Mental Health Counseling Summarization

Aniket Dwivedi, Keshav Sharma, Ritwik Ganguly
IIIT-DELHI

Abstract

In this paper we address the need for effective, concise and automated summarization of mental health dialogues. We worked on the dataset **MEMO_KDD_2022** dataset, introduced in KDD'22 conference. We explore the development of an NLP-based system designed to produce concise yet contextually rich summaries that capture key insights, patient concerns etc. The methodology involves rigorous preprocessing of data and investigate sequence-to-sequence modeling approaches, initially establishing a baseline with *T5-base*. To leverage advancements in large language models specialized for domain-specific understanding, we further explore fine-tuning *t5-large*, *MentalLLama* a Llama 2-based model tailored for mental health analysis. Fine-tuning incorporates instruction-following zero-shot prompt templates suitable for causal language models and employs PEFT techniques, specifically QLoRA (quantization with gradient checkpointing). The project also considers potential enhancements like multi-view summarization offering varying levels of detail to cater to diverse clinical needs. Later the model is evaluated on *BLeU*, *ROUGE* and *BERTScore* metrices demonstrates a significant improvement in summary fluency and accuracy. The expected outcome is an effective tool to aid clinicians, enhance the quality of care documentation, and support better mental health outcomes by reducing the burden of manual summarization.

1 Introduction

Dialogue summarization is a crucial subtask in natural language processing, particularly for domains where **long conversations** must be condensed into actionable, **concise summaries**. In the healthcare sector, especially mental health, summarizing therapy sessions can assist professionals in tracking patient progress, improving communication, and supporting clinical decision-making. Traditional models often fail to capture the nuances, specific counseling components (like symptom discussion, patient discovery, reflection), and structure inherent in therapeutic conversations.

The increasing demand for mental health services necessitates efficient clinical workflows, yet practitioners often face significant burdens in documenting and reviewing extensive counseling session dialogues. These conversations, rich in nuanced information crucial for diagnosis, treatment planning, and tracking patient progress, generate large volumes of unstructured text data, making manual summarization time-consuming and prone to variability. Automated summarization, leveraging recent advancements in Natural Language Processing (NLP) and Large Language Models (LLMs), offers a promising solution to alleviate this documentation load, enhance information retrieval, support clinical decision-making, and ensure continuity of care by providing concise, accurate, and context-aware digests of therapeutic interactions. While the application of LLMs in healthcare shows potential, challenges remain in ensuring reliability, maintaining patient privacy, and capturing the specific sensitivities of mental health discourse. This project aims to develop and evaluate an effective system for summarizing mental health counseling dialogues by exploring sequence-to-sequence architectures, beginning with a *T5-base* (220M) baseline and progressing to fine-tuning a specialized along with its large version *T5-large* (770M), also, instruction-aware LLM *MentaLLaMA* (7B) (Yang et al., 2024) proposed in in WWW'24. Also, we built the *BERT-base* as our fourth baseline model, that outperforms all other base line and reasonably performed good in this case. Our approach incorporates domain-specific data preprocessing, including speaker tagging and careful text normalization, alongside Parameter-Efficient Fine-Tuning techniques like QLoRA to adapt the large model efficiently, ultimately striving to generate summaries that are both clinically useful and computationally feasible.

In our mid-semester phase, we implemented the T5-Base model and achieved reasonable results. However, the limitations in capacity and abstraction motivated an upgrade to T5-Large, which offers significantly more parameters (770M vs. 220M) and hence potentially better performance in capturing complex dependencies and generating more nuanced summaries. Our work investigates whether this upgrade, combined with domain-specific techniques, yields tangible improvements in real-world mental health counseling summarization tasks. We aim to develop a tool that aids clinicians by reducing the burden of manual summarization and enhancing the quality of care documentation.

