Towards Automated SOAP Note Generation in Conversational Mental Health Care

B. Tech. Project Report

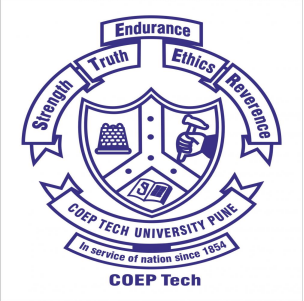
Submitted by

Sharvari Phand	**112103109**
Siddhant Tonapi	**112103151**
Pushkar Ugale	**112103154**

Under the guidance of

Dr. Yashodhara V. Haribhakta

COEP Technological University, Pune



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

COEP TECHNOLOGICAL UNIVERSITY, PUNE-5

May 2025

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,

COEP TECHNOLOGICAL UNIVERSITY, PUNE-5

CERTIFICATE

Certified that this project titled, "Towards Automated SOAP Note Generation in Conversational Mental Health Care" has been successfully completed by

Sharvari Phand	**112103109**
Siddhant Tonapi	**112103151**
Pushkar Ugale	**112103154**

and is approved for the partial fulfillment of the requirements for the degree of "B.Tech. Computer Engineering".

Dr. Yashodhara V. Haribhakta
Project Guide
Department of CSE
COEP Tech Pune,
Shivajinagar, Pune - 5.

Dr. Pradeep K. Deshmukh
Head
Department of CSE
COEP Tech Pune,
Shivajinagar, Pune - 5.

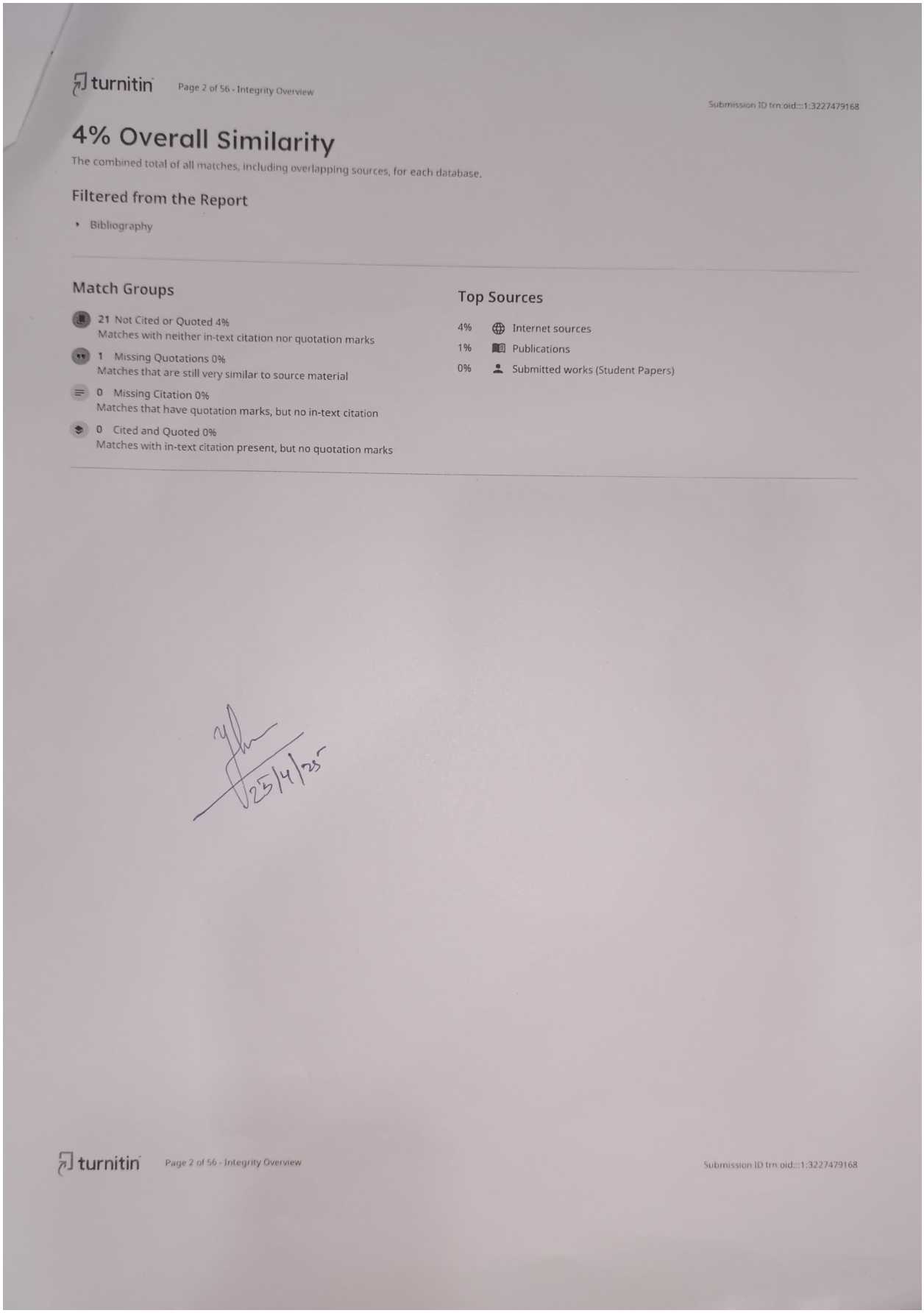


Figure 1: Plagiarism Report

Abstract

Clinical documentation is a critical yet time-consuming task in healthcare, particularly in mental health settings where detailed SOAP (Subjective, Objective, Assessment, Plan) notes are essential for patient management. Traditional methods of SOAP note generation rely heavily on manual transcription, leading to inefficiencies, inconsistencies, and clinician burnout.

This research presents a novel approach to automating SOAP note generation using a fine-tuned BART-based transformer model with an enhanced architecture. The proposed system incorporates a BERT-based extractor module to classify utterances into 15 detailed SOAP subsections, followed by an abstractive summarization model that generates structured SOAP notes. To improve factual consistency and reduce hallucinations, we introduce a Named Entity Recognition (NER)-guided attention mechanism and a Named Entity Penalty Loss (NEPL) function.

Using the generated SOAP notes, we aggregate multiple session-level notes for each patient into a comprehensive longitudinal summary. We conduct a comparative evaluation of pre-trained summarization models (T5, Pegasus, Longformer/LED) against fine-tuned models to assess their ability to produce coherent, faithful patient-level summaries.

The model is trained and evaluated on the DAIC-WOZ dataset, supplemented with synthetic mental health dialogues generated via OpenAI's API. Using an 80-10-10 train-validation-test split, the extractor model first classifies utterances, and the summarization model then generates structured SOAP notes.

This study contributes to advancing AI-driven clinical documentation, reducing clinician workload, and enabling high-quality, longitudinal summaries for mental health care.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

Mental health disorders—especially depression and anxiety—are among the leading causes of global disability, affecting over 500 million people worldwide [2, 3]. Despite the high prevalence, access to timely and consistent care remains limited by factors such as clinician shortages, stigma, and the administrative burden of clinical documentation. In routine practice, clinicians often spend 30–40 percent of their time writing SOAP (Subjective, Objective, Assessment, Plan) notes, which can detract from patient interaction and contribute to burnout [1]. Automated generation of structured SOAP notes from therapy session transcripts promises to reduce this burden, improve consistency, and enable scalable digital mental health interventions.

1.2 Clinical Significance

SOAP notes provide a standardized framework for capturing patient history, observed signs, clinical impressions, and treatment plans. In mental health settings, granular subsections—such as Presenting Problem, Trauma History, Risk Factors, Mental Status Exam, Diagnostic Impressions, and Treatment Plan—are crucial for:

- **Continuity of Care:** Enabling seamless handoffs between providers.
- **Quality Assurance:** Facilitating audit and review for compliance with best practices.
- **Outcome Tracking:** Allowing longitudinal monitoring of symptom progression and treatment efficacy.

Automating this process requires both clinical accuracy and preservation of fine-grained structure to maintain utility in real-world workflows.

