Medical Ontology Integration for Contextualized SOAP Note Synthesis

Dhanush Ravuri
dravuri@umass.edu

Naveen Jarpla
njarpla@umass.edu

Sreehitha R Narayana
snarayana@umass.edu

1 Introduction

Electronic Health Records (EHRs) are integral to modern healthcare, enabling standardized documentation of patient encounters and continuity of care. A large portion of this documentation takes the form of SOAP notes, a structured clinical narratives having the **S**ubjective, **O**bjective, **A**ssessment, and **P**lan components of a doctor consultation. Despite their vital role, SOAP notes are largely manually produced which is time-consuming and can have a huge impact on the amount of time clinicians could potentially spend on face-face interaction and other tasks. Multiple researches have shown that physicians spend an average of two to three hours per day on documentation tasks. Manual entry is not only time-consuming but also prone to errors, potentially affecting the accuracy and consistency of clinical documentation.

Recent advances in Large Language Models (LLMs) have shown promise in automating SOAP note generation from doctor–patient dialogues. However, most of these models often lack domain-specific medical knowledge, which results in factual inconsistencies, hallucinations, and missing important details, especially within the Assessment and Plan sections which require clinical reasoning rather than high level summarization.

An emerging solution lies in integrating structured medical knowledge graphs such as UMLS(Unified Medical Language System) and SNOMED CT into the note-generation. Knowledge graphs encode relationships between diseases, symptoms, and treatments, offering more structured context that can bound LLM outputs in verified medical semantics. Combining these graphs with dialogue derived text enables factual, coherent, and contextually faithful medical notes inline with clinical documentation standards.

Our research therefore addresses a meaningful gap between language generation and structured clinical knowledge integration, offering both theoretical insight and high clinical relevance. Automating high quality note generation not only reduces physician burnout but also contributes to safer, more accurate patient records ultimately enhancing decision-making in healthcare delivery.

Research Questions

This study will investigate the following core questions:

- Can medical dialogues be effectively transformed into structured SOAP notes using large language models (LLMs)?
- How does the inclusion of a medical knowledge graph improve factuality, coherence, and completeness in generated notes?
- What are the common error patterns in baseline LLM-generated SOAP notes, and how can structured knowledge reduce these errors?

2 What We Proposed vs. What We Accomplished

Collect and preprocess clinical dialogue–SOAP note datasets

We initially planned to use large-scale datasets such as MIMIC-III/IV. Although access was successfully obtained, the dataset size and preprocessing requirements exceeded the available computational resources. As a result, we instead used the

MediSOAP dataset, which is specifically designed for SOAP note generation and more feasible under the given compute constraints.

Build and fine-tune baseline sequence-to-sequence models for SOAP note generation

We fine-tuned baseline models, including T5-base, and T5-large, on the MediSOAP dataset to generate structured SOAP notes from medical dialogues and evaluated their performance using standard summarization and semantic similarity metrics.

Integrate medical knowledge graphs (KGs) into the generation pipeline

We incorporated structured medical knowledge into the model pipeline by linking clinical entities from dialogues to ontology-based representations. This integration aimed to ground the generated SOAP notes in medical knowledge and reduce hallucinations, particularly in the Assessment and Plan sections.

Improve model performance through knowledge graph augmentation

While full-scale integration of large knowledge graphs such as UMLS was limited by computational resources, partial knowledge-aware augmentation still improved factual consistency and semantic alignment compared to text-only baselines. Further improvements remain possible but were limited by resource constraints.

Evaluate models using lexical, semantic, and factuality-based metrics

We evaluated baseline, fine-tuned and knowledge-enhanced models using metrics such as ROUGE and BERTScore, along with factual consistency analysis, to assess improvements beyond surface-level text overlap.

Perform error analysis to identify common failure modes

We conducted qualitative error analysis and observed that fine-tuned models often produced missing or clinically vague statements, whereas knowledge-integrated models reduced omissions and improved coherence in clinically critical sections.

3 Related work

Automated generation of structured clinical documentation, particularly SOAP notes, is emerging as a key research topic in medical natural language processing and clinical summarization. The process involves translating multi-turn doctor-patient conversations into short, clinically valid narratives that follow standard medical documentation forms. Recent advances in big language models have prompted researchers to investigate their potential for producing coherent and contextually relevant SOAP notes using encoder-decoder transformer topologies and retrieval-augmented generation (RAG) methods.

Enarvi et al. (2020) framed clinical documentation as an end-to-end sequence-to-sequence generation task, proposing transformer-based models to generate medical reports directly from patient–doctor conversation transcripts. Using a large-scale dataset of orthopedic encounters, their approach demonstrated that neural abstractive models, particularly transformers with copy mechanisms, can outperform extractive baselines in producing fluent and coherent clinical notes. This work provides an important foundation for automated clinical note generation by validating the feasibility of conversation-to-report modeling at scale. However, the generated reports are free-form and do not explicitly enforce standardized clinical structures such as SOAP, which limits their direct applicability to structured documentation tasks where correct section assignment and clinical rigor are essential.

