B. Tech. Project Final Presentation

Towards Automated SOAP Note Generation in Conversational Mental Health Care

Sharvari Phand (112103109)   Siddhant Tonapi (112103151)
Pushkar Ugale (112103154)

Under the guidance of
Dr. Yashodhara V. Haribhakta
COEP Technological University, Pune

May 2025

Outline

Background & Motivation

Mental health disorders—especially depression and anxiety—are among the leading causes of global disability, affecting over 500 million people worldwide [2, 3].

Barriers to care:

- Clinician shortages and long wait times
- Stigma and patient reluctance
- Administrative burden of SOAP documentation

Clinicians spend 30–40 % of their time writing SOAP notes, detracting from patient interaction and contributing to burnout [1]. Automated SOAP note generation can reduce this burden, improve consistency, and scale digital mental health services.

Clinical Significance

SOAP notes standardize:

- **Subjective:** Patient-reported symptoms and history
- **Objective:** Clinician observations and measurements
- **Assessment:** Clinical impressions and diagnoses
- **Plan:** Treatment recommendations and follow-up

Importance in mental health:

- **Continuity of Care:** Seamless handoffs between providers
- **Quality Assurance:** Audit and review for best practices
- **Outcome Tracking:** Longitudinal monitoring of symptoms

Technical Challenges & Scope

Designing a reliable SOAP summarization system requires:

1. **Utterance Segmentation:** Map each turn to one of 15 SOAP subsections
2. **Factual Consistency:** Prevent "hallucinations" of unsupported medical details [4]
3. **Structural Coherence:** Ensure Assessment informs Plan and Objective aligns with Subjective
4. **Data Scarcity & Variability:** Limited labeled transcripts; heterogeneous language

Scope:

- English, text-only therapy session transcripts (no audio/video)
- Focus on 15 fine-grained SOAP subsections
- End-to-end: utterance classification → structured SOAP → longitudinal summary
- Comparative evaluation of T5, Pegasus, and LED (pretrained & fine-tuned)

Literature Review: Clinical Summarization for SOAP Notes

- **Pointer–Generator Networks** [7]
  - Copy–abstract balance
  - $\downarrow$ hallucinations by 25 %
- **Cluster2Sent** [1]
  - Cluster utterances by SOAP section
  - Extract representative turns $\rightarrow$ 1-sentence abstractive decode
  - +8 ROUGE-1 vs. end-to-end; clinicians rate $\uparrow$ coherence
- **Section-Aware BART** [4]
  - One cross-attention block per SOAP section
  - +3–5 pp UMLS concept overlap; $\downarrow$30 % expert-noted errors

Fine-Grained Summarization in Mental Health

- **ConSum** [9]
  - PHQ-9 lexicon $\rightarrow$ filter depression utterances
  - Classify into counseling components $\rightarrow$ summarize each
  - $+7$ ROUGE-1; $\uparrow$ MHIC metric coverage
- **Emotion Tagging** [11]
  - Pre-tag emotional expressions
  - $+15$ % sentiment alignment with therapist ratings

Factual Consistency & Hallucination Control

- **FactCC** [12]: NLI-based detection; $\downarrow$ unsupported facts by 20 %
- **SpanCopy**: entity-aware copying; $\uparrow$ medical entity precision by 2.3 pp
- **Our Approach:**
  - NER-guided attention to up-weight clinical tokens
  - Named Entity Penalty Loss to penalize omissions hallucinations

Comparative Evaluation of Summarization Backbones

- **T5** [10]: text-to-text denoising; up to 1024 tokens; strong fluency
- **PEGASUS** [13]: gap-sentence pretraining; focused abstractive summaries
- **Longformer/LED** [14]: sparse global/local attention; up to 4096 tokens, ideal for long context

Trade-off: T5/PEGASUS excel at moderate length; LED excels at long-form recall.

