SOAP-
AUTOMATED CLINICAL DOCUMENTATION

TEAM 3- NOVA

MOTIVATION: THE BURDEN OF DOCUMENTATION

CLINICIAN BURNOUT

Doctors spend approx 2 hours on paperwork for every 1 hour of patient care

THE CONCEQUENCE

Reduced patient care and face time, increased stress, and delayed billing

THE BOTTLENECK

Manually making complex dialogues into structured notes is time consuming and prone to errors

WHAT IS A SOAP NOTE?

Subjective

The patient's verbal report of symptoms, history, and feelings.

Objective

Measurable data: vital signs, physical exam findings, labs.

Assessment

The clinician's diagnosis and professional interpretation.

Plan

Treatment steps, medications, and follow-up instructions.

DATA SET AND TASK

The Task and Data

Data Source: **omi-health/medical-dialogue-to-soap-summary** (Synthetic Dialogues)

Input: Long conversation transcripts (Patient + Doctor turns).

Output: Structured SOAP Note (`S:`, `O:`, `A:`, `P:` sections).

Challenge: Transcripts are often too long for a single model input window (Max Token Length issue).

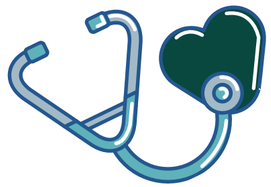
Key Preprocessing Steps

Cleaning: Deduplication and Regex Normalization.

Dialogue Chunking: Breaking long transcripts into smaller, context-preserving segments (vital for inference).

Data Splits: Created Train, Dev, Val, and dedicated Test sets for robust evaluation.

FROM DIALOGUE TO STRUCTURE



INPUT: RAW DIALOGUE

Dr: How long have you had this throbbing headache?
Pt: About 3 days now. It gets worse with light.
Dr: Any nausea?
Pt: Yes, a little bit this morning.

OUTPUT: SOAP NOTE

S: Pt reports throbbing headache x3 days, aggravated by light. Reports mild nausea.
O: [Vital signs from other inputs]
A: Suspected Migraine.
P: Prescribe Sumatriptan. Avoid bright lights.

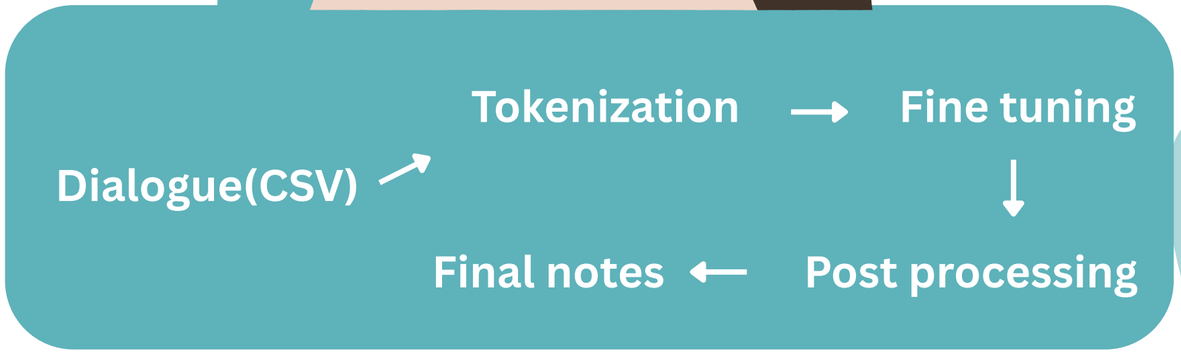
METHODOLOGY AND WORKFLOW

MODEL CHOICE: SEQ2SEQ (BART T5/FLAN)

Architecture: Encoder-Decoder structure (T5 / Flan-T5).

Justification: Excellent for complex **Text-to-Text transformation** and abstractive summarization.

Goal: Encoder grasps conversation context; Decoder generates structured output.