2 Literature Review

The field of automated text summarization has significantly advanced, shifting from primarily extractive methods to more sophisticated abstractive approaches that generate novel sentences. Transformer-based architectures, particularly models like T5, BART, and PEGASUS, have become prominent, demonstrating strong performance on general dialogue summarization benchmarks such as DialogSum. These models operate on the principle of converting NLP tasks into a text-to-text format, and larger models generally show better performance, albeit at higher computational costs.

Dataset Name	Size	Focus
MentalCLOUDS	191 sessions	Component-guided summarization (SH, PD, RT)
MentSum (Sotudeh et al., 2022)	24,000+ posts	Summarization of informal online mental health discussions
MentalChat16K (Xu et al., 2025)	16,000+ pairs	Conversational mental health assistance, empathetic AI solutions
MSE Questionnaire	9,720 utterances	Generation of Mental State Examination summaries
MEMO	Not specified	General counseling summarization

Table 1: Mental health datasets

However, the task of manually summarizing counselling sessions can be time-consuming and demanding for therapists. This administrative burden can potentially detract from the time and energy that mental health professionals can dedicate to direct interaction with clients, and in the face of increasing demand for services and a shortage of clinicians, it may contribute to burnout. (Adhikary et al., 2024).

A key challenge in mental health counseling summarization is capturing domain-specific nuances often missed by general models. Recent studies have started incorporating domain knowledge. For instance, Gaur et al. utilized Patient Health Questionnaire (PHQ-9) knowledge to build knowledge graphs and guide abstractive summarization for counseling sessions. Song et al. employed utterance labels (e.g., Problem Description, Diagnosis) to improve medical dialogue summarization, although their focus remained on extractive outputs. The ConSum model, (Srivastava et al., 2022), specifically builds on these ideas by using PHQ-9 similarity filtering and classifying utterances into 'counseling components' (like Symptom & History, Patient Discovery) to create more clinically relevant abstractive summaries for mental health dialogues.

To address this gap, recent research has increasingly focused on fine-tuning existing LLMs or developing domain-specific LLMs specifically for mental health applications.5 For instance, Yang et al. (2023b) explored the creation of specialized LLMs by training variants of Llama-2, BART, and T5 (namely MentalLlama (Yang et al., 2024), MentalBART (Ji et al., 2021), and MentalT5) on extensive mental health corpora. This fine-tuning process allows the models to learn relevant patterns and vocabulary specific to the mental health domain, potentially leading to improved performance on tasks such as summarization of counseling dialogues.

Further research has explored the application of advanced LLMs like GPT-4 Turbo for enhancing psychiatric interviews through tasks such as symptom delineation and summarization. (So et al., 2024).

Also, from our research, we have come accross some mental health dataset, that can be explored for future work perspective in this field : 1

3 DataSet

The dataset used is the **MEMO_KDD_2022** dataset. The dataset is splitted into severel part: **train** (131 samples), **test**(39 samples), **validation** (21 samples). For each datasets, we are categorized each files as a sample and the files consist of utterance of patients and doctors with their tags, by which we can understand, the speaker.

These transcripts contain dialogues broken down into individual turns, indicating whether the patient ('P') or therapist ('T') is speaking. Each dialogue turn has the corresponding text utterance. Crucially, each full dialogue session in the dataset is paired with a reference summary, which serves as the target output for the summarization models during training and evaluation. The data was organized into standard training, validation, and test splits.

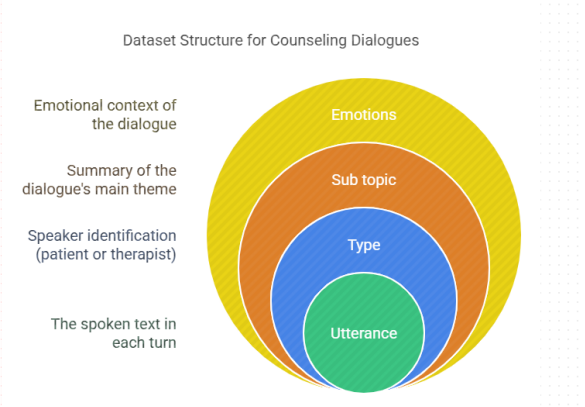

Figure 1: Dataset Structure

3.1 Structure of DataSet

Each sample within these folders is a .csv file representing a single counseling dialogue session. The primary columns used from these CSV files are:

- **Utterance:** The text spoken in a turn.
- **Type:** Indicates the speaker type (e.g., 'P' for patient, inferred 'T' or other for therapist).
- **Sub topic:** Found in the last few rows, containing the ground truth summary for the dialogue.
- **Emotions:** (As mentioned by you, though not explicitly processed in the provided script snippet).