Research Gaps

- **Subsection-Level Segmentation:** Collapse of SOAP into 4 blocks misses clinical subsections (e.g. Presenting Problem, Trauma History). Hierarchical LSTM / BERT-LSTM achieve only 60–70 % on 4-way, ¡50 % on 10+ subsections [1].
- **Hallucinations & Factual Inconsistency:** Abstractive BART/T5 models invent unsupported details (e.g. wrong dosages). 20 % hallucination rate observed for fine-tuned BART [4].
- **Cross-Section Coherence:** Generated sections often misalign (e.g. Assessment Plan), requiring manual correction.

Research Gaps

- **Domain Adaptation & Contextual Grounding:** Vanilla transformers lack sensitivity to psychiatric terms (e.g. "anhedonia"). ClinicalBERT/BioBERT excel on EHR but untested on colloquial therapy transcripts [18, 5].
- **Data Scarcity & Augmentation:** Public corpora (DAIC-WOZ, counseling logs) are small and narrow. Synthetic LLM augmentation risks unnatural dialogue.
- **Evaluation Limitations:** ROUGE/BLEU miss clinical correctness. Domain-aware scores and clinician ratings lack a standardized framework.

Problem Statement

Develop a multi-step pipeline that uses our proposed modified BART architecture to:

1. Generate structured, clinically faithful SOAP notes from session transcripts.
2. Aggregate SOAP notes across multiple therapy sessions into a comprehensive longitudinal summary.
3. Benchmark transformer-based summarizers on both tasks.

Key Objectives

1. **Dataset Expansion:** Augment DAIC-WOZ with synthetic therapy dialogues via OpenAI API.
2. **Fine-Grained Labelling for Mental Health Care:** 15-way *BERT–LSTM* classifier for SOAP subsection tags.
3. **Hallucination Control:** Enhanced *BART* with NER-guided attention & Named Entity Penalty Loss.
4. **Section-Aware Coherence:** Section-specific & fusion cross-attention layers in the decoder.
5. **Longitudinal Summaries:** Chronological concatenation of session notes, summarized by Pegasus, T5, LED.
6. **Comprehensive Evaluation:** ROUGE/METEOR/BERTScore, and entity accuracy.

Methodology: Dataset Creation

- **Base Corpus: DAIC-WOZ Transcripts**
  - Real patient–therapist interviews recorded at USC's Institute for Creative Technologies.
  - Rich in mental-health dialogue: depression, anxiety, counseling nuances.
  - Consists of 189 therapist-patient sessions, conducted by an animated virtual interviewer called Ellie.
  - These interactions range between 7-33 minutes (average is 16 minutes).
- **Synthetic Augmentation via OpenAI API**
  - Prompted OpenAI API to simulate follow-up therapy sessions for each one of the session in original dataset.
  - Total size of dataset after expansion: 626 sessions
  - Generated ground-truth SOAP notes and a target summary for patient-level summarization using OpenAI API.
  - Ensured clinical plausibility by constraining topics to known mental-health scenarios.
  - Expanded dataset size by 200%, improving model generalization.

Data Cleaning & Preprocessing

- **Linguistic Processing**
  - Tokenization and POS tagging
  - Lemmatization to dictionary forms
- **Normalize Text**
  - Expand contractions (e.g. "can't" $\rightarrow$ "cannot")
  - Spell-check and correct typos
- **Clean Filter**
  - Remove stopwords, punctuation, extra whitespace
  - Convert to lowercase
  - Replacing slang words
- **Resolve References**
  - Coreference resolution to replace pronouns with entities
- **Medical NER Extraction**
  - Generic spaCy entities + custom mental-health term lookup