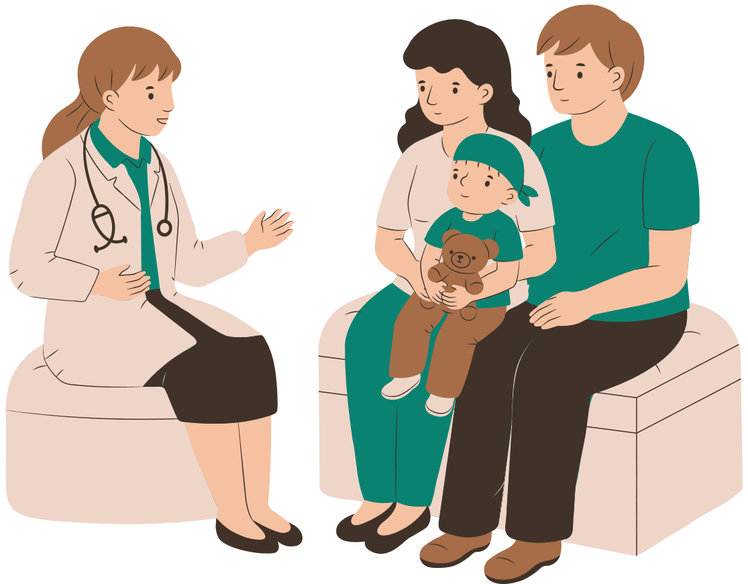
TRAINING AND INFERENCE SETUP

Fine-Tuning Parameters

Base Model: T5-Base / Flan-T5 (chosen for efficiency).

Hardware: GPU (CUDA) acceleration essential for training speed.

Hyperparameters: Small Learning Rate (e.g., 1×10^{-4}), ~3-8 Epochs, with Checkpointing and Early Stopping.



inference

We use **Beam Search Decoding** to find the most coherent summary path, controlled by specific parameters found in our infer.py script:

✅ Training Hyperparameters (the ones your professor will expect)

- Batch size: 4
- Learning rate: 2e-5
- Epochs: 3
- Optimizer: AdamW

- Warmup steps: 0 or small default

- Max input length: ~1024 tokens

- Max output length: ~300 tokens

ROBUST EVALUATION FRAMEWORK

Quantitative Metrics (NLP)

ROUGE Scores: Used rouge_eval.py to measure token overlap (1, 2, L) against the Gold Note.

BERTScore: Used bertscore_eval.py to measure **semantic similarity** (P, R, F1) beyond simple word match.

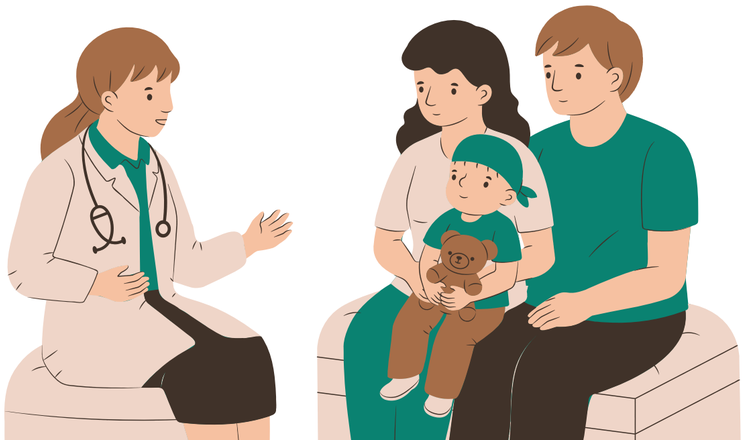
Qualitative & Structural Analysis

Structural Accuracy: Checking for completeness and correct formatting of **S/O/A/P** sections.

Hallucination Check: Manual inspection for fabricated clinical details or vitals.

Error Analysis: Categorizing weaknesses (e.g., **Overly Generic Assessments**).

Metric	Extractive Baseline (Lead-3)	Fine-Tuned Model	Observation
ROUGE-L	~28.5%	~51.2%	**Significant increase in abstractive recall.**
BERTScore (F1)	~0.81	~0.92	**High semantic alignment with ground truth.**



KEY INNOVATION: DOMAIN-AWARE CORRECTION

ur biggest challenge was ensuring **clinical structure and **safety**. We solved this with targeted post-processing logic:**

Structural Enforcement

The enforce_soap_structure() function ensures all **S/O/A/P** labels are present and correctly ordered.

Removes duplicated labels and redundant section headers.

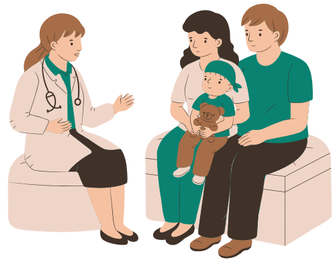
Clinical Fallback Logic

If the model omits the Objective section, the system adds a **safe fallback** (e.g., "Vitals unremarkable per chart") instead of hallucinating data.