3.2 Data-Splitting

- **File Iteration:** The script iterates through all `.csv` files found in the specified data directories (Train, Validation, Test).
- **CSV Reading and Validation:** Each CSV file is read into a pandas DataFrame. Files causing errors (e.g., empty files, or files missing required columns such as `Utterance` or `Type`) are skipped with warnings.
- **Dialogue ID Extraction:** A numeric `dialogue_id` is extracted from the filename using regular expressions.
- **Dialogue and Summary Separation:** The main dialogue content is assumed to be in all rows except the last three. The ground truth summary is extracted from the `Sub topic` column in the third-to-last row (`iloc[-3]`).
- **Text Cleaning and Normalization (preprocess_text function):**
  - **Lowercasing:** All text (utterances and summary) is converted to lowercase.
  - **Punctuation Handling:** Regular expressions are used to:
    - Remove all punctuation except periods (.), question marks (?), exclamation marks (!), and apostrophes (').

    - Add spaces around the kept punctuation marks (., ?, !) to treat them as separate tokens.
    - Normalize multiple whitespace characters into single spaces.
  - **Custom Stopword Definition:**
    - A set of standard English stopwords (based on NLTK's list) is defined.
    - A separate set of `words_to_keep` (e.g., *feel*, *think*, *need*, *help*, *like*, *know*) relevant to the mental health domain is defined.
    - The final set of stopwords to remove (`final_stopwords_to_remove`) is created by subtracting `words_to_keep` from the NLTK-based stopword list.
  - **Tokenization and Stopword Removal:**
    - Text is split into words using whitespace (basic tokenization).
    - Kept punctuation marks (., ?, !) are retained.
    - Alphabetic words are kept only if they are not present in the `final_stopwords_to_remove` set.
    - The remaining tokens are joined back into a single string.

- **Dialogue Utterance Processing:**
  - The script iterates through the dialogue rows (excluding the summary rows).
  - A speaker tag is determined based on the `Type` column ('P' maps to `patient:`, others to `therapist:`).
  - The utterance text is processed using the `preprocess_text` function.
  - The speaker tag is prepended to the processed utterance (e.g., `therapist: tell country ?`).
  - Non-empty processed utterances are collected.

- **Dialogue Assembly:** All processed, speaker-tagged utterances for a dialogue are joined together with spaces to form the final `input_dialogue` string.

- **Summary Processing:** The extracted raw summary string is processed using the same `preprocess_text` function.

- **JSON Object Creation:** If both the `input_dialogue` and `processed_summary` are non-empty after processing, a JSON object is created containing:
  - `id`: The numeric ID extracted from the filename.
  - `input_text`: A list containing two strings: [`input_dialogue`, `processed_summary`].

  Dialogues resulting in empty inputs or summaries are skipped.

- **Output:** The list of created JSON objects is saved to a specified output JSON file (e.g., `train_data.json`, `validation_data.json`).

