Skin-SOAP: A Weakly Supervised Framework for Generating Structured SOAP Notes

Sadia Kamal¹, Tim Oates¹, Joy Wan²

¹Department of Computer Science, University of Maryland, Baltimore County ²Department of Dermatology, Johns Hopkins University School of Medicine, Baltimore, MD sadia1402@umbc.edu, oates@cs.umbc.edu, jwan7@jhmi.edu

Abstract

Skin carcinoma is the most prevalent form of cancer globally, accounting for over \$8 billion in annual healthcare expenditures. Early diagnosis, accurate and timely treatment are critical to improving patient survival rates. In clinical settings, physicians document patient visits using detailed SOAP (Subjective, Objective, Assessment, and Plan) notes. However, manually generating these notes is labor-intensive and contributes to clinician burnout. In this work, we propose skin-SOAP, a weakly supervised multimodal framework to generate clinically structured SOAP notes from limited inputs, including lesion images and sparse clinical text. Our approach reduces reliance on manual annotations, enabling scalable, clinically grounded documentation while alleviating clinician burden and reducing the need for large annotated data. Our method achieves performance comparable to GPT-4o, Claude, and DeepSeek Janus Pro across key clinical relevance metrics. To evaluate this clinical relevance, we introduce two novel metrics Med-ConceptEval and Clinical Coherence Score (CCS) which assess semantic alignment with expert medical concepts and input features, respectively.

1 Introduction

Skin cancer remains one of the most common and deadliest cancers in the United States, with approximately 9,500 new cases diagnosed daily [Rogers et al., 2015], highlighting the crucial role of clinical documentation as the foundation for effective communication, accurate diagnosis, and informed treatment planning. Structured formats like SOAP (Subjective, Objective, Assessment, Plan) notes are widely adopted in the United States to ensure consistency in recording patient encounters and minimizing communication errors among healthcare professionals [Schloss and Konam, 2020]. However, generating these notes is labor-intensive and time-consuming, which reduces direct patient interaction time and significantly contributes to physician burnout [Li et al., 2024] [Biswas and Talukdar, 2024].

Automating the generation of SOAP notes presents a promising solution to reduce administrative burden, improve documentation consistency, and allow clinicians to focus more on patient centred care. Recent advances in large language models (LLMs) have enabled impressive progress in medical natural language processing tasks, including clinical summarization, question answering [Singhal *et al.*, 2023], lab report interpretation [He *et al.*, 2024], and deidentification of sensitive information [Yang *et al.*, 2023]. These models can produce coherent and fluent clinical narratives, making them useful tools for medical documentation. However, general purpose LLMs often lack the domain-specific reasoning required for clinical settings, struggle to understand subtle medical context, and are generally limited to text based inputs. Their performance is further constrained in tasks such as structured note generation, especially when applied to domains like dermatology.

Existing methods for automated SOAP note generation, such as K-SOAP [Li et al., 2024], rely heavily on extensive doctor-patient dialogues and large-scale annotated datasets resources that are particularly limited in dermatology and skin lesion documentation [Wei et al., 2024]. Moreover, capturing both the visual features of skin conditions and the underlying clinical reasoning in a structured format remains a major challenge. To address these limitations, we propose skin-SOAP, a novel weakly supervised multimodal framework that generates structured SOAP notes from limited inputs including lesion images and sparse clinical text. Unlike prior approaches focused solely on text-based SOAP generation or dermatologic diagnosis, our method uniquely integrates retrieval-augmented clinical knowledge, weak supervision, and multimodal synthesis to enable domain-aligned documentation without requiring large-scale annotations. We also introduce two novel metrics to evaluate clinical and semantic quality: MedConceptEval and Clinical Coherence Score (CCS), which go beyond traditional NLP metrics by assessing alignment with defined concepts and feature consistency. By leveraging domain-guided retrieval and pseudolabeling, our framework produces clinically relevant SOAP notes with minimal supervision, offering a scalable solution for dermatology and broader healthcare applications. Our main contributions are:

 We developed skin-SOAP, a weakly supervised multimodal framework for structured SOAP note generation from lesion images and limited clinical text.

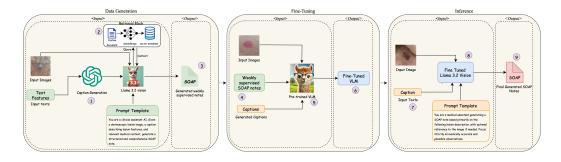


Figure 1: Overview of the proposed skin-SOAP framework, consisting of data generation, fine-tuning, and inference phases.

- We introduced two novel evaluation metrics: MedConceptEval and Clinical Coherence Score (CCS) to evaluate semantic alignment with clinical concepts and input consistency.
- We performed statistical analysis using two-way ANOVA to quantify the effects of SOAP sections and lesion types on semantic similarity scores.
- We conducted qualitative evaluation using an LLM-as-a-Judge framework (Flow-Judge-v0.1) to assess structure, readability, completeness, and clinical relevance.

2 Related Work

Artificial intelligence has been widely applied in the medical domain for tasks such as medical image captioning [Cheng et al., 2025], clinical text summarization [Van Veen et al., 2023], discharge note generation [Jung et al., 2024], and medical question answering [Yan et al., 2024]. Recent efforts have focused on generating clinical notes, especially SOAP notes, from Electronic Health Records (EHRs) and doctor-patient conversations. In [Li et al., 2024], large language models (LLMs) are fine-tuned to generate K-SOAP notes, while [Chen and Hirschberg, 2024] evaluated generative models across general and SOAP-specific formats. Similar work has explored domains like pediatric rehabilitation [Amenyo et al., 2025] and fine-tuning LLMs for efficient SOAP note generation [Leong et al., 2024].