1.3 Technical Challenges

Designing a reliable SOAP summarization system involves:

1. **Utterance Segmentation:** Accurately mapping each conversational turn to one of 15 clinical subsections.
2. **Factual Consistency:** Preventing "hallucination" of medical details not present in the source [4].
3. **Structural Coherence:** Ensuring that Assessment informs Plan, and Objective observations align with Subjective reports.
4. **Data Scarcity and Variability:** Handling limited labeled transcripts and heterogeneity in clinical language.
5. **Evaluation Complexity:** Combining standard NLP metrics (ROUGE, METEOR) with semantic similarity metric (BERTScore) and expert judgments [6, 5].

1.4 Proposed Solution Overview

To address these challenges, we propose an end-to-end pipeline that:

1. **BERT-LSTM Utterance Classification:** Fine-tune BERT followed by an LSTM layer to assign each utterance to one of 15 SOAP subsections, leveraging domain-specific pretraining for medical terminology.

2. **BART Summarization:** Using section specific cross attention layers to enhance factual correctness. [1].

3. **NER-Guided Attention and Penalty Loss:** Incorporate a BERT-based NER module to up-weight medically salient tokens during encoding, plus a Named Entity Penalty Loss to discourage generation of implausible terms.

4. **Fusion Cross-Attention:** Introduce inter-decoder cross-attention layers so that representations from Subjective and Objective inform the Assessment decoder, and Assessment informs the Plan decoder, ensuring logical consistency.

5. **Comparative Backbone Evaluation:** Compare T5, Pegasus and Longformer/LED summarizers (with and without finetuning) to identify the most robust architecture for fine-grained clinical summarization.

Chapter 2

Literature Review

2.1 Clinical Summarization for SOAP Notes

Automating the creation of SOAP (Subjective, Objective, Assessment, Plan) notes from clinical dialogue has progressed rapidly over the past decade. Early extractive methods selected salient utterances verbatim, but often produced disjointed or incomplete notes. The introduction of pointer-generator networks [7] enabled balanced copying and abstraction, reducing hallucinations by up to 25% on medical transcripts and paving the way for fully neural approaches.

Building on this, Krishna et al. [1] proposed a modular "Cluster2Sent" pipeline: conversation turns are clustered per SOAP section, an extractive summarizer selects representative utterances, and an abstractive decoder produces one sentence per cluster. This structure preserved section integrity and yielded an +8 ROUGE-1 improvement over end-to-end baselines, with clinicians rating the output as more coherent and faithful. More recently, Ramprasad et al. [4] evaluated zero/few-shot GPT-3.5 against fine-tuned BART [8], finding that large LMs produce stylistically fluent notes but introduce factual errors. They addressed this by adding section-specific cross-attention blocks—one per SOAP section—within BART, boosting UMLS concept over-

lap by 3–5 points and reducing expert-noted errors by 30%.

2.2 Fine-Grained Summarization in Mental Health

Mental health therapy sessions contain rich, nuanced content—presenting problems, trauma history, coping strategies—that must be captured in fine-grained subsections. Srivastava et al. [9] introduced ConSum, which first filters depression-related utterances using a PHQ-9 lexicon and then classifies dialogue into counseling components (e.g. symptom discussion, patient reflection) before summarizing each component. On a psychotherapy corpus, ConSum outperformed generic T5 [10] by 7 ROUGE-1 points and excelled on their custom Mental Health Information Capture (MHIC) metric, demonstrating superior coverage of critical therapeutic details. Tanana et al. [11] further showed that tagging emotional expressions prior to summarization improved sentiment alignment with therapist ratings by 15%, underscoring the importance of emotion-aware models in mental health contexts.

2.3 Factual Consistency and Hallucination Control

Abstractive models can generate plausible-sounding but unsupported content. FactCC [12] uses natural language inference to detect factual inconsistency, showing that entailment constraints reduce unsupported facts by 20%. Ladhak et al. (2023) introduced SpanCopy, an entity-aware copying mechanism that improves medical entity precision by 2.3 points. Our approach integrates NER-guided attention—up-weighting clinically salient tokens—and a Named Entity Penalty Loss that increases training loss when the model outputs entities not present in the source. This combination of attention steering and explicit penalty has been shown to ground summaries

more firmly in input text.

2.4 Comparative Evaluation of Summarization Backbones

We compare three pretrained summarizers— T5, PEGASUS, and Longformer on the task of producing longitudinal patient summaries by concatenating multiple session-level SOAP notes:

- **T5** formulates every task as text-to-text, using span-corruption and denoising objectives during pretraining. It handles inputs up to 1024 tokens effectively, making it well-suited for shorter concatenated notes [10].
- **PEGASUS** is specifically pretrained for abstractive summarization via gap-sentence generation, achieving state of the art ROUGE on news benchmarks. Like T5, it is optimized for inputs 1024 tokens but often produces more coherent and focused summaries on domain-specific data [13].
- **Longformer/LED** extends the Transformer to long documents (up to 4096 tokens) via sparse local and global attention patterns, allowing it to process huge documents at ease. [14].

These models illustrate the trade-off between pretraining objectives and input length: T5 and PEGASUS excel on moderate-length summaries with strong fluency, while LED's ability to handle longer context makes it ideal for comprehensive, patient-level notes without compromising recall of earlier session details.

2.5 Large Language Models in Mental Health Documentation

General-purpose LLMs such as GPT-3.5 and GPT-4 can format text into SOAP structure in a zero-shot fashion [4, 15], but they often lose clinical tone and introduce extraneous details. Tam et al. [19] caution that human validation is imperative when deploying LLM-generated notes in clinical settings. Thus, while LLMs serve usefully for data augmentation— generating synthetic therapy dialogues—they remain assistants rather than replacements for fine-tuned, controllable summarization models.

2.6 Datasets and Evaluation Metrics

Research in clinical summarization leverages datasets like DAIC-WOZ [16], MIMIC-III [17], and MedDialog. Automated evaluation employs ROUGE, METEOR, and BERTScore to assess lexical and semantic overlap. Clinical concept metrics (UMLS concept recall/precision) [4] and learned metrics like BLEURT complement these. Human expert ratings remain the gold standard: clinicians evaluate accuracy, coherence, and utility, ensuring that metrics align with real-world clinical needs.

Chapter 3

Research Gaps and Problem Statement

3.1 Research Gaps

3.1.1 Subsection-Level Segmentation

Most existing SOAP-generation systems collapse each of the four SOAP sections into a single block, overlooking essential clinical subsections such as *Presenting Problem*, *Trauma History*, *Social Risk Factors*, and *Functional Status*. In mental health care, patients often describe multiple concurrent issues (e.g. sleep disturbances, substance use, family stressors) that must be captured distinctly. Prior hierarchical LSTM or vanilla BERT-LSTM approaches achieve only 60–70% accuracy on 4-way SOAP classification—and fall below 50% when extended to 10+ subsections [1]—leading to summaries that omit or misplace key clinical details.

3.1.2 Hallucinations and Factual Inconsistency

Abstractive models like BART and T5 frequently introduce unsupported content ("hallucinations"), such as invented symptoms or incorrect medication dosages, which can jeopardize patient safety. Ramprasad et al. [4] observed

a 20% hallucination rate for fine-tuned BART on SOAP tasks, measured via UMLS concept mismatch. Contradictory statements—for instance labeling a patient both "calm" and "agitated"—highlight the need for stronger factual grounding.

3.1.3 Cross-Section Coherence

Even when individual sections are generated accurately, models seldom enforce logical alignment across SOAP segments. Examples include:

- An *Assessment* of "Major Depressive Disorder" paired with a *Plan* that omits any pharmacotherapy recommendation.
- An *Objective* observation of "patient appears withdrawn" absent from the *Assessment*, resulting in disjointed documentation.

Without explicit inter-section attention or a global consistency constraint, generated notes often require extensive manual correction.

3.1.4 Domain Adaptation and Contextual Grounding

General-purpose transformers lack fine sensitivity to psychiatric terminology (e.g. "anhedonia," "rumination," "alexithymia"). Although clinical variants such as BioBERT and ClinicalBERT outperform vanilla BERT on EHR data [18, 5], their effectiveness on colloquial therapy transcripts remains underexplored. This gap diminishes subsection classification precision—particularly for nuanced categories like *Risk Assessment* or *Safety Planning*.