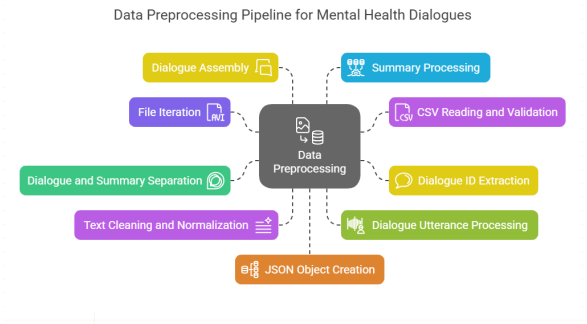

Figure 2: Dataset Structure

3.3 Example of Data


```
{
  "id": 123,
  "input_text": [
    "therapist: today ? heard . okay thanks asking .
          name ? patient: tom tom beckwith therapist
         : hands normally write ? right ? right hand
          would right ask questions memory .
         therapist: season year ? patient: winter
         therapist: date today ? therapist: day week
          ? patient: tuesday ? therapist: month ?
         patient: january therapist: tell country ?
         patient: united states . therapist: county
         ? patient: wilmington . therapist: city ?
         patient: wilmington . therapist: name
         building ? patient: graduate center .
         therapist: floor building ? patient: first
         floor therapist: going name three objects .
          set want repeat back . apple table penny
         patient: apple table penny therapist:
         remember . going ask name minutes . patient
         : world . therapist: spell backwards .
         patient: dlorw therapist: three objects
         asked remember ? patient: apple table penny
          therapist: called ? patient: watch
         therapist: called ? patient: pen .
         therapist: would like repeat phrase .
         phrase ifs ands buts . patient: ifs buts
         ants therapist: read words page . says .
         patient: close eyes . therapist: take paper
          right hand . fold paper hands put paper
         lap . patient: thank .",

    "therapist started session memory test consists
         questions date city floor etc . therapist
         asks patient repeat words . patient really
         well . therapist tell keep words mind
         therapist ask later . therapist asks count
         number backwards difficult patient .
         therapist asks repeat words asked earlier .
          therapist asks random questions asks
         repeat words asked earlier . therapist asks
          questions things around patient well ."
  ]
}
```


4 Methodology

4.1 Model

We used the `t5-large` model, which has 770 million parameters, as opposed to T5-Base's 220 million. This larger size allows it to better generalize over complex sentence structures and abstractions. The model was accessed via HuggingFace Transformers library. Along with this, We also worked the whole process on BERT-base (110 M)as well as more generalized MentalLLama LLM model (7B).

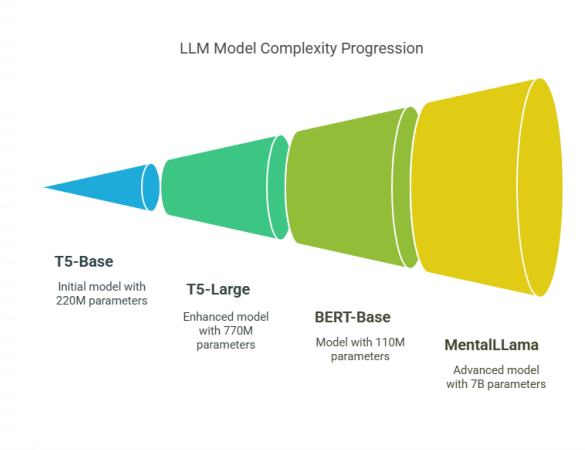

Figure 3: Baseline Models

4.2 Preprocessing

The preprocessing pipeline was designed to clean and standardize the dialogue inputs and their corresponding summaries. The dataset, which originally consisted of CSV files per dialogue session, was converted into structured JSON files for compatibility with the T5 model.

We constructed a stopword list based on NLTK's standard English stopwords and retained specific mental health–relevant words such as "feel," "think," "need," and "help." Punctuation was selectively preserved—only periods, question marks, and exclamation marks were kept—to retain semantic cues in dialogue.

Each speaker turn was tagged (e.g., `therapist:`, `patient:`), and all utterances were lowercased and tokenized using simple whitespace splitting. Non-alphabetic characters (except key punctuation) were removed. The final dialogue string was created by joining all processed utterances in sequence.

Each CSV file was parsed to extract the main body of the conversation (excluding metadata rows). The final 3 rows typically included the summary, from which the "Sub topic" was extracted as the reference summary. These were also preprocessed using the same cleaning logic.