SOAP Sections and Subsections

Subjective	• **Presenting Problem / Chief Complaint**: Main reason for seeking therapy • **Trauma History**: Past traumatic experiences • **Substance Use History**: Alcohol, drugs, smoking, impact • **History of Present Illness (HPI)**: Duration, triggers, progression • **Medical & Psychiatric History**: Past diagnoses, current meds • **Psychosocial History**: Relationships, family, social life • **Risk Assessment**: Suicide risk, self-harm, harm to others
Objective	• **Mental Health Observations**: Mood, affect, cognition, insight • **Physiological Observations**: Sleep, appetite, energy • **Current Functional Status**: Ability to perform daily activities
Assessment	• **Diagnostic Impressions**: Possible or confirmed diagnoses • **Progress Evaluation**: Changes since last session
Plan	• **Medications** • **Therapeutic Interventions**: CBT, DBT, psychoeducation, etc. • **Next Steps**: Goals, journaling, behavior tracking

COEP Tech

Two-Stage Pipeline

1. **Stage 1:** Fine-Grained SOAP Note Generation
   - Classification of utterances using BERT-LSTM.
   - Domain-aware encoder + enhanced decoder BART
   - NER-guided attention & penalty loss
   - Inter-section cross-attention for coherence
2. **Stage 2:** Patient-Level Summarization
   - Concatenate session-level notes per patient
   - Summarize with T5, Pegasus, LED (pretrained vs. fine-tuned)

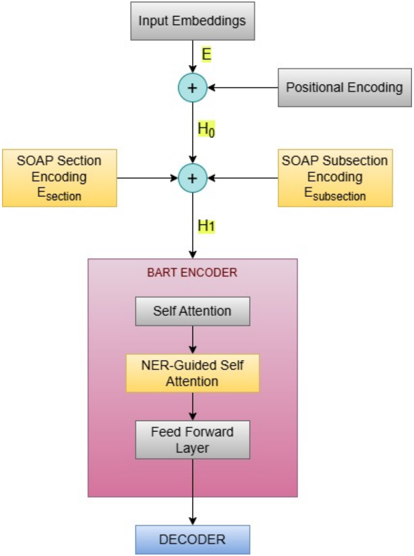
Stage 1: Input Representation

Embedding Sum at Position i:

$$\mathbf{h}_i^{(0)} = E_{\text{tok}}[x_i] + E_{\text{pos}}[i+2] + E_{\text{sec}}[s_i] + E_{\text{sub}}[t_i].$$

- **Token Embeddings** $E_{\text{tok}} \in \mathbb{R}^{V \times d}$
- **Section/Subsection** embeddings for 4 SOAP sections & 15 subsections

BART Encoder with NER-Guided Bias



Custom BART Encoder Overview

- Multi-layer transformer encoder with section/subsection embeddings
- **NER-Guided Attention:**

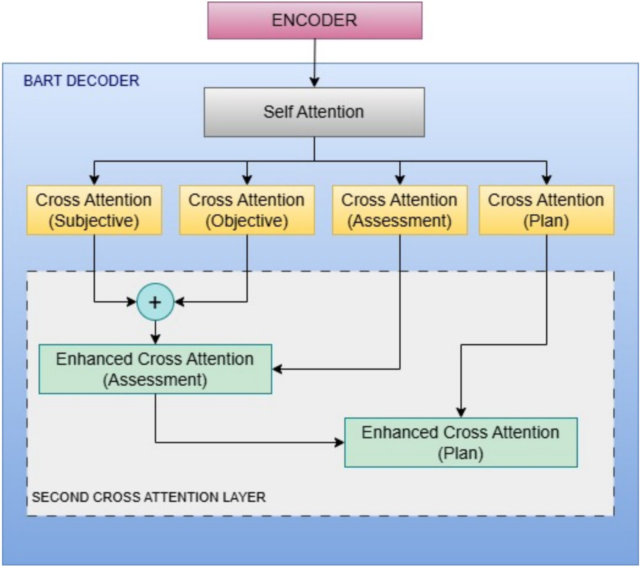
$$A' = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V \; + \; \gamma \, h_{\mathrm{NER}}$$

  - h_{NER}: embeddings of recognized medical entities
  - γ: (=0.1) learnable weight to emphasize clinical tokens
- **Chunking for Long Sequences:**
  - Split input if length $n > L$ (=1024) into segments of size L
  - Independently encode each, then average-pool outputs