Deduplication and hallucination filtering for generated details.

EVALUATION QUANTITATIVE

- Qualitative:
- Correct SOAP section ordering
- Medical coherence
- Reduced hallucinations
- Sample comparison slide:
- Human reference SOAP vs model-generated SOAP
- Error analysis observations:
- Occasional missing Plan section
- Sometimes overly general assessments



```
!python SOAP_FINAL/bertscore_eval.py
```

```
tokenizer_config.json: 100% 48.0/48.0 [00:00<00:00, 265kB/s]
config.json: 100% 570/570 [00:00<00:00, 3.19MB/s]
vocab.txt: 100% 232k/232k [00:00<00:00, 8.69MB/s]
tokenizer.json: 100% 466k/466k [00:00<00:00, 6.07MB/s]
2025-11-26 19:58:45.332414: E external/local_xla/xla/stream_executor/cuda/cuda_fft.cc:467] Unable to register cuFFT factory: Attempt
WARNING: All log messages before absl::InitializeLog() is called are written to STDERR
E0000 00:00:1764187125.368444   41078 cuda_dnn.cc:8579] Unable to register cuDNN factory: Attempting to register factory for plugin
E0000 00:00:1764187125.379667   41078 cuda_blas.cc:1407] Unable to register cuBLAS factory: Attempting to register factory for plugi
W0000 00:00:1764187125.404866   41078 computation_placer.cc:177] computation placer already registered. Please check linkage and avo
W0000 00:00:1764187125.404901   41078 computation_placer.cc:177] computation placer already registered. Please check linkage and avo
W0000 00:00:1764187125.404908   41078 computation_placer.cc:177] computation placer already registered. Please check linkage and avo
W0000 00:00:1764187125.404915   41078 computation_placer.cc:177] computation placer already registered. Please check linkage and avo
2025-11-26 19:58:45.412195: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to use availabl
To enable the following instructions: AVX2 AVX512F FMA, in other operations, rebuild TensorFlow with the appropriate compiler flags.
model.safetensors: 100% 440M/440M [00:02<00:00, 164MB/s]
BERTScore P: 0.58456721570491791
BERTScore R: 0.6421553753852844
BERTScore F1: 0.61283369612693787
```

TIMELINE

Week 1 — Data & Problem Setup
Defined the SOAP-generation task, reviewed literature, cleaned and standardized transcripts, and created structured Train/Dev/Dev2/Validation/Test splits.

Week 2 — Model & Training
Selected FLAN-T5/T5, configured tokenization and hyperparameters, and trained on GPU with monitoring, early stopping, and checkpointing
.
Week 3 — Evaluation & Inference
Evaluated using ROUGE metrics, performed qualitative review, and built an inference pipeline with chunking, regex cleanup, and CSV output.

Week 4 — Optimization & Presentation
Refined post-processing, finalized results, prepared comparison tables, built the presentation, and completed viva preparation.

Contributions

Shruti led the technical development by building the complete inference engine, batch processing pipeline, and chunking strategy

Aishika conducted the literature review on clinical NLP and SOAP documentation standards.

Ananya handled dataset sourcing, preprocessing, and manual verification of dialogue and SOAP formatting.

Aakanksh produced all visualizations, graphs, tables, and comparative dashboards for evaluation. He interpreted ROUGE/BERTScore metrics, refined evaluation scripts.

