Mental Health Counselling Summarization

Abhishek Jha
abhishek22023@iiitd.ac.in

1 Introduction

Mental health has emerged as one of the most pressing global challenges of the 21st century. Mental health counseling typically involves emotionally rich, highly personalized conversations between therapists and clients. These dialogues often contain subtle cues about the individual's psychological state, coping strategies, and behavioral progress over time. However, manually documenting and reviewing these conversations is both labor-intensive and prone to human error or cognitive bias.

Consider a scenario where a therapist meets multiple clients in a single day. Without structured documentation, critical insights from earlier sessions may be overlooked, making it difficult to track patient progress or recall recurring themes. For example, when supporting a client with recurring anxiety, a therapist must identify gradual changes in emotional patterns or evaluate the impact of suggested coping mechanisms. In such cases, a system that can automatically distill key moments from prior sessions can significantly enhance therapeutic continuity and effectiveness.

An **automated summarization system** offers a promising solution by efficiently extracting **clinically relevant insights** from dialogue. For instance, emotionally salient statements such as *"I feel like I'm second guessing everything"* or *"Situations are extremely stressful and hard on you"* can signal a shift in the client's emotional state. Capturing such utterances in a structured summary can help therapists quickly reorient themselves to a patient's history and needs.

This work aims to develop a summarization pipeline tailored for mental health counseling dialogues. By preserving **emotional**, **topical**, and **clinical nuances** in generated summaries, our system is designed to support therapists in delivering more informed, continuous, and responsive care.

2 Related Work

Summarizing mental health counseling conversations presents unique challenges compared to general-purpose dialogue summarization. These dialogues are often emotionally rich and therapeutically complex, requiring systems that can detect subtle affective signals, contextual relevance, and client-specific therapeutic progression. To address these challenges, prior research has explored domain-specific annotation, sentiment-aware modeling, and scalable transformer architectures.

One notable effort is **ConSum** by (Srivastava et al., 2022), which integrates the **Patient Health Questionnaire (PHQ-9)** as a clinical prior to identify relevant utterances within psychotherapy dialogues. Their annotated *MEMO* dataset structures conversations into meaningful therapeutic components such as *Symptom and History (SH)*, *Patient Discovery (PD)*, *Reflecting (RT)*, and *Discussion Filler (DF)*. This hierarchical framework helps the model focus on clinically salient information, leading to summaries that were both coherent and validated by domain experts.

To further enhance summary relevance, (Ma et al., 2018) proposed a unified model that jointly performs abstractive summarization and sentiment classification. By using a multi-view attention mechanism, their architecture learns distinct representations for each task, allowing sentiment signals to guide the summarization process. This joint formulation improves summary fidelity while ensuring that emotionally charged content particularly valuable in mental health contexts is retained.

Complementing these domain-specific approaches, Guo et al. (2021) introduced *LongT5*, a transformer model optimized for long-sequence inputs. By combining PEGASUS style pretraining with the Transient Global (TGlobal) attention mechanism, LongT5 achieves high performance on both summarization and question-answering tasks.

Its ability to handle extended inputs makes it especially suited for processing therapy conversations, which can span multiple turns and complex topics.

Together, these studies demonstrate that effective summarization in mental health contexts requires both content-aware filtering and architectural scalability. Our work builds on these insights by incorporating clinically grounded relevance cues—derived from sentiment intensity and PHQ-9 indicators—into the summarization pipeline to better capture therapeutic significance in generated summaries.

3 Dataset

Our dataset consists of 190 therapist-client counseling conversations addressing various mental health concerns. It is partitioned into training (131), validation (21), and testing (39) subsets with a 70:10:20 split. Each conversation is stored in structured CSV files. The essential annotated fields for modeling are described in Table 1.

Table 1: Dataset Fields and their Usage

Field	**Purpose**
Type	Identify speaker (Therapist/Client)
Utterance	Actual spoken text
Dialogue_Act	Shows speaker's intent (question/ answer/ affirmation)
Sub Topic	Represents thematic subdivisions (routine, story, symptom, inactive)
Emotion	Numeric emotion value, useful for sentiment analysis
Summary	Reference summaries
Primary Topic	Main session theme
Secondary Topic	Supporting session theme

These structured annotations enable effective summarization by capturing relevant conversational details.