Custom BART Decoder Structure

- Four independent decoders: Subjective, Objective, Assessment, Plan
- Each decoder attends to shared encoder outputs $\mathbf{H}^{(N)}$
- Ensures section-specific language modeling

Custom BART Decoder Diagram



Inter-Decoder & Fusion Cross-Attention

- **Inter-Decoder Attention:**

$$A_{\mathrm{enh}} = \mathrm{softmax}\left(\frac{Q_A\,(S\|O)^\top}{\sqrt{d}}\right)(S\|O)$$

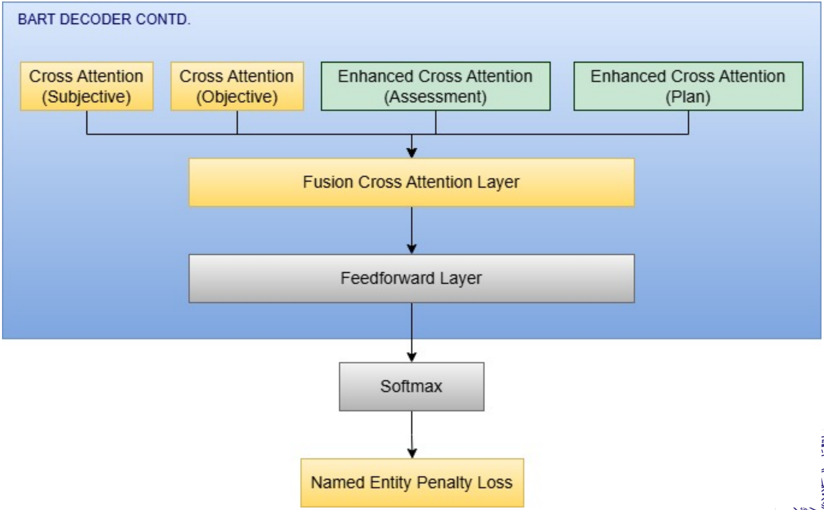
  propagates information from Subjective $\rightarrow$ Objective $\rightarrow$ Assessment

- **Fusion Attention:**

$$F = \sum_i \alpha_i\,\mathrm{Attn}_i,$$

  where the weights α_i are learned to combine section signals

Inter-Decoder and Fusion Cross-Attention Diagram



Loss Functions & Decoding

Total Loss:

$$L = L_{\mathrm{CE}} + \lambda \left| \{\text{missed}\} \cup \{\text{hallucinated}\} \right|, \quad \lambda = 5 \times 10^{-2}.$$

- $\{\text{missed}\}$: entities in input but omitted
- $\{\text{hallucinated}\}$: entities generated without basis
- **Decoding:** beam search (block repeats, length norm, no early EOS)

Stage 2: Patient-Level Summarization

1. **Aggregate** all session-level SOAP notes into a unified input (excluding empty sections), and append each sessions with session markers.
2. **Inference using the following models (both pre-trained and fine-tuned):**
   - **T5** – A text-to-text transformer that treats summarization as a generative task. Effective for general-purpose summarization.
   - **PEGASUS** – Designed specifically for abstractive summarization by masking and predicting key sentences (gap-sentence objective).
   - **Longformer/LED** – Uses sparse global attention to process long-context inputs efficiently, ideal for multiple session inputs.
3. **Why these models?**
   - T5 and PEGASUS are strong baselines for abstractive summarization.
   - LED is chosen for its ability to handle long, structured SOAP input spanning multiple sessions.
   - Fine-tuning allows each model to better adapt to the clinical domain and our SOAP-specific structure.