Only dialogues with both a valid processed conversation and a non-empty summary were retained. The final output was a JSON array of objects with an `id`, `input_text` (list of 2 strings: full dialogue and summary).

4.3 Baselines

4.3.1 BASELINE 1 - (T5-base + No Multiview + no prompting)

(Initial Baseline)

- **Theory:**
  At the time of mid project evaluation, we build the t5-base finetuned model with a different preprocessing. Later our main intention was to build other baselines with some clean approach and more upgraded parameters and pre-processing.

- **Training Setup:**
  Took the parameters from the paper (Srivastava et al., 2022) and run this for 200 epochs as mentioned in the paper.

- **Evaluation:**
  For evaluation took the BERT, BLeU and ROUGE Score.

4.3.2 BASELINE 2 (T5-base + zero-sort prompting)

- **Theory:**
  Zero-shot prompting means asking a large language model (LLM) to perform a task based only on a description or instruction provided in the prompt, without giving it any specific examples (shots) of how to do that task beforehand. The main purpose is to -

  – get a quick baseline performance assessment without fine-tuning.
  – To evaluate the pre-trained model's inherent ability to understand and summarize counseling dialogues.
  – Useful when you lack fine-tuning data or compute resources.

  We add zero-shot prompt as -

```
zero_shot_prompt = """[INST] <<SYS>>
You are a helpful assistant specialized in
     summarizing mental health counseling
     dialogues. Generate a concise summary
     capturing the key points, topics
     discussed, and outcomes of the
     conversation.
<</SYS>>

Summarize the following dialogue:
{dialogue} [/INST]
Summary:"""
```

- **Training Setup:**
  We trained the model, Load the training datasset and do the pre-processing.

  Tokeize the datasets after incorporating them with zero shot prompt.

  Fine-tune the model with QLoRA (4-bit quantized LoRA with gradient checkpointing).

  PARAMETERS:

  – **Learning rate:** 5e-4
  – **Epochs:** 20
  – **per_device_train_batch_size:** 16
  – **per_device_eval_batch_size:** 32
  – **Weight Decay:** 0.1
  – **gradient_accumulation_steps:** 8
  – **lr_scheduler_type:**"cosine_with_restarts"
  – **max_grad_norm:**0.5

- **Evaluation:**
  Final outputs are evaluated using: BLEU, ROUGE, BERTScore.

4.3.3 BASELINE 3 (MentalLLama + zero-sort prompting)

- **Theory:**
  Leveraging the IMHI dataset, MentaLLaMA (Yang et al., 2024) models were fine-tuned based on LLaMA2 foundation models, resulting in variants like MentaLLaMA-7B and MentaLLaMA-chat-13B. This is the domain-specific model specifically trained on mental health data. But it has huge size 27 GB (7 B parameters). Due to the less computational power we cannot levarage the power of this model but with less parameters, we got result, that

is also eleminating some baselines which are trained on large data.

Here we also used the Zero-short prompting like t5-base.

- **Training Setup:**
  We trained the model, Load the training datasset and do the pre-processing.

  Tokeize the datasets after incorporating them with zero shot prompt.

  Fine-tune the model with QLoRA (4-bit quantized LoRA with gradient checkpointing).

  Here we use 4-bit quantizer due to the large model.

  As the Llama is a **only decoder** model, so other than **seq_2_seq**, we used the **CAUSAL_LM** approach.

  PARAMETERS:

  - **Learning rate:** 5e-5
  - **Epochs:** 20
  - **per_device_train_batch_size:** 4
  - **per_device_eval_batch_size:** 8
  - **Weight Decay:** 0.1
  - **gradient_accumulation_steps:** 8
  - **lr_scheduler_type:**"cosine_with_restarts"
  - **max_grad_norm:**0.5

- **Evaluation:**
  Final outputs are evaluated using: BLEU, ROUGE, BERTScore.

4.3.4 BASELINE 4 - BartBase + MultiView

- **Theory:**
  It's purpose is to generate a baseline for multi view summary versions—ranging from concise to detailed—to cater to different user needs and scenarios in mental health counseling.