Building upon these advances, Leong et al. (2025) proposed MediNotes, a generative AI framework for automatic SOAP note generation based on doctor-patient conversations. Their system used a transformer-based LLM to record conversational turns and provide structured SOAP outputs, while integrating a lightweight retrieval component to retrieve further clinical context. The model exhibited high syntactic coherence and sectional structure, suggesting that pre-trained sequence-to-sequence models can accurately capture discourse semantics. However, MediNotes relied mostly on unstructured text retrieval and lacked clear medical ontological knowledge, resulting in factual errors and the occasional creation of wrong or unsubstantiated diagnoses.

Our project addresses this gap by integrating structured medical knowledge graphs to enhance factual grounding and clinical accuracy.

Addressing structural coherence from a different angle, Krishna et al. (2021) developed SOAP notes generation from doctor-patient conversations using Modular Summarization Techniques. Their system divided the summary work into extractive and abstractive steps: it first collected key utterances per SOAP section, clustered semantically related utterances, and then created concise phrases for each cluster with an abstractive model. This modular strategy increased factual consistency and coherence over end-to-end abstractive models, and expert evaluation revealed superior alignment with the desired SOAP structure. However, the system still relied mostly on surface-level utterance grouping and did not employ precise medical ontologies, allowing for factual gaps in clinical reasoning.

To improve the structural and semantic alignment of generated notes, Li et al. (2025) introduced the CliniKnote dataset, which pairs multi-turn clinical dialogues with fully annotated SOAP notes. They proposed the K-SOAP format, augmenting traditional SOAP notes with a keyword section to facilitate quicker information retrieval. To improve data quality, their pipeline integrates clinical Named Entity Recognition (NER) and Relation Extraction (RE), allowing automatic alignment of conversation utterances with structured note sections. This approach effectively reduces manual annotation effort and enhances the semantic structure of generated notes. However, it still relies on surface-level entity and relation extraction without explicitly leveraging structured medical knowledge graphs, which can limit the factual grounding and clinical accuracy of the outputs.

Integrating medical knowledge graphs with big language models has shown great promise in increasing the accuracy and comprehension of clinical documentation, especially for diagnosis prediction. Recent research(Gao et al. (2025)) shows that structured knowledge from resources such as the Unified Medical Language System(UMLS) can be obtained, evaluated based on patient context, and then included into model prompts to produce more factual and clinically appropriate outputs. This method improves diagnostic reasoning and decreases hallucinations, but its success is dependent on the quality of the knowledge graph and the complexity of prompt engineering.

While these studies focus on improving factual correctness and semantic structure, a parallel line of research emphasizes evaluation frameworks for assessing the quality and factual accuracy of generated clinical notes. Abacha et al. (2023) provided an extensive set including lexical metrics (ROUGE-L, BLEU, METEOR), embedding-based semantic similarity (BERTScore, MoverScore) and knowledge-grounded measures, including UMLS-based entity and relation alignment and a clinical application of FactCC to detect factual errors. Their findings revealed that traditional summarization metrics such as ROUGE or BLEU correlate weakly with expert clinical assessments, while knowledge-grounded and ontology- and entity-aware metrics offer a more faithful evaluation of factual completeness and diagnostic validity. The study also highlighted that no single metric suffices for comprehensive evaluation, advocating for hybrid metric frameworks. Drawing on these insights, we can better inform the design of our evaluation pipeline and ensure that the generated clinical documentation aligns with both linguistic quality and domain-specific clinical correctness.

Recent work by Dahlberg et al. (2025) provides a critical re-examination of standard evaluation protocols for clinical text. In their systematic review of 37 studies, they demonstrated that lexical overlap metrics like ROUGE and BLEU frequently penalize meaning-preserving paraphrases a common occurrence in LLM generated notes rendering them insufficient proxies for clinical quality. To address this, they proposed a standardized 'layered evaluation strategy' that prioritizes semantic metrics (e.g., BERTScore) alongside an 'LLM-as-Evaluator' framework. Their experimental benchmark harmonized human evaluation criteria into four core dimensions: correctness, completeness, conciseness, and fluency. This validates our project's decision to supplement traditional metrics with LLM based judges to better capture the clinical nuance and reasoning capabilities of our models

4 Hypothesis

Training LLMs for SOAP note generation typically requires massive labeled datasets, extensive computational resources, and reliance on proprietary closed-source APIs, creating significant barriers to scalability, reproducibility, cost-effective implementation, and broad adoption across diverse clinical environments. Our hypothesis is: *Can we design a hybrid, lightweight, knowledge-informed pipeline that leverages finetuned T5 models in combination with UMLS-based knowledge graph integration to generate clinically coherent SOAP notes from raw doctor–patient transcripts, achieving performance comparable to or better than standard baselines?*

Specifically, we investigate whether supplementing a strong baseline LLM (e.g., T5-Large) with domain-informed embeddings and knowledge-guided post-processing can:

- Generate high-quality, clinically coherent SOAP notes directly from doctor-patient transcripts.
- Enhance factual grounding and clinical relevance through structured knowledge integration.
- Maintain fluency, coherence, and interpretability of generated notes while remaining resource-efficient.