Train–Validation–Test Split

- **Training (80%)**: Utterance classifier + BART summarizer
- **Validation (10%)**: Hyperparameter tuning (LR, dropout, loss weights)
- **Test (10%)**: Final unseen evaluation

Patient-Level Summaries

- **Training (90%)**
- **Test (10%)**

Model Training and Evaluation Overview

1. **SOAP Section Summarization:**
   - Train a custom **BART-based model** on labeled SOAP section data.
   - Tune hyperparameters: NER-loss weight, beam size.
2. **Patient History Summarization:**
   - Compare **T5, PEGASUS, and LED (pretrained vs fine-tuned)** for generating longitudinal summaries.

Evaluation Metrics:

- **ROUGE-1/2/L:** Measures word/phrase overlap between generated and reference summaries.
  - *ROUGE-1*: Unigram (word-level) overlap.
  - *ROUGE-2*: Bigram overlap.
  - *ROUGE-L*: Longest common subsequence.
- **METEOR:** Considers synonyms and stemmed matches; aligns semantically similar words more robustly.
- **BERTScore:** Uses contextual embeddings from BERT to compare semantic similarity between predicted and gold summaries.

Training Configuration & Infrastructure

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

Hardware & Software	
GPU	NVIDIA T1000 8 GB
CPU	Ryzen 5700X 8-core
RAM	32 GB
Frameworks	PyTorch 2.0, HuggingFace
Env.	Google Colab / VS Code

Experimental Setup Summary

- **Data:** DAIC-WOZ + synthetic, 80/10/10 split
- **Training:** BERT-LSTM extractor → custom BART model
- **Infrastructure:** GPU acceleration, PyTorch ecosystem
- **Metrics:** ROUGE, METEOR, BERTScore
- Enables robust, reproducible evaluation of fine-grained SOAP summarization

Results & Discussion Overview

- BERT–LSTM Threshold Analysis
- Session-level Summarization (Predicted vs. Gold Labels)
- Key Session-level Findings
- Patient-level “Summary-of-Summaries” Evaluation
- Key Takeaways and Trends

Threshold Analysis: BERT–LSTM Classification

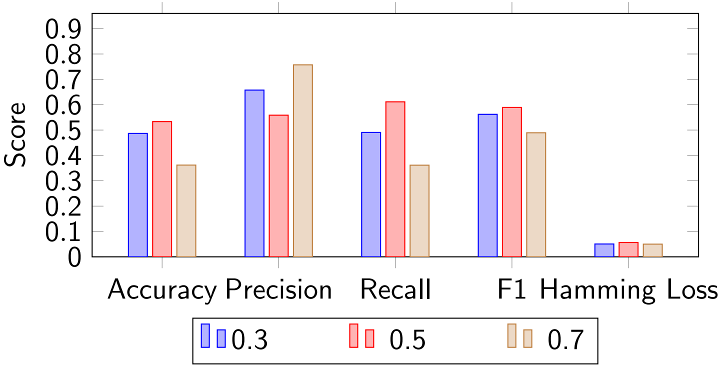


Figure: Performance at three decision thresholds.

- ↑ Precision/↓ Recall as threshold ↑
- F_1 peaks at 0.5 best balance
- Hamming Loss very low (0.05–0.06)

Session-level Summarization (Gold Labels)

Model	R-1	R-2	R-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: With ground-truth SOAP subsection labels.

Session-level Summarization (Predicted Labels)

Model	**R-1**	**R-2**	**R-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: When subsections are predicted by BERT–LSTM.

Semantic Similarity: BERTScore-F1

Model	**BERTScore-F1**
Standard BART	0.3854
T5 (pre-trained)	0.2817
Custom BART	0.3662

All scores computed with the BERTScore-F1 metric (higher = better semantic alignment).

Session-level Key Findings

- **Content Recall:** Custom BART leads on ROUGE-1 in both settings.
- **Phrase Precision:** T5 tops ROUGE-2/ROUGE-L closer phrasing.
- **Semantic Fit:** METEOR highest for T5; Custom BART narrows gap.
- **Label Quality Impact:** Gold labels boost all models but rankings stay the same.