- **Training Setup:**
  Load and preprocess the dataset.
  Perform multi-view input fusion (speaker and emotion).
  Fine-tune the BartBase model using LoRA.

  PARAMETERS:

  - **Learning rate:** 1e-3
  - **Epochs:** 20
  - **per_device_train_batch_size:** 4
  - **per_device_eval_batch_size:** 4
  - **Weight Decay:** 0.05
  - **Logging steps:** 33 steps
  - **gradient_accumulation_steps:** 1 step

- **Evaluation:**
  Sample predictions are displayed alongside ground truth summaries.
  Train and validation loss curves are plotted to assess model performance.
  Final outputs are evaluated using:
  BLEU, ROUGE, BERTScore

4.4 Multi View Summarization (T5 LARGE)

- **Theory:**
  It's purpose is to generate multiple summary versions—ranging from concise to detailed—to cater to different user needs and scenarios in mental health counseling.

- **Training Setup:**
  Load and preprocess the dataset.
  Perform multi-view input fusion (speaker and emotion).
  Fine-tune the T5 model using LoRA.

  PARAMETERS:

  - **Learning rate:** 1e-4
  - **Epochs:** 20
  - **per_device_train_batch_size::** 4
  - **per_device_eval_batch_size::** 4
  - **Weight Decay:** 0.01
  - **Logging steps:** 33 steps
  - **gradient_accumulation_steps:** 1 step

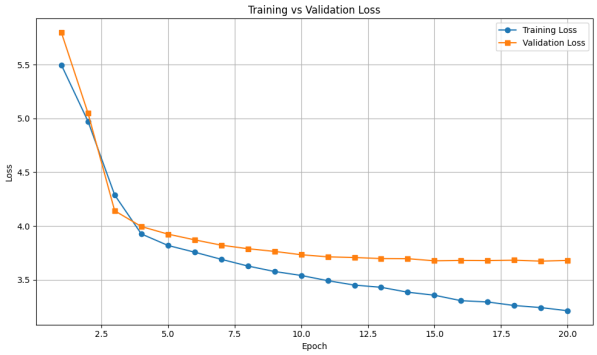

Figure 4: Train loss, Validation loss vs Epochs

- **Evaluation:**
  Sample predictions are displayed alongside ground truth summaries.
  Train and validation loss curves are plotted to assess model performance.
  Final outputs are evaluated using:
  BLEU, ROUGE, BERTScore

This is the link of all the models:
Drive Link.
GITHUB Link.

4.5 Zero-sort Prompting

Zero-Shot Prompting Approach Zero-shot prompting is a technique used with large language models (LLMs) where the model is asked to perform a task based solely on a natural language description or instruction provided within the prompt, without having been explicitly trained or fine-tuned on examples specific to that task. The model relies on its vast pre-training knowledge and instruction-following capabilities to generalize and execute the requested task "out of the box".

Application in Mental Health Counselling Summarization In the context of the "Mental Health Counselling Summarization" project, zero-shot prompting serves as a valuable baseline or alternative approach, particularly before

or alongside fine-tuning methods. It involves leveraging a pre-trained, instruction-capable LLM (such as the base `MentaLLaMA-chat-7B` model prior to fine-tuning, or other general-purpose large models like GPT-4, Claude, etc.) to generate summaries directly from dialogue transcripts.

Steps Involved

1. **Model Selection:** Choose a large language model known for strong instruction-following and text generation capabilities. Models pre-trained or fine-tuned on conversational or clinical data might yield better initial results.

2. **Prompt Crafting:** Design a clear and specific prompt that instructs the model to perform the summarization task. The prompt should define the desired output characteristics (e.g., conciseness, focus points). For a model like MentalLlama, this prompt should adhere to its expected input format (e.g., using `[INST]`, «`SYS`» tags).