5 Baselines

To evaluate the effectiveness of our proposed methods, we compare against two baseline models that represent widely used paradigms in clinical text generation: T5-Base and T5-Large, both general-purpose sequence-to-sequence models. These baselines allow us to assess the impact of model capacity, parameter-efficient fine-tuning, and knowledge augmentation relative to standard modeling approaches.

5.1 T5-Base

T5-base is a general-purpose encoder–decoder Transformer model trained under a unified text-to-text framework, where all tasks are formulated as conditional generation problems. This formulation makes T5-base a natural baseline for dialogue-to-text generation tasks, including clinical summarization.

In this baseline, T5-base is fine-tuned to map raw doctor–patient dialogues directly to SOAP notes without any explicit structural constraints or external knowledge integration. The model receives the dialogue text as input and autoregressively generates the corresponding clinical note.

T5-base is chosen as a baseline because it represents a standard, widely used sequence-to-sequence architecture and provides a strong general-domain reference point. It allows us to evaluate how much structure and clinical relevance can be learned from supervised training alone, without relying on increased model capacity or external medical knowledge.

T5-base is fine-tuned using supervised learning with cross-entropy loss. The following hyperparameters are used:

- Optimizer: AdamW
- Learning rate: 1×10^{-3}
- Batch size: 8
- Number of epochs: 3
- Maximum input length: 1024 tokens
- Maximum output length: 512 tokens
- Decoding strategy (evaluation): Beam search with 3 beams

Hyperparameters are selected based on validation performance. No hyperparameters are tuned on the test set.

The T5-base model generates fluent text but does not consistently produce outputs in the SOAP format. Instead, it primarily performs unstructured summarization, with generated outputs resembling dialogue summaries or continuations rather than explicitly segmented clinical notes. This behavior highlights the limitations of naïve sequence-to-sequence fine-tuning for structured clinical documentation.

5.2 T5-Large

T5-large is a higher-capacity variant of the T5 encoder–decoder Transformer model, trained under the same unified text-to-text framework as T5-base. With substantially more parameters, T5-large is better suited to capturing long-range dependencies and complex output structures in conditional generation tasks.

In this baseline, T5-large is fine-tuned on the dialogue-to-SOAP note generation task using the same supervised learning setup as T5-base. The model takes raw doctor–patient dialogues as input and generates the corresponding SOAP note without incorporating external medical knowledge or additional structural constraints.

T5-large is chosen to evaluate the impact of increased model capacity on structured clinical note generation. By comparing T5-large with T5-base, we can assess whether scaling the model alone improves structural adherence and content organization, without changing the training objective or input representation.

T5-Large is fine-tuned using supervised learning with cross-entropy loss, using the hyperparameter settings established for T5-Base. Compared to T5-base, T5-large demonstrates improved awareness of the SOAP format and more frequently produces outputs with identifiable section structure. However, the generated content largely remains summarization-focused and does not consistently exhibit strong clinical reasoning. This suggests that increasing model capacity improves structural organization but is insufficient on its own to ensure factual grounding or clinically reliable note generation.

5.3 Dataset Splits and Evaluation Protocol

The dataset is split into training, validation, and test sets using a fixed random seed to ensure reproducibility. When an explicit validation split is not provided, 10% of the training data is held out for validation. The validation set is used exclusively for hyperparameter selection and early stopping and model selection.

The test set is used only once for final evaluation. No model parameters or decoding hyperparameters are tuned on the test set, ensuring a fair and unbiased comparison.

In summary, the chosen baselines represent two complementary perspectives:

- T5-base serves as a low-capacity baseline, highlighting the impact of model size and parameter efficiency on structured SOAP note generation.
- T5-large serves as a high-capacity baseline, estimating the upper-bound performance achievable by general-purpose sequence-to-sequence models.

Both baselines lack explicit structural enforcement and external knowledge grounding, providing a meaningful contrast to our proposed LoRA-based and knowledge-grounded approaches.

6 Approach

Approach 1: Baseline model

Our first approach evaluates whether a high-capacity, general-purpose sequence-to-sequence model can generate structurally coherent SOAP notes directly, without task-specific training. We examine the extent to which the model's pre-trained knowledge alone captures both the summarization objective and the canonical SOAP format when applied to clinical dialogue inputs.