4 Methodology

We propose a multi-stage summarization framework tailored for mental health counseling conversations. Our pipeline begins with preprocessing and restructuring of raw dialogues, followed by the computation of utterance-level relevance scores grounded in clinical and emotional cues. These enriched inputs are then used to fine-tune state-of-the-art abstractive summarization models, enabling them to generate contextually aware and clinically meaningful summaries.

4.1 Preprocessing and Structuring Dialogues

We begin by removing filler words and non-essential entries such as those labeled `primary_topic`, `secondary_topic`, and `summary`. All emotion values are averaged to compute a *global emotion score*, representing the emotional tone of the entire dialogue.

Each utterance is then restructured with a consistent format. For example:

> *Therapist [DialogueFunction=question] (Sub_topic=routine) (Emotion=1): How have you been sleeping lately?*

This structure is used to build a flattened version of the dialogue, facilitating input to downstream models while preserving emotional and topical context. The processed output from this stage serves as the input for our relevance modeling pipeline described next.

4.2 Modeling Relevance using Sentiment and PHQ-9 Cues

To identify clinically meaningful content within conversations, we introduced a scoring mechanism that combines emotional intensity with mental health symptom relevance. This is done through two primary signals: sentiment classification and PHQ-9 detection.

First, we fine-tuned a **RoBERTa-base sentiment classifier** on our dialogue dataset to classify each utterance into one of three sentiment categories: positive, neutral, and negative. From the model's output probabilities, we compute a sentiment intensity score as follows:

$$\text{Sentiment Score} = P_{\text{positive}} - P_{\text{negative}}$$

Utterances with a high absolute sentiment score are treated as emotionally charged and potentially more relevant.

In parallel, we implemented a **PHQ-9 detector** based on the clinically validated nine-item depression scale. Each utterance is evaluated against two types of PHQ-9 signals:

- *Keyword Match Score*, based on the presence of depression-related terms such as "worthless", "fatigue", or "better off dead".
- *Embedding Similarity Score*, calculated as the cosine similarity between the utterance and sentence embeddings of PHQ-9 item descriptions (e.g., "Feeling down, depressed, or hopeless") using the `all-MiniLM-L6-v2` model.

The overall PHQ score for an utterance is computed as a weighted sum of both:

$$\text{PHQ Score} = \alpha \cdot \text{Embedding Score} + \beta \cdot \text{Match Score}$$

where $\alpha = 1.0$ and $\beta = 0.5$ in our experiments.

To compute the final **relevance score**, we combine both components as follows:

$$\text{Rel Score} = 2 \cdot |\text{Sentiment Score}| + 1.5 \cdot \text{PHQ Score}$$

Utterances with relevance scores above a predefined threshold are tagged as `(Relevance=High)`, while others are tagged as `(Relevance=Low)`. This relevance label is appended directly to the utterance in the preprocessed dataset (from Section 4.1) and serves as an auxiliary signal during summarization model training.

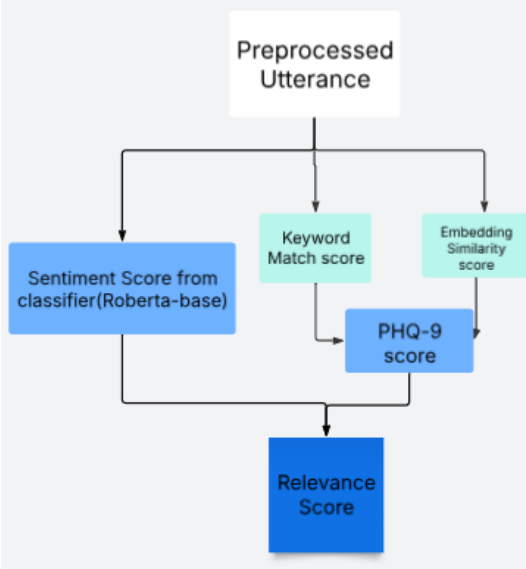

Figure 1: Relevance scoring pipeline using sentiment classifier and PHQ-9 signals.

4.3 Baselines

Our methodology begins with strong pretrained abstractive summarization models, which serve not only as benchmarks but also as foundational architectures for our subsequent enhancements. We selected **Pegasus-large** and **T5-large** as our baseline models due to their proven effectiveness on long-form summarization tasks and their flexibility for fine-tuning.

Pegasus-large is specifically pre-trained for summarization using a gap-sentence generation objective, making it well-suited for tasks involving information-dense inputs such as therapeutic dialogues.