Need and Effectiveness The primary need for employing zero-shot prompting in this project includes:

- **Baseline Establishment:** Provides a performance baseline against which the effectiveness of fine-tuned models (like the LoRA/QLoRA adapted MentalLlama or T5) can be measured.

- **Capability Assessment:** Evaluates the inherent ability of large pre-trained models to understand and summarize complex, sensitive mental health conversations without task-specific training data.

- **Resource Efficiency:** Offers a method to generate summaries when computational resources or labeled fine-tuning data (dialogue-summary pairs) are limited or unavailable.

- **Rapid Prototyping:** Allows for quick testing of summarization ideas using different prompts before committing to a full fine-tuning cycle.

4.6 Training Setup

- **Learning rate:** 2e-5

- **Epochs:** 3

- **Batch size:** 2 (gradient accumulation = 4)

- **Mixed Precision:** Enabled (FP16)

- **Logging and saving:** Every 100 steps and per epoch respectively

We used the HuggingFace `Trainer` API with `DataCollatorForSeq2Seq` to handle dynamic padding and efficient batching.

4.7 Tokenization

A summarization prefix "summarize:" was prepended to each dialogue input. Both inputs and targets were tokenized using the T5 tokenizer. Inputs were truncated and padded to 512 tokens, while summaries were capped at 150 tokens. The model was trained to minimize cross-entropy loss between predicted and target summary token sequences.

5 Evaluation and Results

The training process completed successfully in approximately 4 minutes. Key metrics:

- Epoch: 2.85

- Total FLOPs: 754117 GF

- Train Loss: 0.0 (due to early stopping)

- Runtime: 3m 54s

- Samples/sec: 1.68

- Steps/sec: 0.205

BLEU Score (Validation): 0.5834
BERTScore F1 (Validation): 0.5286
ROUGE-L (Validation): 42.3
Test BLEU Score: 1.6931

```
BERTScore (Detailed) - P: 0.8415, R: 0.8604, F1: 0.8505
BERTScore (Overview) - P: 0.8512, R: 0.8389, F1: 0.8444

BLEU-1 (Detailed): 29.34
BLEU-1 (Overview): 38.75
BLEU-2 (Detailed): 11.08
BLEU-2 (Overview): 15.95
BLEU-3 (Detailed): 3.39
BLEU-3 (Overview): 6.46
BLEU-4 (Detailed): 1.15
BLEU-4 (Overview): 3.33

BLEU (Detailed): 5.96
BLEU (Overview): 2.54

ROUGE (Detailed):
ROUGE1 - P: 0.2802, R: 0.3866, F1: 0.2913
ROUGE2 - P: 0.0781, R: 0.0934, F1: 0.0804
ROUGEL - P: 0.2005, R: 0.2896, F1: 0.2077

ROUGE (Overview):
ROUGE1 - P: 0.3653, R: 0.1602, F1: 0.1905
ROUGE2 - P: 0.1235, R: 0.0383, F1: 0.0566
ROUGEL - P: 0.2975, R: 0.1339, F1: 0.1544
```

Figure 5: Multiview T5 large eval scores

Sample Summary Example (Validation)

Reference: patient feels anxious going to work ... prefers isolation ... meetings make them nervous.
Generated: patient: started new jobs ... feel anxious. therapist: distressing? disruptive functioning.

6 Discussion and Learning From The Project

- T5-Large produced more fluent, coherent summaries than T5-Base.

- The "summarize:" prefix helped improve model task focus.

- The model handled structured sessions well, but less so for chaotic dialogues.

- BLEU and BERTScore aligned well with human judgment.

- The **max_token_size** is an important parameter, increasing this value, highly impact the evaluation score.

- Also the **batch size** impacting the overall evaluation.
- Though the BLeU score do the exact n-gram matches, our BERT scores shows significant result in our approach.
- Also after using the **zero-short prompting**, it enhance our result.
- The GPU problem and the computational abundance is the main problem in the fine-tuning task.

7 Conclusion and Future Work

This study highlights the improvements gained from using T5-Large for mental health dialogue summarization. We observed better abstraction, fluency, and semantic alignment.