We employ T5-Base and T5-Large, pretrained text-to-text Transformer models, and evaluate them directly on the dialogue-to-SOAP note generation task without task-specific training. The task is formulated as conditional generation, where the input is the raw multi-turn clinical dialogue and the model produces the corresponding SOAP note.

We reformulate the input using an instruction-style prompt explicitly specifying the desired output structure, e.g.:

"Generate a structured medical SOAP note from the following doctor–patient dialogue."

Let d denote a doctor–patient dialogue and s its reference SOAP note. The model learns a mapping:

$$\hat{s} = f_\theta(d)$$

where f_θ represents the T5-large model with parameters θ.

Although the pretrained models generate fluent and contextually relevant medical text, they fail to consistently produce outputs that conform to the SOAP format when evaluated directly. The generated sequences often resemble unstructured continuations or abstractive summaries of the dialogue, without clear separation between Subjective, Objective, Assessment, and Plan sections. This observation suggests that pretraining alone does not provide sufficiently strong document-level structural priors for reliable SOAP note generation.

In preliminary zero-shot evaluations, both T5-Base and T5-Large showed similar problems: they often failed to clearly separate the Subjective, Objective, Assessment, and Plan sections, especially for longer dialogues, and their output lengths and organization were inconsistent. These results suggest that pretrained capacity of the models alone is not enough to reliably generate structured clinical notes.

Approach 2: Parameter Efficient Fine Tuning with LoRA

The limitations observed in Approach 1 motivate the introduction of stronger task conditioning and a more controlled fine-tuning strategy. We hypothesize that parameter-efficient adaptation, combined with explicit task prompting, can improve structural adherence while maintaining computational efficiency and training stability.

Starting from the T5-Base and T5-Large, we apply Low-Rank Adaptation (LoRA) to selectively adapt the model. Training is restricted to the LoRA adapter parameters, with the base pretrained model frozen. Only the adapter weights are updated, enabling the model to capture task-specific patterns while preserving the general knowledge of the pretrained backbone.

This approach yields a marked improvement in format compliance, with a significant portion of generated outputs exhibiting explicit SOAP section headers. In these evaluations, T5-Base consistently underperformed T5-Large in maintaining the SOAP structure. It often omitted or merged sections and produced brief or inconsistent notes, especially for longer dialogues, showing difficulty in handling long-context clinical information. T5-Base also had higher variance across samples in output length, section order, and content coverage, making it an unreliable baseline. Therefore, T5-Base was excluded from subsequent experiments, and all further analyses focus on T5-Large.

However, two key failure modes persist in T5-Large:

1. **Incomplete structure**, where model occasionally omits the Assessment or Plan section, leaving notes otherwise complete.
2. **Clinical hallucinations**, including unsupported diagnoses, medications, or treatment plans not grounded in the source dialogue.

These findings suggest that while LoRA based fine-tuning improves structural alignment and efficiency, the model remains reliant on latent knowledge encoded during pretraining, limiting its ability to consistently enforce factual correctness.

Approach 3: Knowledge-Graph–Augmented Generation

Clinical SOAP note generation—particularly the Assessment and Plan sections—requires domain-specific reasoning that is often under-specified in raw doctor–patient dialogues. Language models fine-tuned solely on dialogue–note pairs tend to extrapolate based on latent correlations learned during pretraining, which frequently results in hallucinated diagnoses or treatment recommendations. We hypothesize that explicitly grounding the generation process in structured medical knowledge graphs (KGs) can constrain the model's output space, improve factual consistency, and enhance clinical plausibility.

System Formulation

Let a medical dialogue be represented as a sequence of utterances:

$$d = \{u_1, u_2, u_3, ...u_n\}$$

The objective is to generate a structured SOAP note *s* conditioned on both the dialogue *d* and a set of clinically relevant knowledge triples retrieved from a medical knowledge graph.

Step 1: Medical Entity Extraction

We first extract a set of medical entities from the dialogue using a biomedical named entity recognition (NER) model:

$$E = \{e_1, e_2, ...e_m\}$$

Each entity e_i corresponds to a clinically meaningful concept such as a symptom, disease, medication, or procedure, and is normalized to a canonical identifier (e.g., a UMLS Concept Unique Identifier). This normalization enables direct alignment between free-text dialogue and structured medical knowledge.

Step 2: Knowledge Graph Subgraph Retrieval

Let the medical knowledge graph be defined as $G = (V, R)$, where V denotes medical concepts and R denotes typed relations (e.g., associated_with, treats, causes). For each extracted entity $e_i \in E$, we retrieve a localized subgraph:

$$G_{e_i} = \{(e_i, r, v) r \in R, v \in V, dist(e_i, v) \leq k\}$$

where k is a fixed hop limit controlling subgraph size. The final knowledge context is constructed as:

$$G_E = \bigcup_{i=1}^{m} G_{e_i}$$

This retrieval strategy captures clinically relevant associations while limiting the inclusion of unrelated or noisy concepts.