T5-large, by contrast, adopts a universal text-to-text framework that generalizes well across various NLP problems, including summarization, question answering, and translation.

Initially, we fine-tuned these models on our counseling dataset without any domain-specific augmentations. The goal was to establish a reference point for model performance using only the structured dialogues (from Section 4.1) as input and the ground-truth summaries as targets.

4.4 Relevance-Guided Summarization

Following the baseline evaluation (Section 4.3), we extend our methodology by fine-tuning a single model—**Flan-T5-large**—on a relevance-augmented version of the training data. Unlike the baselines, which were trained on the original preprocessed dialogues, this new setup incorporates additional supervision through utterance-level annotations encoding clinical and emotional salience.

Relevance-Aware Input Representation: As described in Section 4.2, each dialogue utterance is scored using a hybrid mechanism that integrates sentiment intensity with PHQ-9 symptom cues. Based on the final relevance score, utterances are labeled as either `(Relevance=High)` or `(Relevance=Low)`. These labels are appended along with metadata including speaker type, dialogue function, sub-topic, and emotion value. An example of this enriched representation is shown below:

> *Therapist [DialogueFunction=question] (Sub_topic=routine) (Emotion=1) (Relevance=High): How have you been sleeping lately?*

This structured format helps the model focus on utterances that are therapeutically significant and emotionally informative.

Modeling Strategy: We fine-tune the **Flan-T5-large** model using these relevance-tagged dialogues as input, while keeping the decoder targets unchanged. No changes were made to the model architecture. Flan-T5 was chosen due to its instruction-tuned design and its strong performance on dialogue-centric NLP tasks. Its training on a diverse set of summarization, instruction-following, and conversational datasets makes it well-suited for capturing nuanced therapeutic context when guided by structured supervision.

Objective: This approach aims to assess

whether lightweight, input-level relevance supervision can guide the model to generate more clinically meaningful and emotionally resonant summaries, without requiring architectural modifications. Comparative results against the original baseline models (trained without relevance annotations) allow us to evaluate the effectiveness of this preprocessing strategy.

5 Experimental Setup

5.1 Baseline Setup

We fine-tuned two pretrained abstractive summarization models — **T5-large** and **Pegasus-large** — on the counseling dialogue dataset to establish performance baselines. These models were trained using dialogue-summary pairs after applying the preprocessing steps described earlier, which included flattening the dialogues, removing filler words, and truncating input/output sequences.

Tokenization: We used the official tokenizers provided by Hugging Face: `T5Tokenizer` for T5-large and `PegasusTokenizer` for Pegasus-large. T5 inputs were prefixed with `summarize:` to conform to its text-to-text formulation.

Input Formattin: All input sequences were truncated or padded to a maximum length of 512 tokens. Target summaries were similarly limited to 128 tokens. During training, label padding tokens were masked to exclude them from the loss computation.

Training Configuration: Both models were trained for a maximum of 10 epochs using the `AdamW` optimizer and early stopping based on validation loss (patience = 3). Beam search with 4 beams and a maximum generation length of 128 tokens was used during inference. The objective function was cross-entropy loss without label smoothing.

Hardware: All experiments were conducted on a single NVIDIA RTX A6000 GPU (48GB). We used mixed-precision training to accelerate computations and reduce memory consumption. All models were implemented using the PyTorch framework in combination with Hugging Face Transformers.

Model-specific hyperparameters are summarized below:

This baseline setup serves as the foundation for evaluating the impact of our proposed enhancements, which introduce clinically relevant features into the summarization pipeline.

Table 2: Model-specific training hyperparameters for baselines

Parameter	Pegasus-large	T5-large
Batch Size	8	4
Learning Rate	5×10^{-5}	3×10^{-5}
Epochs	10	10
Scheduler	StepLR	Linear Warmup

5.2 Relevance-Guided Fine-Tuning Setup

To evaluate the effectiveness of our proposed preprocessing method, we fine-tuned the **Flan-T5-large** model exclusively on the relevance-enriched dialogue dataset described in Section 4.2. In contrast, the baseline models—**Pegasus-large** and **T5-large**—were trained only on the original preprocessed dialogues and not re-trained on the enriched data. This distinction allows us to isolate the impact of incorporating clinical and emotional relevance cues into the input.

Flan-T5's fine-tuning was preceded by an extensive grid search over hyperparameters, aimed at optimizing model performance for summarizing emotionally rich, multi-turn counseling conversations. The tuning process spanned both training and decoding parameters.