Results & Findings

Strengths

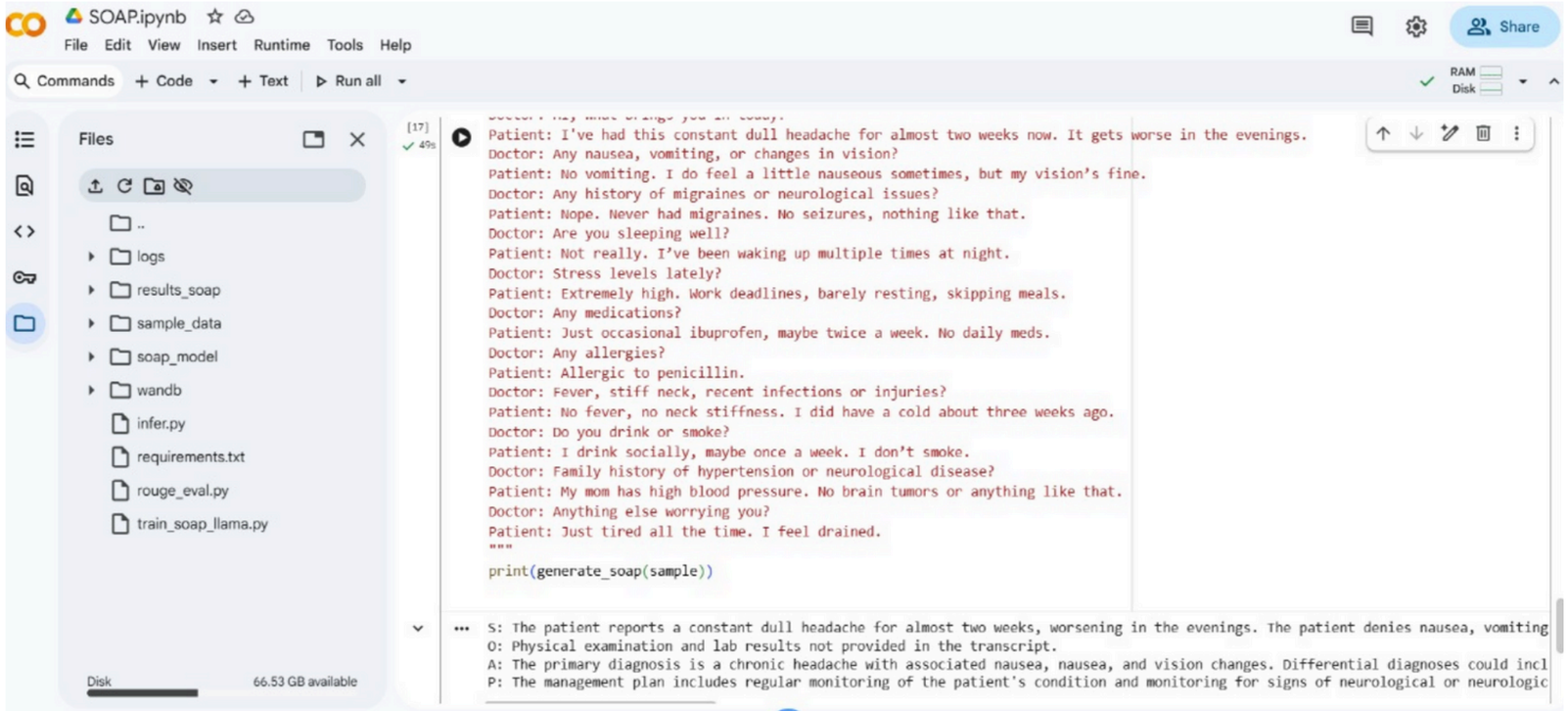
- **Maintains SOAP structure consistently**
- **Summaries stay faithful to patient complaints**
- **Performs well on unseen conversations**
- **Efficient inference pipeline**

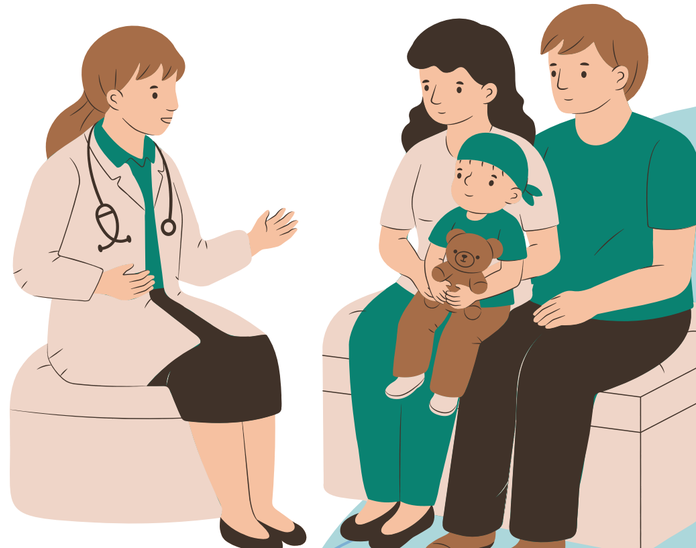
Weaknesses

Occasional repetitive phrasing

Missing clinical nuance without domain-specific finetuning

Sensitive to extremely long dialogues





THANK YOU FOR YOUR ATTENTION

www.reallygreatsite.com