Step 3: Knowledge Linearization and Prompt Construction

Since the underlying language model operates over textual inputs, the retrieved subgraph G_E is linearized into natural language statements. Each knowledge triple $(h, r, t) \in G_E$ is transformed into a sentence, for example: *"Chest pain is associated with myocardial infarction."*

The final model input is constructed as:

$$x = [Instruction; d; LinearizedKnowledge(G_E)]$$

where Instruction explicitly specifies structured SOAP note generation. The model then performs conditional generation:

$$\hat{s} = f_\theta(x)$$

Relative to the prior approaches, knowledge-augmented generation demonstrates higher structural completeness with consistent inclusion of all SOAP sections, substantially reduced hallucination frequency particularly in diagnostic assessments and treatment plans, and improved clinical plausibility and internal consistency.

Algorithm 1 Knowledge-Enhanced SOAP Note Generation

```
Require: Dialogue $d$, Knowledge Graph $G$, hop limit $k$
Ensure: Generated SOAP note $\hat{s}$
Extract medical entities $E \leftarrow \text{NER}(d)$
Initialize knowledge set $G_E \leftarrow \emptyset$
for each entity $e \in E$ do
    Retrieve subgraph $G_e \subseteq G$ within $k$ hops
    Update $G_E \leftarrow G_E \cup G_e$
end for
Linearize $G_E$ into textual knowledge statements
Construct prompt $x = [\text{Instruction}; d; \text{Knowledge}]$
Generate SOAP note $\hat{s} = f_\theta(x)$
return $\hat{s}$
```

These results indicate that explicit incorporation of structured medical knowledge provides an effective inductive bias, enabling the model to produce more factually grounded and clinically coherent SOAP notes.

The progression of approaches reveals a clear trade-off between model capacity, structural control, and factual grounding. Direct fine-tuning fails to enforce document structure, while parameter-efficient adaptation improves format compliance but not factual reliability. Knowledge-enhanced prompting offers the most robust performance, underscoring the importance of integrating external clinical knowledge for high-stakes medical text generation tasks.

7 Evaluation

The following sections detail the comprehensive metrics used to assess the quality of generated medical responses, focusing on linguistic quality, entity accuracy, and hallucination mitigation

7.1 Natural Language Generation Metrics

To evaluate the quality of the generated responses at the character level, we employ standard Natural Language Generation (NLG) metrics. Specifically, we utilize ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy). We calculate BLEU-1, BLEU-2, BLEU-3, and BLEU-4 to measure precision across different n-grams, as well as ROUGE-1, ROUGE-2, and ROUGE-L to assess the overlap with reference texts. These metrics help benchmark the general linguistic fluency and overlap with ground truth responses

7.2 Text Semantic Similarity

Beyond character overlap, we assess the overall semantic similarity between the generated text and the target text using BertScore. This metric leverages the pre-trained BERT model to compute the semantic distance between texts, providing a more robust measure of similarity than traditional n-gram matching by accounting for the actual semantics conveyed in the response.

7.3 LLM-as-a-Judge

Recognizing that conventional metrics often neglect the contextual background of historical dialogue, we construct a judge based on Large Language Models (LLMs) to evaluate response quality. Utilizing Gemini as an experienced doctor, we employ a specific prompt template to score responses.

7.3.1 Fluency Assessment

We evaluate the linguistic quality and professional tone of the generated response to ensure it adheres to clinical standards. The model assigns a score from 0 to 10, where 0 represents incoherent or grammatically fractured text, and 10 represents a response that is perfectly fluent, natural, and indistinguishable from one written by a human physician.

7.3.2 Consistency Verification

We evaluate if the response aligns logically with the subsequent standard physician responses, ensuring it contains the necessary key information and questions. This is also scored on a scale of 0 to 10, where a higher score indicates stronger alignment with the ground truth.

7.4 QA-based Factual Consistency

We measure factual consistency using a reference free question answering pipeline designed to verify that information in the generated note is grounded in the original clinical dialogue. To avoid model bias, we utilize a fine tuned distinct BART-large model, to generate factual questions based on the generated note. We then employ a separate RoBERTa-base QA model to answer these questions using the original clinical dialogue as the context. We define the factual consistency score as the percentage of questions where the answer derived from the source dialogue matches the answer derived from the generated note (measured via F1 overlap). This independent verification ensures the note contains only information supported by the source text.

8 Data

We consider two clinical data sources in this work: MIMIC and MediSOAP, each serving a distinct role in shaping our experimental design. While both datasets are relevant to SOAP note modeling, only MediSOAP was actively used in training and evaluation due to computational constraints.