The best-performing configuration, selected based on validation set performance, is presented in Table 3. These settings balance coverage, fluency, and repetition control in the generated summaries.

Table 3: Best-performing hyperparameters for Flan-T5 on relevance-enriched data

Hyperparameter	Value
Learning Rate	2.37×10^{-5}
Batch Size	2
Weight Decay	0.0312
Warmup Ratio	0.0312
Beam Width	2
No-Repeat n-gram Size	4
Repetition Penalty	2.20
Length Penalty	1.56

This relevance-guided fine-tuning setup enables Flan-T5 to attend more selectively to high-signal utterances, ultimately improving the informativeness, coherence, and clinical utility of the generated summaries.

6 Results

6.1 Evaluation Metrics

To evaluate the quality of generated summaries, we adopt a suite of widely used metrics that assess both lexical overlap and semantic similarity with reference summaries. Given the sensitive and context-dependent nature of mental health conversations, our choice of metrics aims to capture both surface-level fidelity and deeper semantic alignment.

ROUGE-1, **ROUGE-2**, and **ROUGE-L** are recall-oriented metrics that measure the overlap of unigrams, bigrams, and longest common subsequences respectively between the generated and reference summaries. These metrics are effective for evaluating how much of the essential content from the reference is retained in the prediction, which is crucial in high-stakes domains like counseling where missing key points could lead to misinterpretation.

BLEU is a precision-based metric that captures how many n-grams in the generated summary match the reference summary. While originally designed for machine translation, BLEU remains a useful complementary measure for assessing the fluency and correctness of phrasing in generated summaries.

BERTScore leverages contextual embeddings from large pretrained language models to compute semantic similarity between tokens in generated and reference summaries. It is particularly useful in our domain, where surface-level wording may vary, but the underlying meaning should remain intact. We report the F1 variant of BERTScore to capture the harmonic mean of precision and recall.

Together, these metrics provide a comprehensive evaluation: ROUGE and BLEU focus on lexical and structural fidelity, while BERTScore and BLEURT assess the semantic correctness and fluency — both of which are essential for faithfully summarizing nuanced therapy dialogues.

6.2 Results: Baselines vs Proposed Model

We evaluate the performance of the two baseline summarization models — **T5-large** and **Pegasus-large** — alongside our proposed approach, **RoBERTa-base-sentiment + PHQ-9 + Flan-T5**, on the test set using both lexical and semantic evaluation metrics, as outlined earlier. The goal is to assess the effectiveness of relevance-guided supervision in enhancing summary informativeness and emotional fidelity.

The table below reports the evlaluation metrics as mentioned above. The Flan-T5 model is trained on the relevance-enriched dataset, where utterances are tagged with clinically meaningful cues based on sentiment intensity and PHQ-9 signal matching.

Table 4: Performance Comparison: Baseline Models vs Proposed Model

Metric	Pegasus	T5	FlanT5 with rel score
ROUGE-1	31.35	35.04	39.31
ROUGE-2	11.18	12.30	15.30
ROUGE-L	19.68	21.53	24.70
BLEU	2.63	2.62	5.04
BERTScore	85.78	86.69	87.08

The loss plots in Figures 2 and 4 indicate steady convergence without overfitting, validating the use of early stopping during training. These plots further suggest that the models effectively learn from the enriched input signals introduced through the relevance-based annotation pipeline.