Patient-level "Summary-of-Summaries"

Model & Setting	R-1	R-2	R-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
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
LED (standard)	0.3022	0.1151	0.1714	0.1504	0.8492
LED (fine-tuned)	**0.5254**	**0.2404**	**0.2788**	**0.3105**	**0.8993**

Table: Chronological multi-session summarization results.

Patient-level Key Observations

- Fine-tuning yields large gains (e.g. +22.3 pp ROUGE-1 for LED).
- T5 struggles on long inputs; LED's sparse attention handles them best.
- LED (fine-tuned) tops all metrics best multi-session coverage.
- Pegasus (fine-tuned) closes gap, balancing fluency and faithfulness.

Summary of Findings

- **Stage 1 (Session-level):**
  - BERT-LSTM classifier mapped utterances into 15 SOAP subsections with high precision.
  - Custom BART (NER-guided attention + fusion cross-attention + NEPL) achieved
    ROUGE-1 = 0.5140 vs. 0.4417 (std. BART), METEOR = 0.3017 vs. 0.2501.
  - Comparable ROUGE-2/ROUGE-L to T5, but with fewer unsupported clinical statements.
- **Stage 2 (Patient-level):**
  - Fine-tuned Longformer/LED led with
    ROUGE-1 = 0.5254, METEOR = 0.3105, BERTScore = 0.8983.
  - All fine-tuned models outperformed their respective pretrained models.

Utterance Classification Example

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

Custom BART–Generated SOAP Note (Part 1)

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.

Custom BART–Generated SOAP Note (Part 2)

Subsection	Generated Note
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.

Patient-level summary (Generated by Fine-tuned LED)

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

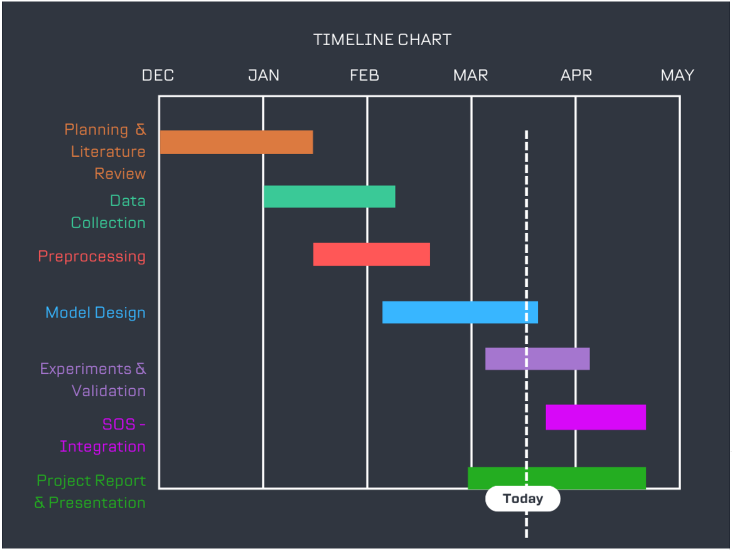
Key Contributions

- **Section-Aware Summarization:** Custom BART maximizing content recall while minimizing hallucinations.
- **Inter-Section Coherence:** Fusion cross-attention layers align Subjective, Objective, Assessment, Plan.
- **Longitudinal Summaries:** "Summary-of-summaries" framework evaluating T5, PEGASUS, LED end-to-end.
- **Reproducibility:** Open-source code and annotated dataset for further research.

Limitations & Future Work

- **Data Diversity:** Extend beyond DAIC-WOZ and synthetic to varied, real-world transcripts.
- **Clinical Validation:** Conduct blinded clinician studies to assess safety and utility.
- **Multimodal Integration:** Incorporate audio/video cues (prosody, expressions) for richer context.
- **Interactive Summarization:** Enable real-time, incremental note generation with clinician feedback.

Project Timeline



[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.

References II

[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.

References III

[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.

Questions & Discussion

Thank You!

Questions?