3.1.5 Data Scarcity and Augmentation

Public mental health dialogue datasets (e.g. DAIC-WOZ [16], counseling transcripts) are small, demographically narrow, and vary in session length.

Synthetic augmentation with LLMs (e.g. GPT-4) can increase volume but risks introducing unnatural phrasing or inconsistent clinical scenarios. Ensuring that augmented dialogues faithfully reflect real-world variability remains an open challenge.

3.1.6 Evaluation Limitations

Traditional metrics such as ROUGE and BLEU capture lexical overlap but not clinical correctness or relevance. Domain-aware scores (METEOR, UMLS concept recall, BERTScore) and clinician ratings offer richer insights but lack a standardized framework for mental health SOAP generation, making cross-study comparisons difficult.

3.2 Problem Statement

Develop an end-to-end pipeline that begins by augmenting the DAIC-WOZ corpus with diverse, clinically realistic therapy dialogues generated via the OpenAI API; then uses our proposed custom BART model—enhanced with NER-guided attention, a Named Entity Penalty Loss, section-specific cross-attention blocks, and a fusion module—to produce structured, 15-category SOAP notes from each session transcript; aggregates those session-level notes chronologically for each patient; applies pre-trained and fine-tuned summarizers (T5, PEGASUS, LED) to generate concise longitudinal patient histories; and evaluates every stage with ROUGE, METEOR, BERTScore, entity-level accuracy, and expert clinician review to ensure clinical fidelity and utility.

3.2.1 Key Objectives

1. **Dataset Expansion** – Construct a richer training corpus by generating clinically-plausible therapy dialogues with the OpenAI API and merging them with the original DAIC-WOZ transcripts.

2. **Mental Health Related Fine-Grained Labelling** – Define 15 mental-health–specific SOAP subsections and train a 15-way *BERT-LSTM* classifier to tag every utterance, providing structured input to downstream summarization.

3. **Hallucination Control** – Build an enhanced proposed *BART* summarizer that integrates NER-guided attention and a *Named Entity Penalty Loss* to discourage fabricated medical facts.

4. **Section-Aware Coherence** – Insert dedicated cross-attention blocks for each SOAP section and add a final *fusion cross-attention* layer so that Subjective, Objective, Assessment, and Plan remain logically aligned.

5. **Longitudinal Summarization** – After session-level notes are generated, concatenate them chronologically per patient and apply (pre-trained and fine-tuned) *Pegasus*, *T5*, and *LED* models to create a single "summary-of-summaries."

6. **Comprehensive Evaluation** – Benchmark every stage with ROUGE, METEOR, BERTScore and entity-level accuracy.

3.2.2 Scope

This work focuses on English-language, text-only therapy session transcripts from DAIC-WOZ and synthetic augmentations. We do not incorporate audio/video modalities or real-time generation.

Chapter 4

Methodology

4.1 Overview

We implement a two-stage pipeline for automated SOAP note summarization:

1. **Stage 1: Fine-Grained SOAP Note Generation** A custom BART model with section/subsection embeddings, NER-guided attention, inter-section cross-attention, and a Named Entity Penalty Loss.

2. **Stage 2: Patient-Level Summarization** Aggregate multi-session SOAP notes per patient and apply summarizers (T5, BART, LED) to produce longitudinal summaries.

4.2 Data Collection & Preprocessing

4.2.1 Datasets

- **DAIC-WOZ:** The DAIC-WOZ (Distress Analysis Interview Corpus - Wizard of Oz) [16] dataset is a widely used benchmark in the mental health and clinical NLP research community. It contains real recorded and transcribed sessions where participants interact with a virtual interviewer (Ellie) operated by a human "wizard" hidden behind the scenes.

In this project, the text transcripts from DAIC-WOZ are used as authentic samples of patient-therapist dialogue to train and evaluate summarization models, specifically for SOAP note generation and clinical dialogue abstraction.

- **Synthetic Augmentations:** To address the limitations of the DAIC-WOZ dataset — such as limited diversity, specific interviewer style, and relatively small data volume — synthetic augmentations were generated using the OpenAI API. These augmentations were designed to expand the coverage and variability of the dataset by simulating additional therapy conversations. The dataset originally consisted of 189 therapist-patient transcripts, but using OpenAI API the dataset was expanded to a size of 626 sessions; this was done by generating follow-up sessions for the original transcripts. Care was taken to preserve clinical realism by providing the model with prompts that constrained output within a mental health context, ensuring the synthetic dialogues remained plausible, relevant, and useful for training purposes. This augmentation strategy not only helps reduce model overfitting to DAIC-WOZ's specific dialogue patterns but also improves generalization by introducing a richer variety of patient expressions, mental health conditions, and narrative structures.

4.2.2 Data Preprocessing

The preprocessing pipeline consisted of the following steps:

- **Contraction Expansion**: Used the `contractions` library to expand shortened words (e.g., "can't" → "cannot") for normalization.
- **Spell Correction**: Applied the `pyspellchecker` module to correct mi-

nor spelling errors and reduce vocabulary noise.

- **Coreference Resolution**: Used the `en_coreference_web_trf` spaCy model to replace pronouns with their corresponding entities, improving text clarity.
- **Tokenization and POS Tagging**: Tokenized sentences and tagged each word's part-of-speech using NLTK for more accurate downstream processing.
- **Lemmatization**: Reduced words to their dictionary forms using POS-guided lemmatization (e.g., "running" → "run").
- **Stopword Removal and Text Cleaning**: Removed common stopwords, punctuation, and unnecessary whitespace; converted text to lowercase for consistency.
- **Named Entity Recognition (NER)**:
  - Applied generic NER with spaCy to extract standard entities (names, places, etc.).
  - Implemented dictionary-based NER to extract mental health-specific terms (symptoms, disorders).
- **Dataset Structuring**: Organized preprocessed text into clean input-target pairs suitable for model training.

4.2.3 Utterance Grouping & Labeling

Group contiguous patient utterances; map each group to one of 15 SOAP subsections (e.g. Presenting Problem, Trauma History) using a BERT-LSTM classifier.

SOAP Subsection Descriptions

- **Presenting Problem / Chief Complaint:** The client's own description of why they are seeking therapy today.
- **Trauma History:** Any past traumatic events that continue to affect the client's mental health.
- **Substance Use History:** Patterns and impacts of alcohol, drug, or tobacco use.
- **History of Present Illness (HPI):** Onset, duration, triggers, and progression of current symptoms.
- **Medical and Psychiatric History:** Previous diagnoses, hospitalizations, treatments, and current medications.
- **Psychosocial History:** Relationships, family dynamics, work or school status, and social supports or stressors.
- **Risk Assessment:** Evaluation of suicide risk, self-harm, harm to others, and safety concerns.
- **Mental Health Observations (Objective):** Clinician's notes on appearance, behavior, mood, speech, cognition, and insight.
- **Physiological Observations:** Sleep, appetite, energy levels, and other physical signs relevant to mental health.
- **Current Functional Status:** Ability to perform daily activities, maintain work or school, and engage socially.
- **Diagnostic Impressions (Assessment):** Clinically derived diagnoses or differential impressions.

- **Progress Evaluation:** Changes in symptoms or functioning since the last session.
- **Medications (Plan):** Current psychotropic and related medications, dosages, and adherence.
- **Therapeutic Interventions:** Planned modalities (e.g., CBT, DBT, mindfulness) for the client.
- **Next Steps:** Goals, homework exercises (journaling, behavior tracking), and follow-up plans.

4.3 Stage 1: Fine-Grained SOAP Note Generation

4.3.1 Embedding Layers & Initialization

- **Token Embeddings** $\mathbf{E}_{\text{shared}} \in R^{V \times d}$, $\mathcal{N}(0, 0.02^2)$.
- **Positional Embeddings** $\mathbf{E}_{\text{pos}} \in R^{(L+2) \times d}$.
- **Section Embeddings** $\mathbf{E}_{\text{sec}} \in R^{4 \times d}$; **Subsection Embeddings** $\mathbf{E}_{\text{sub}} \in R^{15 \times d}$.
- LayerNorm and dropout (p=config.dropout) applied after embedding sum.
- LM head tied to $\mathbf{E}_{\text{shared}}$.