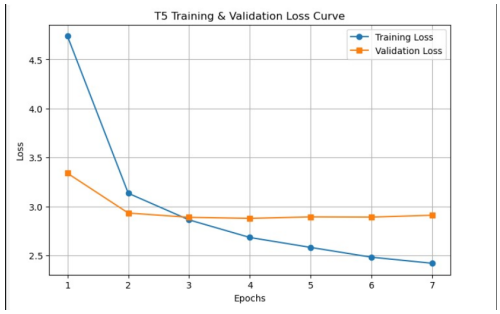

Figure 2: Training and validation loss curves for T5

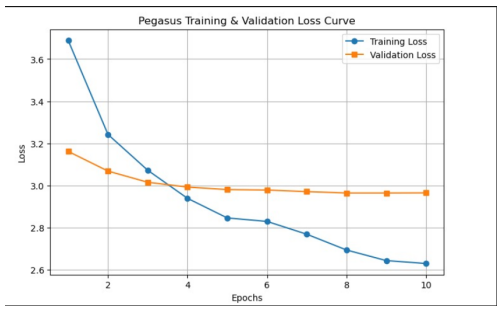

Figure 3: Training and validation loss curves for Pegasus

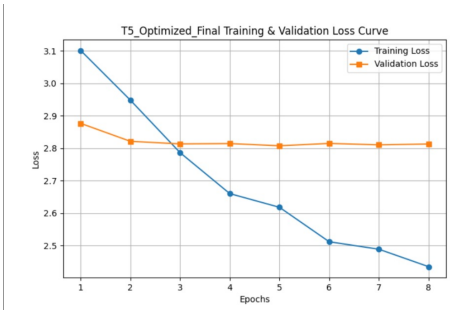

Figure 4: Training and validation loss curves for Roberta-base-sentiment + PHQ9 + FlanT5

7 Observations

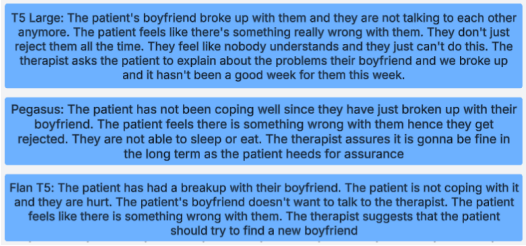

Figure 5: Generated summaries by Pegasus, T5, and the RoBERTa-base-sentiment + PHQ-9 + Flan-T5

We observe consistent improvements when comparing the baseline models—**Pegasus-large** and **T5-large**—with the proposed relevance-guided model, **Roberta-base-sentiment + PHQ9 + FlanT5**. These observations span across quantitative performance, training dynamics, and qualitative generation quality.

Quantitative Evaluation: As shown in Table 4, the relevance-aware FlanT5 model outperforms both baselines across all metrics. It achieves the highest ROUGE-1 (39.31), ROUGE-2 (15.30), and ROUGE-L (24.70), indicating stronger n-gram overlap with reference summaries. The BLEU score also improves substantially (5.04 vs. $\sim$2.6), reflecting enhanced precision in sequence generation. Moreover, the BERTScore of 87.08 surpasses that of T5 (86.69) and Pegasus (85.78), signifying better semantic alignment.

Training Behavior: Figures 2–4 illustrate the training and validation loss curves. While the baseline models show early stagnation in validation loss, the FlanT5 model exhibits a stable decline, suggesting improved generalization. We attribute this to the relevance-augmented input, which emphasizes salient clinical and emotional content, making learning more focused and efficient.

Qualitative Comparison: Figure 5 compares summaries produced by the three models. Baselines tend to capture surface-level details but often miss nuanced therapeutic signals. In contrast, the relevance-guided FlanT5 model accurately summarizes key emotional states and counselor interventions, reflecting deeper comprehension of conversational context.

Key Insight. These findings demonstrate that incorporating relevance supervision at the input level significantly enhances summarization performance. Without modifying model architecture, our approach enables the model to prioritize emotionally and clinically meaningful utterances, resulting in summaries that are both informative and contextually aligned with mental health counseling needs.

8 Conclusion and Future Work

In this work, we presented a relevance-guided summarization framework tailored for mental health counseling conversations. By integrating sentiment analysis and PHQ-9 symptom cues at the input level, our approach introduces lightweight supervision that enhances the model's ability to prioritize emotionally and clinically important utterances. We demonstrated that this preprocessing strategy—applied to a strong instruction-tuned model like Flan-T5—leads to consistent improvements over traditional baselines (T5 and Pegasus) across both lexical and semantic evaluation metrics.

Our results underscore the importance of domain-specific cues in dialogue summarization and suggest that emotional and therapeutic relevance can be effectively encoded without altering model architecture. The observed improvements in ROUGE, BLEU, and BERTScore metrics confirm the utility of relevance-aware input representations in generating more coherent, emotionally resonant, and clinically useful summaries.

As future work, we could explore enhancing the sentiment analysis model by incorporating more nuanced psychological indicators or extending the relevance scoring mechanism to include other mental health assessments. Additionally, experimenting with different model architectures or incorporating external knowledge sources could further refine the summarization process and improve the quality of the generated summaries.

References

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916*.

Shuming Ma, Xu Sun, Junyang Lin, and Xuancheng Ren. 2018. A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. *arXiv preprint arXiv:1805.01089*.

Aseem Srivastava, Tharun Suresh, Sarah P Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Counseling summarization using mental health knowledge guided utterance filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3920–3930.