	(1.15, 0.37)	(1.55, 0.60)	(1.10, 0.30)
Best model generated summary	(3.85, 0.74)	(4.10, 0.55)	(4.10, 0.55)	(4.00, 0.56)	(2.20, 0.89)	(1, 0)	(1.55, 0.75)	(1.0, 0.00)

(a) Human evaluation scores obtained by averaging the ratings provided by two clinician on ten conversations

	Completeness (μ, σ)	Relevance (μ, σ)	Fluency (μ, σ)	Clarity (μ, σ)	Missingness (μ, σ)	Hallucination (μ, σ)	Contradiction (μ, σ)	Repetition (μ, σ)
Reference summary	(5.00, 0.00)	(5.00, 0.00)	(4.50, 0.86)	(4.93, 0.25)	(1.03, 0.18)	(1.06, 0.36)	(1.63, 0.99)	(1.10, 0.40)
Best model generated summary	(4.26, 1.01)	(4.13, 1.13)	(4.67, 0.61)	(4.9, 0.30)	(2, 1.33)	(1.03, 0.18)	(2.36, 1.37)	(1.16, 0.59)

(b) Human evaluation scores obtained by averaging the ratings provided by three non-clinician on ten conversations

Table A.4: Human evaluation scores. For *Completeness*, *Relevance*, *Fluency*, and *Clarity*, a rating closer to 5 indicates the best, whereas for *Missingness*, *Hallucination*, *Contradiction*, and *Repetition*, a rating closer to 1 is preferable.

	Annotator 1	Annotator 2
Annotator 1	1.00	0.45
Annotator 2	0.45	1.00

(a) Cohen's Kappa Coefficient

	Annotator 1	Annotator 2
Annotator 1	1.00	0.87
Annotator 2	0.87	1.00

(b) Pearson's Correlation Coefficient

Table A.5: Inter-rater Reliability (Clinical Annotators)

	Annotator 1	Annotator 2	Annotator 3
Annotator 1	1.00	0.70	0.43
Annotator 2	0.70	1.00	0.42
Annotator 3	0.43	0.42	1.00

(a) Cohen's Kappa Coefficient

	Annotator 1	Annotator 2	Annotator 3
Annotator 1	1.00	0.97	0.75
Annotator 2	0.97	1.00	0.75
Annotator 3	0.75	0.75	1.00

(b) Pearson's Correlation Coefficient

Table A.6: Inter-rater Reliability (Non-Clinical Annotators)

Conversation	Generated Summary
Doctor: What is your patient ID? Patient: 1001 Doctor: What is your age? Patient: 32 Doctor: What is your gender? Patient: Female Patient: "Okay" Doctor: "Hello" Doctor: "What are your main problems recently?" Patient: "I haven't been feeling well recently, and I feel a little tight in my chest" Doctor: "Have you ever gone to the hospital to see a doctor?" Patient: "Not yet, I don't have much time recently" Patient: "Maybe it will take two weeks to go" Doctor: "Hmm, let's take some time to see if you have any emotional problems recently" Patient: "There's nothing wrong with my mood, I just feel mentally tired recently" Doctor: "Do you feel tired without doing anything?" Patient: "I feel like this, I don't want to move" Doctor: "Then do you feel like you don't want to work?" Patient: "I don't have enough energy to work" Patient: "Yes" Patient: "But I have to work" Patient: "It's quite stressful" Doctor: "Have you ever felt that you have lost interest in your past hobbies?" Patient: "I should still be interested" Patient: "I just don't have enough time to develop my hobbies" Doctor: "Sleep, eat, etc." Doctor: "Is everything normal?" Patient: "fairly normal" Doctor: "Will you feel dizzy or nauseous?" Patient: "I get dizzy occasionally" Doctor: "Do you feel lack of confidence? You are always worried about not doing well" Patient: "No, I don't have time to worry about this or that" Doctor: "It sounds like you are doing well lately" Doctor: "The consultation ends here"	Patient is a 32 year old girl. She is feeling a little tight in her chest and hasn't been feeling well recently. She hasn't gone to the hospital to see a doctor. Her mood is good and she feels mentally tired. She feels tired without doing anything and doesn't want to work. She also feels lack of confidence. She has lost interest in her past hobbies.

Table A.7: BART-large-CNN generated summary on a sample Chinese psychological conversation taken from (Yao et al., 2022) - I

Conversation	Generated Summary
Doctor: What is your patient ID? Patient: 1010 Doctor: What is your age? Patient: 21 Doctor: What is your gender? Patient: Female Patient: "Hello" Patient: "I feel tired recently" Doctor: "How long has this situation been going on?" Patient: "About a week or so" Doctor: "What's the reason?" Patient: "Because I am a student, I have been under a lot of pressure to study recently" Doctor: "Okay, it's normal to have academic pressure. You can usually find more people to talk to." Patient: "Okay" Doctor: "Are you feeling unhappy recently?" Patient: "No" Patient: "Just a little tiring easily" Doctor: "So you can still do the things you were interested in before?" Patient: "Nothing has changed" Doctor: "Do you feel inferior easily?" Patient: "Some people feel that they are always not as good as others" Doctor: "It's okay, everyone has their own strengths, don't envy others" Patient: "Yeah" Doctor: "Has there been any change in appetite?" Patient: "No" Doctor: "What about sleep?" Patient: "Neither" Doctor: "Well, I think you don't have a big problem. Students are all under academic pressure." Doctor: "Remember to go out for a walk more often and relax" Doctor: "Maybe talk to your friends more" Patient: "Okay, thank you doctor" Doctor: "Then you can do more things you like" Doctor: "You can exercise more and write to help relax" Patient: "Yeah, I will do it" Doctor: "Then it's over"	Patient is a 21 year old girl. She is feeling tired and under a lot of pressure to study. Her mood is normal and nothing has changed. She does not feel unhappy, just tired and tiring easily. Her appetite and sleep is not affected. She feels better by going for a walk, talking to friends, exercising and writing.

Table A.8: BART-large-CNN generated summary on a sample Chinese psychological conversation taken from (Yao et al., 2022) - II