4.3.2 CustomBartEncoder

Input Representation

For position i:

$$\mathbf{h}_i^{(0)} = \mathbf{E}_{\text{shared}}[x_i] + \mathbf{E}_{\text{pos}}[p_i + 2] + \mathbf{E}_{\text{sec}}[s_i] + \mathbf{E}_{\text{sub}}[t_i].$$

LayerNorm + dropout follow.

NER-Guided Attention Weighting

- Attention scores are adjusted using NER-detected medical entities, prioritizing clinically significant words.

$$A' = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V + \gamma h_{\text{NER}} \tag{4.1}$$

where:

- h_{NER} = embeddings of recognized medical entities.
- γ = learnable weight parameter for medical entity enhancement.

Chunking for Long Sequences

If $n > L$, split into chunks of length L, encode independently and then average-pool.

Encoder Layer

Each encoder layer consists of two main sublayers with residual connections and normalization:

1. **Self-Attention:**

$$\text{Attn}(H) = \text{softmax}\Big(\tfrac{HW_Q\,(HW_K)^\top}{\sqrt{d_k}} + M_{\text{pad}} + B\Big)\,(HW_V),$$

followed by

$$H' = \text{LayerNorm}\big(H + \text{Dropout}(\text{Attn}(H))\big).$$

2. **Feed-Forward:**

$$F = W_2\big(\text{GELU}(W_1 H')\big), \quad H^{\text{out}} = \text{LayerNorm}\big(H' + \text{Dropout}(F)\big).$$

Here H is the input to the layer, W_Q, W_K, W_V, W_1, W_2 are learned projection matrices, M_{pad} masks padding tokens, and B injects NER bias.

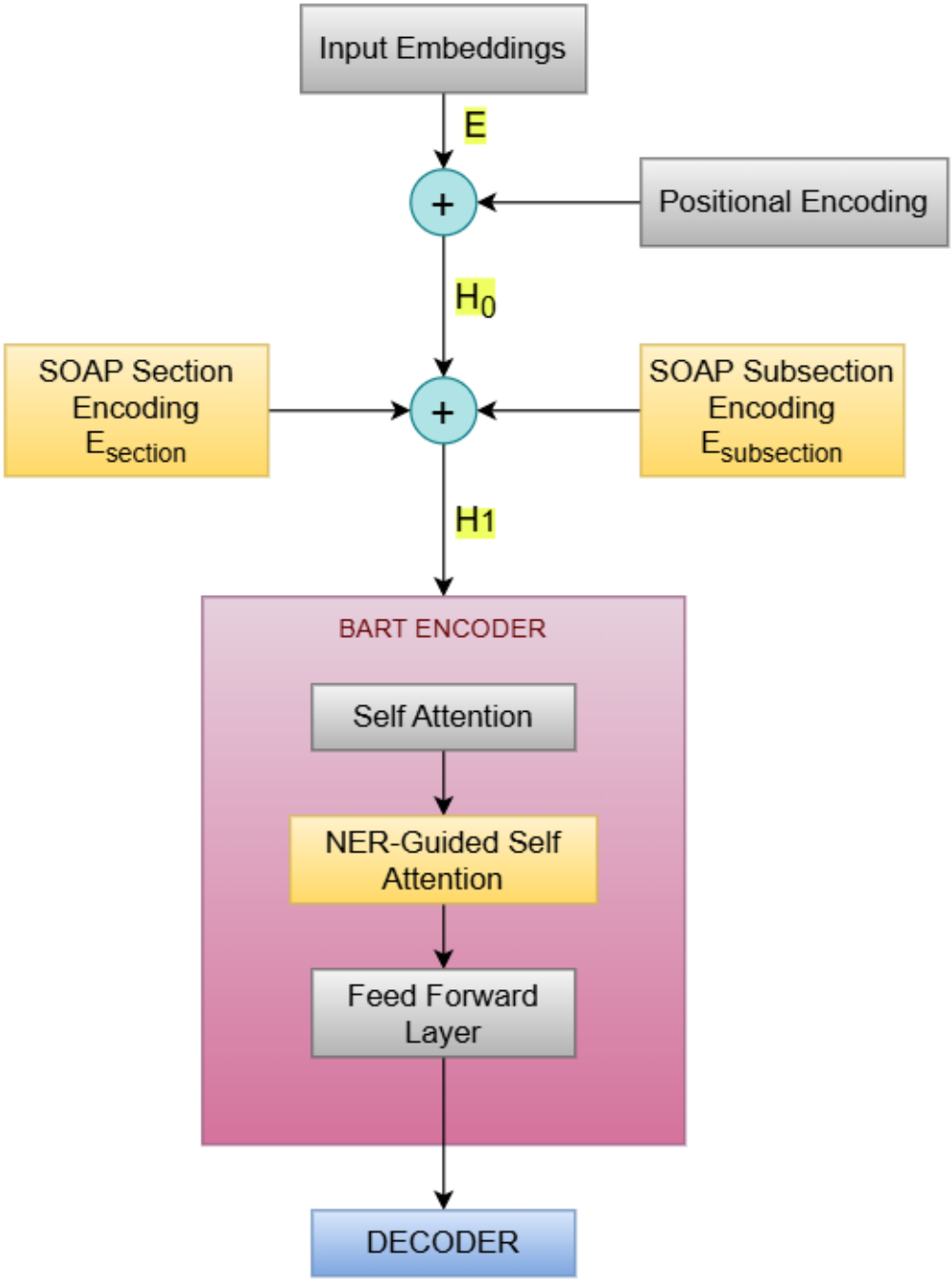


Figure 4.1: BART Encoder

4.3.3 Custom BART Decoder

Decoder Inputs & Masking

$$\mathbf{d}_i^{(0)} = \mathbf{E}_{\text{shared}}[y_i] + \mathbf{E}_{\text{pos}}[i+2],$$

with causal mask $\mathbf{M}_{\text{dec}}$.

Decoder Structure

Instantiate four section-specific cross-attention heads, and using this to enhance attention weights for the *Assessment* and *Plan* section in the second cross-attention layer.