The MIMIC dataset comprises large-scale, de-identified electronic health records from hospital encounters, containing diverse clinical notes authored by healthcare professionals across multiple specialties that reflect authentic documentation practices and exhibit substantial variability in structure, terminology, and detail. Although MIMIC represents a valuable source of real-world clinical text, its scale and unstructured nature impose significant computational demands. Preprocessing of the MIMIC dataset was attempted but could not be completed due to computational limitations. As a result, MIMIC was not used in training or evaluation experiments in this work.

The MediSOAP dataset was used as the primary data source for SOAP note generation. It consists of paired doctor–patient conversations and corresponding clinician-authored SOAP notes, explicitly segmented into Subjective, Objective, Assessment, and Plan components. MediSOAP provides clear structural supervision that enables models to learn section boundaries, content allocation, and clinically coherent note construction. In this work, MediSOAP was used for:

- Training and alignment of structured SOAP note generation models.
- Evaluation of section-wise completeness and clinical coherence.
- Analysis of dialogue-to-SOAP mapping behavior.

We selected MediSOAP due to its explicit SOAP annotation schema and manageable scale, which makes it well-suited for structured clinical generation under limited computational resources.

9 Results

We evaluate the proposed approaches using both qualitative inspection and quantitative evaluation. Results are organized to mirror the progression of methods described, highlighting how increasing levels of task conditioning and external knowledge grounding influence the quality of generated SOAP notes. We first present representative generated outputs to illustrate qualitative differences across approaches, followed by a comparative quantitative analysis using automatic and LLM-based evaluation metrics.

9.1 Qualitative Analysis of Generated SOAP Notes

Qualitative analysis was conducted to assess structural adherence, content allocation, and clinical plausibility across the different approaches. This analysis focuses on identifying recurring strengths and failure modes in generated SOAP notes.

Approach 1 (Baseline) produces fluent and contextually relevant medical text; however, outputs frequently fail to conform to the SOAP format. Generated notes often resemble unstructured summaries of the dialogue, with inconsistent or missing section boundaries.

Approach 2 (LoRA-Based Fine-Tuning) demonstrates improved structural compliance compared to the baseline. Many generated outputs explicitly include SOAP section headers, and Subjective and Objective content is generally well formed. Two consistent failure modes are observed in the outputs: incomplete coverage of later sections (Assessment and Plan) and clinical hallucinations, such as unsupported diagnoses or treatment recommendations not grounded in the dialogue. These observations suggest that while parameter-efficient fine-tuning improves format adherence, it does not adequately constrain factual reasoning.

Approach 3 (Knowledge-Graph–Augmented Generation) yields the most structurally complete and clinically coherent SOAP notes. All four SOAP sections are consistently generated, with improved logical flow and clearer separation of clinical content. Qualitative inspection reveals a marked reduction in hallucinated diagnoses and treatment plans, particularly within the Assessment and Plan sections. Generated recommendations are more directly supported by information present in the dialogue or retrieved medical knowledge, supporting the effectiveness of explicit knowledge grounding.

9.2 Impact of Knowledge-Graph Augmentation

To evaluate the effect of knowledge integration in Approach 3, we examine its impact on factual consistency and clinical plausibility in generated SOAP notes. Conditioning generation on both the dialogue and linearized knowledge graph facts promotes medically grounded diagnostic reasoning and treatment planning.

Compared to Approaches 1 and 2, knowledge-augmented generation demonstrates improved internal consistency, especially in clinically sensitive sections. Assessment statements more frequently align with reported symptoms, and treatment plans are less speculative and more conservative. These findings indicate that external knowledge serves as an effective inductive bias, constraining the model's output space and mitigating over-reliance on latent correlations learned during pretraining.

9.3 Quantitative Comparison Across Approaches

Table 1 presents the quantitative evaluation across all three approaches. Metrics include surface-level similarity (ROUGE-L, BLEU), semantic alignment (BERTScore-F1), factual reliability (Hallucination), and LLM-judge scores assessing consistency and fluency.

The baseline (T5-Large) achieves moderate semantic similarity (BERTScore 0.713) and surface-level overlap (ROUGE-L 0.393, BLEU 0.1843), but exhibits high hallucination (0.55) and lower consistency (5.26) and fluency (6.92), reflecting poor structural adherence and occasional unsupported clinical statements.

The LoRA-based model (PEFT T5-Large) improves semantic alignment (BERTScore 0.764) and slightly increases ROUGE-L (0.405), while hallucination decreases to 0.43 and consistency and fluency improve to 5.97 and 7.31, respectively. Despite these gains, factual reliability and BLEU (0.1611) indicate persistent gaps in fully capturing dialogue content.

The knowledge-graph–augmented model (FT-KG T5-Large) demonstrates the strongest performance, with highest ROUGE-L (0.438) and BERTScore (0.873), lowest hallucination (0.21), and markedly higher section-wise consistency (7.84) and competitive fluency (7.73). These results indicate that explicit knowledge grounding effectively improves structural completeness, factual reliability, and clinical plausibility.