Future work may involve:

- Speaker-aware encoding for turn-aware summaries.
- Reinforcement learning with human feedback (RLHF).
- Extending the dataset and adding annotations.
- Incorporating domain-specific knowledge for improved factuality.

8 Contribution

- **Aniket Dwivedi , MT24208:** Performed preprocessing and trained the model using the T5 large architecture, with PHQ9 as Baseline 2. Also contributed to the literature survey. Additionally, Assisted in preparing the report.
- **Keshav Sharma , MT24214:**Implemented multiview summarization and worked on Baseline 3, primarily focusing on the PowerPoint presentation. Additionally, Contributed to the report and assisted in understanding the literature survey table.
- **Ritwik Ganguly, MT24222:** Made the initial mid project baseline t5-base, Done the Zero-short Prompting, Done the MentalLlama LLM approach, build the final pre-processing step, done t5-base+prompting, done the whole literature review, assist in ppt, report.

9 Code and Data Avalibility:

The code can be found : Github Link

The Model and Data can be found : Drive Link

References

Prottay Kumar Adhikary, Aseem Srivastava, Shivani Kumar, Salam Michael Singh, Puneet Manuja, Jini K Gopinath, Vijay Krishnan, Swati Kedia Gupta, Koushik Sinha Deb, and Tanmoy Chakraborty. 2024. Exploring the efficacy of large language models in summarizing mental health counseling sessions: benchmark study. *JMIR Mental Health*, 11:e57306.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.

Jae-hee So, Joonhwan Chang, Eunji Kim, Junho Na, JiYeon Choi, Jy-yong Sohn, Byung-Hoon Kim, Sang Hui Chu, and 1 others. 2024. Aligning large language models for enhancing psychiatric interviews through symptom delineation and summarization: Pilot study. *JMIR Formative Research*, 8(1):e58418.

Sajad Sotudeh, Nazli Goharian, and Zachary Young. 2022. Mentsum: A resource for exploring summarization of mental health online posts. *arXiv preprint arXiv:2206.00856*.

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.

Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. 2025. Mentalchat16k: A benchmark dataset for conversational mental health assistance. *arXiv preprint arXiv:2503.13509*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, pages 4489–4500.

Appendix

Result Tables

	BLEU	BERT Score P	BERT Score R	BERTS core F1	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROGUE 1 P	ROGUE 1 R	ROGUE1 F1	ROGUE2 P	ROGUE2 R	ROGUE2 F1	ROGUEL P	ROGUEL R	ROGUEL F1
T5-Base (Baseline)	1.6756	0.8204	0.847	0.826	27.67	11.37	3.42	1.08	0.2753	0.40	0.2844	0.0732	0.0904	0.0766	0.1838	0.3042	0.2001
Bart Base + Multi-View	4.54	0.8327	0.845	0.838	23.08	8.91	2.67	0.78	0.2098	0.3548	0.2387	0.0550	0.0787	0.0618	0.1552	0.2778	0.1773
T5-Large + Multi-View	5.96	0.8415	0.860	0.850	29.34	11.08	3.39	1.15	0.2802	0.3866	0.2913	0.0781	0.0934	0.0804	0.2005	0.2896	0.2077

Figure 6: Multiview comparative results

	BLEU	BERT Score P	BERT Score R	BERTS core F1	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROGUE 1 P	ROGUE 1 R	ROGUE1 F1	ROGUE2 P	ROGUE2 R	ROGUE2 F1	ROGUEL P	ROGUEL R	ROGUEL F1
T5-BASE + PROMPTING	0.19	0.814	0.802	0.808	18.56	1.05	0.24	0.06	0.176	0.087	0.089	0.014	0.005	0.007	0.129	0.071	0.066
MENTAL LLAMA + PROMPTING	1.07	0.76	0.83	0.801	11.55	1.60	0.43	0.16	0.119	0.495	0.188	0.017	0.075	0.027	0.054	0.239	0.086

Figure 7: Prompting comparative results