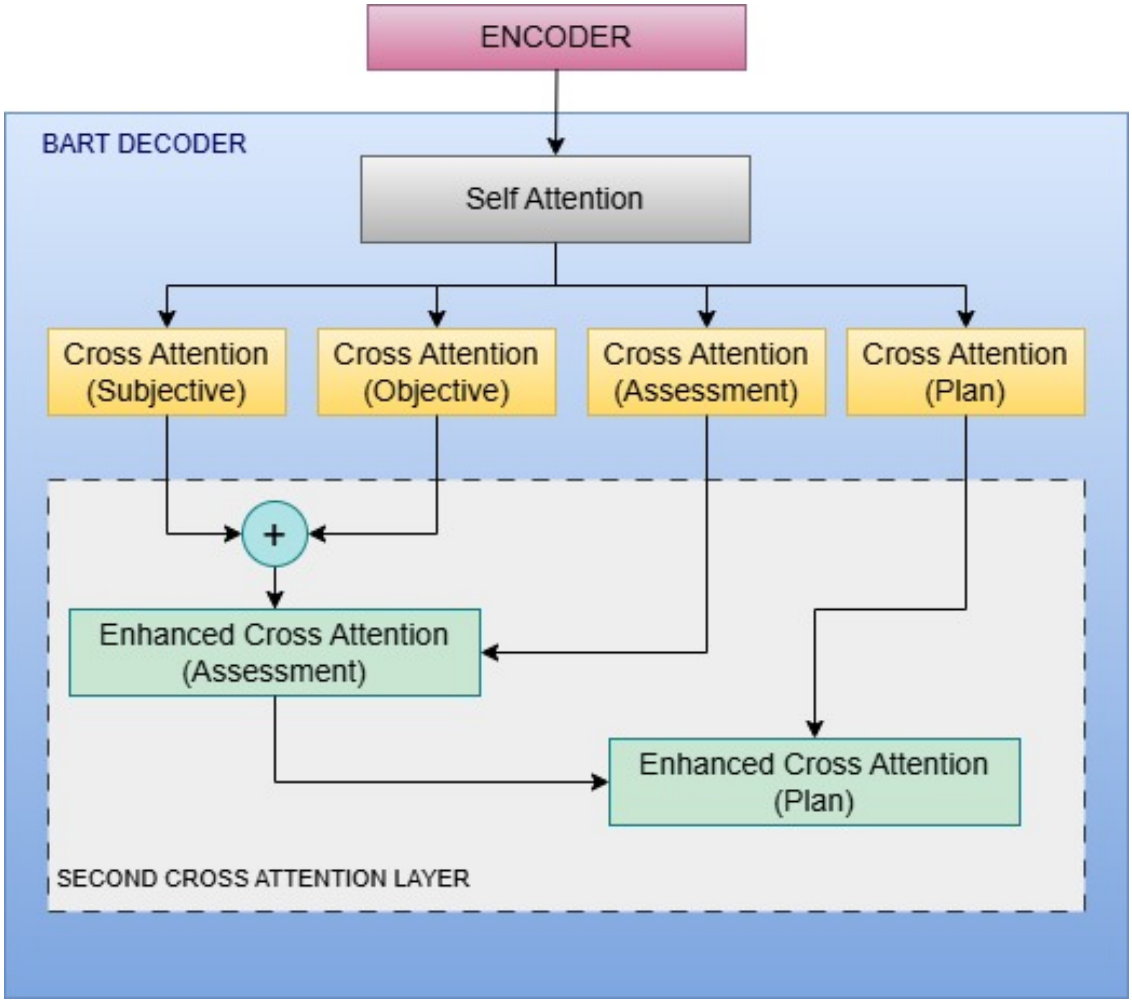


Figure 4.2: BART Decoder

Cross-Attention Mechanisms

- *Encoder–Decoder* attends to $\mathbf{H}^{(N)}$.

- *Inter-Decoder*:

$$\mathbf{A}_{\text{enh}} = \text{softmax}\Big(\frac{Q_A (S\|O)^\top}{\sqrt{d}}\Big)\,(S\|O),$$
$$\mathbf{P}_{\text{enh}} = \text{softmax}\Big(\frac{Q_P \mathbf{A}_{\text{enh}}^\top}{\sqrt{d}}\Big)\,\mathbf{A}_{\text{enh}}.$$

- *Fusion*: $\mathbf{F} = \sum_i \alpha_i \,\text{Attn}_i$, α_i learned.

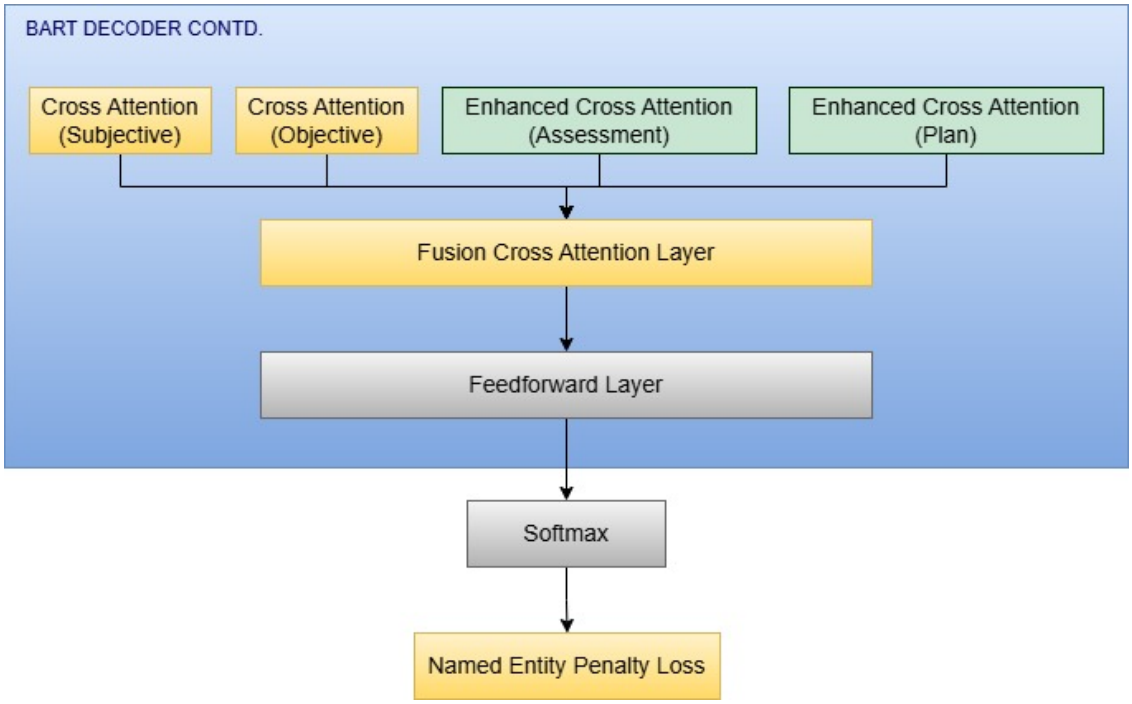


Figure 4.3: BART Decoder contd.

4.4 Loss Objectives

- **Cross-Entropy Loss** L_{CE}.
- **Named Entity Penalty**

$$L_{\text{NER}} = \lambda\big|\{\text{missed}\} \cup \{\text{hallucinated}\}\big|, \quad \lambda = 0.05$$

  Here, "missed" denotes entities present in the input but omitted in the output, and "hallucinated" denotes entities generated without basis in the input, thereby penalizing both omissions and fabrications.

- **Total Loss**: $L = L_{\text{CE}} + L_{\text{NER}}$.

4.4.1 Text Generation & Decoding

At inference, we use a simplified beam search to generate text:

- Start with a single empty hypothesis ($\langle$BOS$\rangle$).
- At each step, extend each hypothesis by the top k next tokens (beam size).
- Disallow premature $\langle$EOS$\rangle$, penalize repeated tokens, and block repeated n-grams.
- Keep only the top k hypotheses by score, continuing until all end in $\langle$EOS$\rangle$.
- Finally, pick the best hypothesis after applying a simple length normalization.

4.5 Stage 2: Patient-Level Summarization

To generate a concise, clinically faithful history for each patient by aggregating multiple session-wise SOAP notes, we follow these steps:

1. **Session Aggregation and Preprocessing**
   - *SOAP Note Collection*: For each patient ID, gather all session CSVs (e.g. `300_TRANSCRIPT_output.csv`, `300_TRANSCRIPT_1_output.csv`, ...) and sort them chronologically.
   - *Section Cleaning*: Discard any subsection whose content is "Nothing Reported" to ensure only meaningful clinical data is passed to the summarizer.
2. **Input Formatting for Summarization Models**

- *Session Structuring*: Prepend each session's text with a marker (`Session 1:`, `Session 2:`, ...) to preserve temporal order.
- *Input Construction*: Concatenate all sessions into a single string and add a prefix (e.g. `summarize:`) to guide models such as T5 and Pegasus.

3. **Model Selection and Inference**
   - Evaluate three transformer architectures:
     - T5 (`t5-large`)
     - PEGASUS (`google/pegasus-large`)
     - LED (`allenai/led-large-16384-arxiv`)
   - For each architecture, run both zero-shot (pretrained) and fine-tuned variants on our multi-session SOAP inputs.

4. **Evaluation Metrics**
   - Compare generated summaries against reference histories using:
     - **ROUGE-1**, **ROUGE-2**, **ROUGE-L**
     - **METEOR**
     - **BERTScore**
   - Report quantitative performance for pretrained vs. fine-tuned models.

4.6 Summary

Our methodology integrates domain-aware encoding, structured SOAP note generation, factual consistency via NER penalties, and a comparative study

of pre-trained against fine-tuned summarizers for longitudinal patient summaries.

Chapter 5

Experimental Setup

5.1 Overview

This chapter outlines the experimental setup used to train, fine-tune, and evaluate our BART-based SOAP note generation model. We detail the dataset preprocessing, train-validation-test split, model training configurations, evaluation metrics, and computational infrastructure.