Overall, the results show a clear progression: naive fine-tuning offers moderate similarity but poor structure, LoRA-based adaptation improves fluency and partial factual reliability, and knowledge-graph augmentation achieves the most robust and clinically coherent SOAP note generation.

10 Error Analysis

To better understand the limitations of our baseline models and proposed approach, we conducted a qualitative manual error analysis on a small set of representative dialogue examples. Rather than relying solely on automatic metrics, we examined generated SOAP notes in detail and compared them against reference notes to identify recurring failure modes. This analysis focuses on three representative samples that illustrate common structural and semantic errors observed across models.

10.1 Error Types in Baseline Models

Across the baseline models, we observe several consistent failure patterns.

Lack of structural separation: In fine-tuned T5-base, generated outputs often resemble unstructured summaries of the dialogue rather than clearly segmented SOAP notes. Subjective and Objective information is frequently merged, and Assessment and Plan sections are either missing or overly generic. This behavior is evident in several samples, where important clinical observations that belong in the Objective section are incorrectly placed under Subjective.

Over-summarization and omission of key details: Even when structure is partially present, fine-tuned models tend to compress complex clinical narratives into high-level summaries. In some samples, while major surgical interventions are mentioned, the temporal sequence of diagnosis, surgery, and postoperative management is poorly represented. This suggests difficulty in handling long, multi-stage clinical encounters.

Hallucinated or weakly supported diagnoses: Fine-tuned models occasionally introduce diagnoses or differential considerations that are not clearly supported by the dialogue. For example, in a sample, the Assessment includes irrelevant differential diagnoses such as cardiac devices or implants, which are not mentioned in the dialogue or reference note. This indicates reliance on spurious correlations rather than grounded clinical reasoning.

10.2 Error Types in the Knowledge-Grounded Approach

While the knowledge-grounded approach improves overall structure and completeness, it also exhibits distinct limitations.

Entity overgeneralization: In some cases, the entity extraction and knowledge retrieval pipeline introduces concepts that are only weakly related to the patient's condition. In a sample, the generated Assessment labels the condition as "chronic mesenteric gastroenteritis," which does not appear in the reference and likely arises from loosely related gastrointestinal entities retrieved from the knowledge graph.

Blurring of diagnostic granularity: In a sample, although the generated SOAP note correctly identifies hearing loss and otosclerosis, the Assessment section remains overly broad, describing mixed conductive and sensorineural hearing loss without explicitly emphasizing bilateral otosclerosis and right-sided semicircular canal dehiscence as primary diagnoses. This suggests difficulty in prioritizing clinically salient concepts when multiple related entities are present.

Propagation of upstream extraction errors: Errors in named entity recognition can propagate downstream into the generated SOAP note. For example, fragmented or partial entities such as truncated diagnostic terms may result in awkward or imprecise phrasing in the final output, affecting clarity and clinical precision.

10.3 Summary

From the manual analysis of representative generated SOAP notes, several recurring error patterns were identified across models:

Table 1: Evaluation Metrics

Metric	Mean (FT-KG)	95% CI (FT-KG)	Mean (FT)	95% CI (FT)	Mean (T5)	95% CI (T5)
ROUGE-L	0.438	[0.423, 0.454]	0.405	[0.389, 0.420]	0.393	[0.376, 0.409]
BLEU	0.243	[0.221, 0.266]	0.161	[0.142, 0.182]	0.184	[0.160, 0.207]
BERTScore	0.873	[0.857, 0.889]	0.764	[0.746, 0.781]	0.713	[0.695, 0.730]
Hallucination Rate	0.210	[0.181, 0.244]	0.430	[0.395, 0.468]	0.550	[0.508, 0.589]
Consistency	7.84	[7.41, 8.27]	5.97	[5.52, 6.39]	5.26	[4.78, 5.71]
Fluency	7.73	[7.29, 8.14]	7.31	[6.92, 7.68]	6.92	[6.48, 7.31]

- Clinical Accuracy Errors:
  - Occasional inaccuracies in reported medications or treatment details
  - Confusion in sequencing of clinical events and procedures
  - Infrequent errors in reporting laboratory values or examination findings
- Section-Level Organization Errors:
  - Overlap of information across multiple SOAP sections
  - Misplacement of diagnostic statements within the Plan section
  - Variability in formatting consistency across generated notes
- Incomplete Information Coverage:
  - Missing or vague follow-up instructions in some cases
  - Partial omission of less prominent symptoms
  - Inconsistent representation of past medical history

These patterns indicate that, although the models demonstrate strong language generation capabilities, maintaining precise clinical structure and comprehensive coverage remains a challenge, particularly in complex or multi-stage clinical encounters.

11 Contributions of group members

The work for this project was divided evenly among the three group members, with each person contributing to both the implementation and the final report.