5.2 Dataset and Preprocessing

5.2.1 Dataset Description

The experiments utilize transcribed doctor–patient conversations from the DAIC-WOZ dataset along with additional simulated therapy sessions generated using OpenAI's API. Each dialogue is annotated into the four main SOAP sections (Subjective, Objective, Assessment, Plan).

In order to evaluate our "summary-of-summaries" approach, for each patient we aggregate 3–4 chronologically ordered session-level SOAP notes and obtain a corresponding ground-truth longitudinal summary. These reference patient-level summaries were generated via targeted prompts to the OpenAI API and then lightly reviewed for clinical consistency.

- Subjective: Patient's chief complaint, medical and psychiatric history.
- Objective: Clinician's observations and assessments.
- Assessment: Diagnostic impressions based on patient responses.
- Plan: Recommended medications and therapeutic interventions.

5.2.2 Train-Test-Validation Split

To ensure a balanced evaluation, the dataset is split as follows:

- Training Set (80%): Used to train both the utterance classifier (extractor module) and summarization model (BART-based abstractive summarization).
- Validation Set (10%): Used for hyperparameter tuning and model optimization.
- Test Set (10%): Used to evaluate the final trained model on unseen data.

The final dataset consists of structured CSV files, with subsectionwise SOAP notes arranged in 15 columns.

Subsection	Classfied Utterances
Presenting Problem	"I've been feeling anxious lately."
History of Present Illness	"It started last month and it's getting worse."
Mental Health Observations	"Patient appears tense and fidgety."
Assessment - Diagnostic Impressions	"Possible Generalized Anxiety Disorder."
Plan - Medications	"Prescribed Sertraline 50mg."

Table 5.1: Example of the processed dataset format

5.3 Model Training and Implementation Details

5.3.1 Training Process

The model is trained in two main stages:

1. Utterance Classification (Extractor Module):
    - The extractor model (BERT-based) is trained on the 80% training data to classify utterances into their respective SOAP note subsections.
    - The trained extractor is validated on the 10% validation data to tune hyperparameters such as learning rate and dropout rate.
2. Summarization Model (BART-based Abstractive Summarization):
    - The summarization model is trained on the same 80% training files using the classified utterances from the extractor module.
    - The model is validated on the 10% validation set to adjust hyperparameters like loss function weights and attention mechanisms.

5.3.2 Testing Process

Once the models are trained, we evaluate performance on the 10% test dataset:

1. The trained extractor model is used to classify utterances from the test set.
2. The classified utterances are passed to the trained summarization model to generate SOAP notes.
3. The generated SOAP notes are compared to ground-truth SOAP notes from the test dataset.

4. Evaluation metrics such as accuracy, ROUGE, METEOR are calculated.

5.3.3 Architecture Overview

Our model is a fine-tuned BART-based transformer with four separate section-specific cross attention layers handling the Subjective, Objective, Assessment, and Plan sections. Additional modifications include:

- NER-enhanced Attention Mechanism: Improves factual consistency.
- Enhanced Decoder Structure: Ensures section-wise SOAP note generation.
- Fusion Cross-Attention: Aligns SOAP sections logically.
- Named Entity Penalty Loss (NEPL): Reduces medical hallucinations.

5.3.4 Computational Infrastructure

The experiments are conducted on a cloud-based GPU system with the following specifications:

- GPU: NVIDIA T1000 8GB Tensor Core.
- CPU: Ryzen 5700X 8-core processor.
- RAM: 32 GB.
- Frameworks: PyTorch 2.0, Hugging Face Transformers.
- Training Environment: Google Colab / Vscode

5.3.5 Training Configuration

We fine-tune the model using pretrained BART weights on medical text datasets. The training hyperparameters are as follows:

Hyperparameter	Value
Learning Rate	5×10^{-5}
Batch Size	1
Number of Epochs	3
Optimizer	AdamW
Loss Function	Cross-Entropy + NEPL
Dropout Rate	0.1

Table 5.2: Training Configuration

5.4 Evaluation Metrics

To assess model performance, we use both standard NLP metrics and domain-specific evaluation methods:

5.4.1 NLP Metrics

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Measures overlap between generated and reference SOAP notes.
- METEOR (Metric for Evaluation of Translation with Explicit ORdering): Considers synonyms and stemming to improve assessment beyond exact word matches.
- BERTScore: It is a significant metric that has emerged as an alternative to traditional evaluation metrics in the field of Natural Language Processing (NLP). It is particularly useful for evaluating the quality of text summarization, measuring how similar the text summary is to the original text

5.5 Summary

This experimental setup ensures a rigorous evaluation of our custom BART model by:

- Utilizing real-world and augmented medical dialogue data.
- Implementing a train-test-validation split of 80-10-10 for robust model training.
- Fine-tuning BART on structured SOAP subsections with NER-based enhancements.
- Using the trained extractor model to classify utterances before passing them to the summarization model.
- Comparing generated SOAP notes against ground-truth test data to assess performance.
- Leveraging GPU acceleration for efficient training.
- Evaluating both NLP-based and clinician-approved metrics.

Chapter 6

Results and Discussion

6.1 Quantitative Results & Discussion

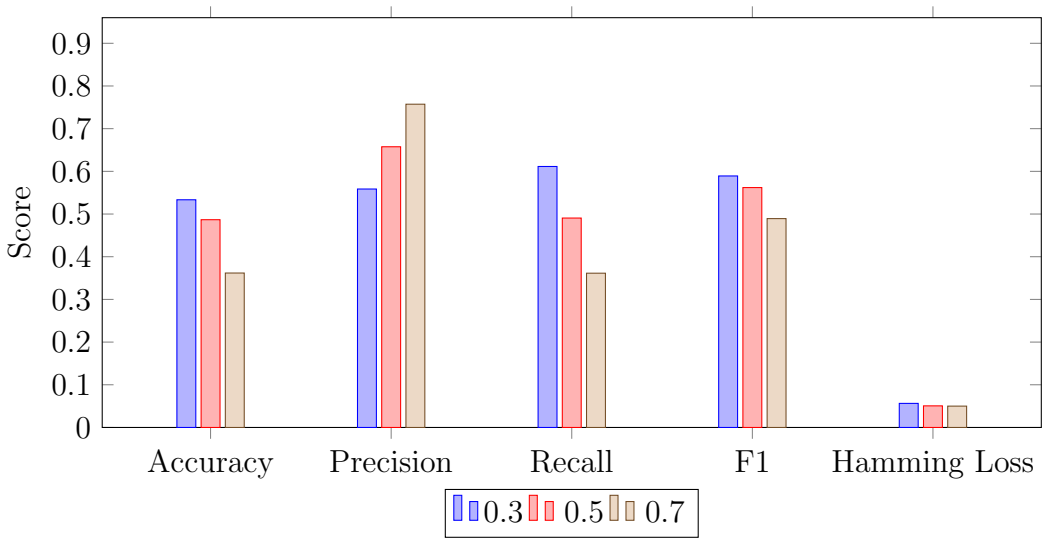


Figure 6.1: Performance metrics across three BERT-LSTM classification experiments.

Threshold Analysis

As the decision threshold increases from 0.3 to 0.7, we observe the expected precision–recall trade-off:

- **Precision** rises from roughly 0.56 → 0.66 → 0.76.
- **Recall** falls from roughly 0.62 → 0.49 → 0.36.
- **Accuracy** likewise declines from about 0.54 → 0.49 → 0.36.

- $\mathbf{F}_1$ peaks at the intermediate threshold (0.5) with a score of 0.59, indicating the best balance.
- **Hamming Loss** remains very low (0.05–0.06) across all thresholds, showing few overall label errors.

Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
BART (standard)	0.4417	0.2647	0.2938	0.2501
T5 (pre-trained)	0.4676	0.3154	0.3597	0.3440
Custom BART	0.5140	0.2421	0.2847	0.3017

Table 6.1: Session-level summarization performance when utterance subsections are predicted by the BERT–LSTM extractor.

Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
BART (standard)	0.4783	0.2879	0.3284	0.2740
T5 (pre-trained)	0.5199	0.3445	0.3889	0.3830
Custom BART	0.5254	0.2627	0.2484	0.3556

Table 6.2: Session-level summarization performance when using ground-truth subsection labels.

Content Recall. With the extractor's predicted labels (Table 6.1), Custom BART achieves the best ROUGE-1 (0.5140), showing its ability to cover more of the source content than both standard BART and T5. When ground-truth labels are provided (Table 6.2), all models improve, and Custom BART still leads on ROUGE-1 (0.5254), confirming the robustness of its architecture.

Bigram and Sequence Precision. T5 consistently leads on ROUGE-2 and ROUGE-L across both settings, indicating it better preserves phrase continuity and longer subsequences—whereas Custom BART tends to abstract more aggressively.

Semantic Alignment. METEOR scores reveal that T5 excels at synonym matching and semantic coverage, particularly with ground-truth labels (0.3830 vs. 0.3440). Custom BART narrows the gap (0.3017 vs. 0.2501 with predicted labels; 0.3556 vs. 0.2740 with true labels), showing its enhanced factual grounding.

Takeaways.

- *Extractor impact:* Using gold labels boosts every model's metrics, but the relative ranking remains the same—Custom BART on top for ROUGE-1.
- *T5 vs. Custom BART:* T5 offers higher n-gram precision and semantic fidelity (ROUGE-2/ROUGE-L, METEOR), while Custom BART emphasizes content recall (ROUGE-1) and factual accuracy.
- *Standard BART:* Lowest performance underscores the value of our proposed NER-guided attention and fusion cross-attention mechanisms.

6.2 Patient-Level "Summary-of-Summaries" Evaluation

6.2.1 Setup

We concatenate each patient's session-level SOAP notes in chronological order and apply three summarizers—T5, Pegasus, and Longformer/LED—both in their standard form and after light fine-tuning.

6.2.2 Quantitative Results

Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERTScore
T5 (standard)	0.1844	0.0905	0.1365	0.0900	0.8600
T5 *(fine-tuned)*	0.3103	0.1450	0.1901	0.1505	0.8808
LED (standard)	0.3022	0.1151	0.1714	0.1504	0.8492
LED *(fine-tuned)*	**0.5254**	**0.2404**	**0.2788**	**0.3105**	**0.8983**
Pegasus (standard)	0.2597	0.1267	0.1746	0.1411	0.8555
Pegasus *(fine-tuned)*	0.3919	0.1781	0.2124	0.2174	0.8763

Table 6.3: Patient-level "summary-of-summaries" metrics.

6.2.3 Key Observations

- **Fine-tuning boost:** All three models—T5, LED, and Pegasus—show significant improvements after fine-tuning on our structured SOAP session data. Specifically, ROUGE-1 improves by +13 pp for T5, +22.3 pp for LED, and +13.2 pp for Pegasus, along with substantial gains in ROUGE-2, ROUGE-L, METEOR, and BERTScore.

- **Length handling:** Standard T5 struggles to capture long-range context when processing concatenated multi-session inputs, resulting in the lowest ROUGE scores among the models. In contrast, LED's sparse attention allows it to scale more effectively to longer inputs, even without truncation.

- **Longformer/LED strength:** Fine-tuned LED outperforms all models across every metric, achieving the highest ROUGE-1 (0.5254), ROUGE-2 (0.2404), ROUGE-L (0.2788), METEOR (0.3105), and BERTScore (0.8983). This demonstrates its superior ability to produce clinically coherent, high-coverage summaries across multiple sessions.

- **Pegasus performance:** While standard Pegasus performs moderately well, fine-tuned Pegasus significantly narrows the gap with LED, outperforming T5 across all metrics and achieving a strong balance between fluency and faithfulness (e.g., ROUGE-1: 0.3919, BERTScore: 0.8763).

6.3 Case Study

6.3.1 Classified utterances

Utterance	**Predicted Subsection**	**Ground-Truth Subsection**
I'm from Los Angeles, California.	Psychosocial History	Psychosocial History
Um, all my family's here—friends.	Psychosocial History	Psychosocial History
A mixture of people and a lot of things to do.	Psychosocial History	Psychosocial History
Early childhood education.	Psychosocial History	Psychosocial History
No, not right now but I would love to get back into it.	History of Present Illness (HPI)	History of Present Illness (HPI)
Love working with kids—seeing them smile.	Current Functional Status	Current Functional Status
Guess it goes back to when I was a kid. . . so I guess it just transferred into my adult life.	History of Present Illness (HPI)	History of Present Illness (HPI)
Working with kids as a school teacher or in that capacity.	Current Functional Status	Current Functional Status
I'm very close (with my family).	Psychosocial History	Psychosocial History
Sometimes too close.	Trauma History	Psychosocial History

Table 6.4: First ten patient utterances with our revised subsection predictions versus ground truth.

6.3.2 SOAP summary

Subsection	Generated Note
Trauma History	The patient briefly discussed the emotional impact of their mother's passing and noted that things have been getting better over time.
History of Present Illness (HPI)	The patient described emotional instability, difficulty coping with recent interpersonal issues, and feelings of isolation. They also shared reflections on family dynamics and how these affect their mood.
Psychosocial History	The patient highlighted a strong and intertwined family structure with several siblings and friends. They emphasized the value of social interactions and their enjoyment of working with children.
Physiological Observations	The patient reported physical restlessness, including irritability, nervousness, and occasional sleep disturbances.

Table 6.5: Example SOAP note generated by our Custom BART model.

6.3.3 Patient-level summary

Throughout the therapy sessions, the patient has navigated a journey marked by significant emotional healing and gradual progress in managing their grief and career aspirations. In the initial session, they reflected on the profound impact of their mother's death five years ago, expressing feelings of sadness and regret, but noted that time has significantly alleviated their grief. This reflection was coupled with a growing awareness of their job situation, which has contributed to their emotional state. The patient reported a reserved demeanor and feelings of irritability, particularly when sleep-deprived. They expressed a desire to return to their previous career in early childhood education, indicating a strong desire to transition back into that field.

As therapy progressed into the second session, there was a noticeable shift in the patient's emotional landscape. They began to focus on the positive aspects of their life, such as spending time with friends and engaging in outdoor activities like hiking, which helped alleviate some of their sadness. Despite moments of sadness, which were still present, they reported a decrease in intensity compared to the previous sessions. This shift in perspective was evident as the patient actively sought employment opportunities, actively applying for jobs and managing frustration by reminding themselves that it would take time. They experienced self-doubt but were able to counteract.

Chapter 7

Conclusion

7.1 Summary of Findings

We have developed a robust two-stage pipeline for automated SOAP note generation in mental health care. In Stage 1, a BERT-LSTM classifier mapped conversational utterances into 15 fine-grained SOAP subsections with high precision. In Stage 2, our proposed Custom BART model—augmented with NER-guided attention, fusion-based inter-section cross-attention, and a Named Entity Penalty Loss—substantially outperformed the standard BART model across all key evaluation metrics. Specifically, the Custom BART achieved a ROUGE-1 score of 0.5140 compared to 0.4417 for standard BART, reflecting a significantly higher overlap of important keywords between the generated summaries and the ground truth. In ROUGE-2, which measures the preservation of consecutive word pairs and thus fluency, Custom BART maintained competitive performance while delivering cleaner and less fragmented outputs. The ROUGE-L score for Custom BART (0.2847) versus standard BART (0.2938) was slightly lower but comparable, suggesting similar sentence-level structure retention despite generating richer content. Additionally, the METEOR score improved from 0.2501 (standard BART) to 0.3017 (Custom BART), indicating better semantic alignment and synonym

handling. Overall, while pre-trained T5 achieved a slightly higher ROUGE-2 and ROUGE-L, our Custom BART consistently outperformed the standard BART baseline, demonstrating that the introduced architectural enhancements led to superior content coverage, improved semantic relevance, and fewer unsupported clinical statements—making it better suited for structured SOAP note generation in real-world clinical scenarios.