- Sreehitha: Worked on data collection and preprocessing, including cleaning the dialogue data and formatting inputs and outputs for SOAP note generation. Also built and trained the T5-base LoRA fine-tuned model and performed evaluation of all models using relevant performance metrics.

- Naveen: Focused on building and training the models, including the T5-Large and additional LoRA fine-tuned variants. Managed the training workflow, conducted systematic hyperparameter tuning using the validation set, and performed comparative analysis across model variants to optimize performance.

- Dhanush: Implemented the knowledge graph integration pipeline by performing medical entity extraction from clinical dialogues, retrieving clinically relevant subgraphs from the knowledge graph, and linearizing the extracted knowledge into structured textual representations for prompt construction.

- All members: Contributed to error analysis and report writing.

12 Conclusion

In this project, we explored multiple approaches for automatic generation of structured SOAP notes from doctor–patient dialogues, ranging from standard sequence-to-sequence fine-tuning to parameter-efficient adaptation and knowledge-grounded generation. This work provided practical experience with modern NLP techniques and highlighted the challenges of applying them to structured clinical documentation.

A key takeaway is that generating fluent medical text does not guarantee structural correctness or factual reliability. Although pretrained language models were able to produce coherent outputs, consistently enforcing the SOAP format proved more difficult than expected. While LoRA-based fine-tuning improved format compliance, hallucinated clinical details, particularly in the Assessment and Plan sections, remained a significant challenge.

The most effective results were achieved through knowledge-grounded generation using medical knowledge graphs. Incorporating structured clinical knowledge led to more complete SOAP notes and a noticeable reduction in hallucinations, demonstrating the importance of domain grounding for clinical text generation.

Contrary to the assumption that maximizing external knowledge always yields better results, we found that a filtered Knowledge Graph (KG) outperformed a comprehensive one. Surprisingly, creating intentional gaps in the provided knowledge forced the LLM to shift from passive information retrieval to active clinical reasoning and more robust SOAP note generation.

Future work could focus on tighter integration of knowledge graphs with language models, improved evaluation methods for factual consistency, and extensions to longer or real-time clinical encounters. Overall, this project shows that reliable clinical note generation requires both strong language modeling and explicit incorporation of domain knowledge.

13 GitHub Repository:

Click here

14 AI Disclosure

- Did you use any AI assistance to complete this proposal? If so, please also specify what AI you used.
  - Yes, ChatGPT

If you answered yes to the above question, please complete the following as well:

- If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which part of the proposal is associated with which prompt.
  - Clean up the following Python code: remove unused imports and variables, standardize names, fix formatting and indentation, add brief comments, and remove any temporary or debugging code while keeping the functionality intact.
  - Rewrite the following paragraph to improve clarity, logical flow, and academic style. Maintain the original meaning and content while enhancing readability, sentence structure, and formal tone.
  - Review the following report draft for clarity, coherence, grammar, and academic style. Suggest improvements to sentence structure, paragraph flow, and formal tone, and highlight any sections that may need additional explanation or reorganization.
- **Free response:** For each section or paragraph for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good output, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to generate new text, check your own ideas, or rewrite text?
  - AI was a supportive tool, enhancing efficiency, improving readability, and ensuring consistency, but human oversight was essential to verify accuracy, maintain technical correctness, and align outputs with project-specific goals.

References

Abacha, A. B., wai Yim, W., Michalopoulos, G., and Lin, T. (2023). An investigation of evaluation metrics for automated medical note generation.

Dahlberg, A., Käennemi, T., Winther-Jensen, T., Tapiola, O., Luisto, R., Puranen, T., Gordon, M., Sanmark, E., and Vartiainen, V. (2025). Measuring the quality of ai-generated clinical notes: A systematic review and experimental benchmark of evaluation methods.

Enarvi, S., Amoia, M., Del-Agua Teba, M., Delaney, B., Diehl, F., Hahn, S., Harris, K., McGrath, L., Pan, Y., Pinto, J., Rubini, L., Ruiz, M., Singh, G., Stemmer, F., Sun, W., Vozila, P., Lin, T., and Ramamurthy, R. (2020). Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In Bhatia, P., Lin, S., Gangadharaiah, R., Wallace, B., Shafran, I., Shivade, C., Du, N., and Diab, M., editors, *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online. Association for Computational Linguistics.

Gao, Y., Li, R., Croxford, E., Caskey, J., Patterson, B. W., Churpek, M., Miller, T., Dligach, D., and Afshar, M. (2025). Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI*, 4:e58670.

Krishna, K., Khosla, S., Bigham, J., and Lipton, Z. C. (2021). Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4958–4972, Online. Association for Computational Linguistics.

Leong, H. Y., Gao, Y. F., Ji, S., Kalaycioglu, B., and Pamuksuz, U. (2025). A gen ai framework for medical note generation.

Li, Y., Wu, S., Smith, C., Lo, T., and Liu, B. (2025). Improving clinical note generation from complex doctor-patient conversation.