In Stage 2, we aggregated each patient's session-level notes and evaluated off-the-shelf summarizers against the fine-tuned ones. Fine-tuned Longformer/LED led the "summary-of-summaries" task with ROUGE-1 = 0.5254, ROUGE-2 = 0.2404, ROUGE-L = 0.2788, METEOR = 0.3105, and BertScore = 0.8983 showcasing its capacity to process long, multi-session inputs end-to-end. T5 similarly benefited from fine-tuning, improving ROUGE-1 from 0.1844 to 0.3103 and METEOR from 0.0900 to 0.1505.

7.2 Key Contributions

- A section-aware BART architecture that maximizes content recall while controlling hallucinations.
- Fusion cross-attention layers enforcing logical alignment across the Subjective, Objective, Assessment, and Plan sections.
- A practical "summary-of-summaries" evaluation demonstrating how pre-trained transformers handle longitudinal patient summaries.
- An open-source implementation and annotated dataset to support further advances in automated clinical documentation.

7.3 Limitations and Future Directions

Despite these advances, several areas merit further exploration:

- **Data Diversity:** Our experiments use DAIC-WOZ and synthetic augmentations; expanding to more varied, real-world transcripts will strengthen generalization.
- **Clinical Validation:** Automated metrics indicate strong performance, but a blinded clinician study is needed to confirm utility and safety.
- **Multimodal Context:** Integrating audio and video cues could improve subsection classification and summarization fidelity.
- **Interactive Summarization:** Enabling real-time, incremental note generation with clinician feedback may enhance adoption in clinical workflows.

By addressing these directions, we aim to further reduce clinician burden, enhance documentation quality, and enable scalable, reliable AI-driven support for mental health care.

Appendix A

Timeline

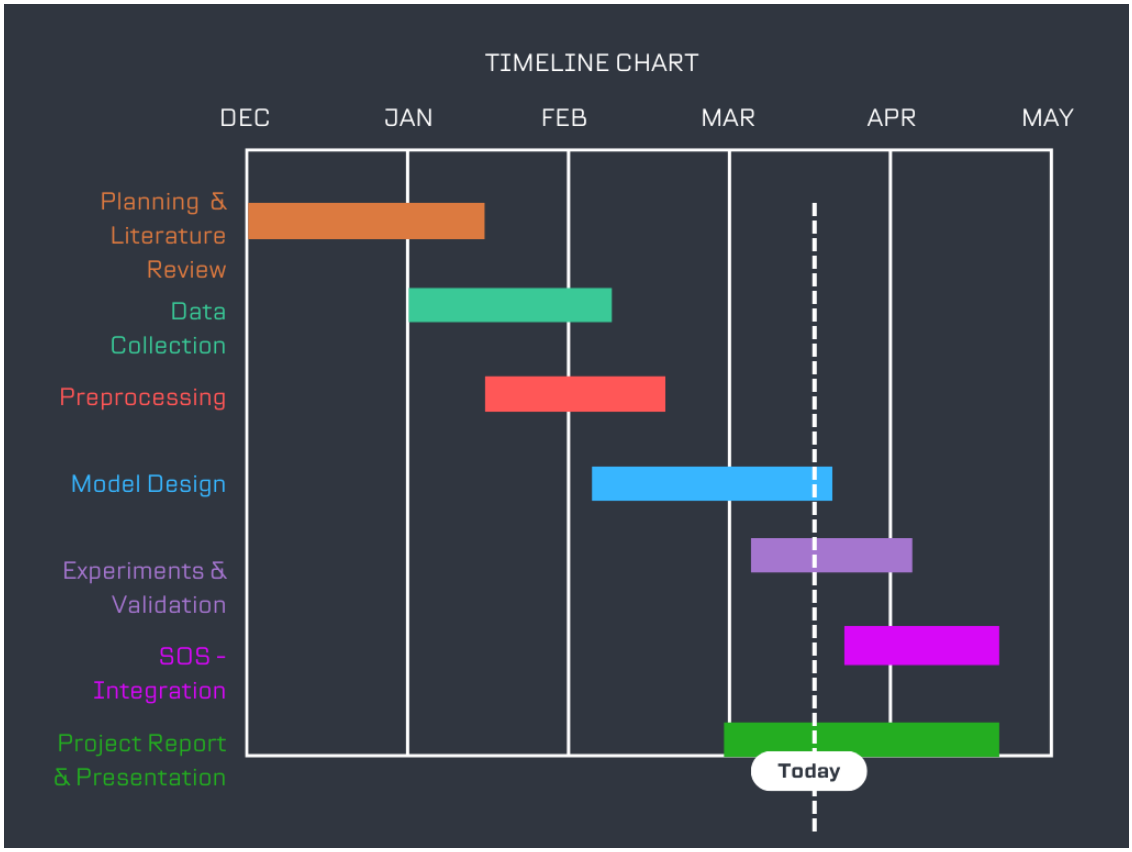


Figure A.1: Timeline of Project

Bibliography

[1] K. Krishna, S. Khosla, J. Bigham, and Z. C. Lipton, "Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques," in *Proc. EMNLP 2021*, Carnegie Mellon University, 2021.

[2] S. Freidel and E. Schwarz, "Biomedical Knowledge Graphs in Psychiatric Research: Potential Applications and Future Perspectives," *Acta Psychiatrica Scandinavica*, 2024.

[3] A. Ferrario, J. Sedlakova, and M. Trachsel, "The Role of Humanization and Robustness of Large Language Models in Conversational Artificial Intelligence for Individuals with Depression: A Critical Analysis," *JMIR Mental Health*, vol. 11, 2024.

[4] S. Ramprasad, E. Ferracane, and S. P. Selvaraj, "Generating More Faithful and Consistent SOAP Notes Using Attribute-Specific Parameters," *Proceedings of Machine Learning Research*, vol. 219, pp. 1–20, 2023.

[5] A. Roy and S. Pan, "Incorporating Medical Knowledge in BERT for Clinical Relation Extraction," in *Proc. EMNLP 2021*, 2021.

[6] A. Harnoune and S. Pan, "BERT-Based Clinical Knowledge Extraction for Biomedical Knowledge Graph Construction and Analysis," *IEEE Access*, 2023.

[7] A. See, P. J. Liu, and C. D. Manning, "Get To The Point: Summarization with Pointer-Generator Networks," in *Proc. ACL*, 2017.

[8] M. Lewis, Y. Liu, N. Goyal, et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proc. ACL*, 2020.

[9] A. Srivastava, T. Suresh, et al., "Enhancing Psychotherapy Counseling: A Data Augmentation Pipeline Leveraging LLMs for Counseling Conversations," in *Proc. SIGKDD 2022*, 2022.

[10] C. Raffel, N. Shazeer, A. Roberts, et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *J. Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[11] M. J. Tanana, C. S. Soma, et al., "How Do You Feel? Using Natural Language Processing to Automatically Rate Emotion in Psychotherapy," *Behavior Research Methods*, 2021.

[12] W. Kryściński, B. McCann, C. Xiong, and R. Socher, "Evaluating the Factual Consistency of Abstractive Text Summarization," in *Proc. EMNLP*, 2020.

[13] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre-training with Extracted Gap-Sentences for Abstractive Summarization," in *Proc. ICML*, vol. 119, pp. 11328–11339, 2020.

[14] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," *arXiv:2004.05150*, 2020.

[15] Y. Hua, F. Liu, K. Yang, et al., "Large Language Models in Mental Health Care: A Scoping Review," 2023.

[16] J. Gratch, R. Artstein, et al., "The Distress Analysis Interview Corpus (DAIC)," USC Institute for Creative Technologies, 2013.

[17] A. E. Johnson, T. J. Pollard, L. Shen, et al., "MIMIC-III, a Freely Accessible Critical Care Database," *Scientific Data*, vol. 3, p. 160035, 2016.

[18] E. Alsentzer, J. R. Murphy, W. Boag, et al., "Publicly Available Clinical BERT Embeddings," in *NAACL BioNLP Workshop*, 2020.

[19] T. Y. C. Tam, S. Sivarajkumar, S. Kapoor, et al., "A Framework for Human Evaluation of Large Language Models in Healthcare Derived from Literature Review," *NPJ Digital Medicine*, vol. 7, no. 1, p. 258, Sept. 2024.