Earlier approaches to SOAP note generation, such as [Schloss and Konam, 2020; Singh et al., 2023], used sequence-to-sequence models and transformer architectures trained on large dialogue datasets. Although general-purpose LLMs have improved clinical NLP tasks [Singhal et al., 2023; Yang et al., 2023], they often lack domain-specific reasoning and structured multimodal output capabilities. While methods like [Ramprasad et al., 2023] have attempted to improve faithfulness in text-based settings, these methods heavily rely on large annotated dialogue corpora, which are difficult to obtain due to privacy concerns [Chen and Hirschberg, 2024], and they remain text-centric without leveraging multimodal information, which is critical to fields like dermatology. SkinGPT-4 [Zhou et al., 2024] recently demonstrated that integrating clinical images and text using a multimodal LLM improves dermatological diagnostic reasoning. However, it is primarily designed for diagnostic prediction and lacks structured documentation capabilities. Specifically, it does not generate clinically formatted SOAP notes or support section-level reasoning required for clinical documentation. In contrast, our work focuses on end-to-end generation of structured SOAP notes from limited multimodal inputs, introducing weak supervision and domain-guided retrieval to ensure both clinical reliability and scalability.

3 Methodology

We proposed a three-phase weakly supervised multimodal framework for generating clinically structured SOAP notes from limited dermatologic inputs. As illustrated in Fig. 1, our method consists of: (1) a data generation phase to synthesize weakly supervised SOAP notes using generative captioning and retrieval-augmented knowledge integration, (2) a fine-tuning phase to adapt a vision-language model using the synthesized notes, and (3) an inference phase to generate high-quality structured SOAP notes from new patient data.

3.1 Dataset

We use the PAD-UFES-20 dataset [Pacheco *et al.*, 2020], which consists of 2,298 dermoscopic images along with structured metadata for 1,641 skin lesions collected from 1,373 patients. The lesions are classified into six types: Basal Cell Carcinoma (BCC), Melanoma (MEL), Squamous Cell Carcinoma (SCC, including Bowen's disease), and three non-cancerous conditions: Actinic Keratosis (ACK), Seborrheic Keratosis (SEK), and Nevus (NEV).

Approximately 58% of the samples are confirmed through biopsy, while the remaining cases are clinically diagnosed based on expert consensus. Each lesion is linked to a CSV file containing 26 structured clinical features, which include lesion characteristics (such as size and anatomical location), patient demographics, symptom information (such as itching, bleeding, or changes in appearance), and family medical history.

3.2 Data Generation

Due to the limited availability of large-scale annotated SOAP note datasets in dermatology, we employ a weak supervision strategy to synthesize training data. Each training sample consists of a lesion image paired with structured clinical features, such as lesion diameter, biopsy status, and symptom descriptors. We first use GPT-3.5 [Brown *et al.*, 2020] to

generate a clinical caption summarizing these structured attributes into a coherent description of the lesion. To improve the clinical relevance and factual reliability of the generated notes, we design a retrieval-augmented generation framework [Lewis et al., 2020]. The generated caption is used as a query to retrieve semantically relevant passages from a curated vector database built with ChromaDB. This database indexes document chunks extracted from authoritative medical sources, including the South Texas Skin Cancer Institute [Skin Cancer Specialists, 2024], the National Cancer Institute [National Cancer Institute, 2024], the American Cancer Society [American Cancer Society, 2024], and the UK's National Health Service [NHS, 2024], covering lesion types, diagnostic criteria, symptomatology, and treatment guidelines. The retrieved context are concatenated with the original caption and provided as input to the pre-trained Vision-LLaMA 3.2 model, guided by a structured prompting template that encourages SOAP format outputs. This design addresses common limitations of pre-trained language models, such as outdated knowledge and hallucinated reasoning, and enables the model to generate reliable and clinically grounded, weakly supervised SOAP notes.

3.3 Fine-Tuning

We fine-tune the Vision-LLaMA 3.2 model using the synthesized dataset, where the lesion image and generated caption are treated as multimodal inputs, and the weakly supervised SOAP note serves as the training target. The model is optimized to produce structured outputs following the standard SOAP note format illustrated in Fig 2.

Component	Subcategories	Description
Subjective (S)	Chief Complaint	The main reason for the patient's visit.
	Medical History	Relevant medical history, including previous diagnoses, treatments, or family history.
Objective (O)	On Examination	Detailed findings from the physical and dermoscopic examination, such as dots/globules, pigmented network etc.
	Observation	Additional clinical observations, including lesion shape, size, or symmetry.
Assessment (A)	Investigation	Results of diagnostic tests or imaging (e.g., dermoscopic findings).
	Diagnosis	The clinician's interpretation of the findings, identifying the primary condition (e.g., melanoma, atypical nevus).
	Assessment	A summary of the condition, including differential diagnoses and overall severity or staging (if applicable).
Plan (P)	Treatment Plan	Recommended actions such as excision, medication, or lifestyle modifications.
	Patient Education	Patient Education & Guidance provided to the patient regarding self-care and monitoring.

Figure 2: Structured SOAP note components.

Parameter-Efficient Fine-Tuning (PEFT)

To reduce computational costs without compromising model performance, we employ Parameter-Efficient Fine-Tuning (PEFT) strategies. Specifically, we use Quantized Low-Rank Adaptation (QLoRA) [Dettmers *et al.*, 2023], which introduces trainable low-rank matrices into transformer layers. Instead of updating all model parameters, QLoRA injects low-rank decompositions into specific modules such as the query, key, value, and output projections. The low-rank adaptation update is defined as:

$$\Delta W = AB \tag{1}$$

where $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times d}$ are the learnable low-rank matrices with $r \ll d$.

The adapted weight matrix during fine-tuning becomes:

$$W' = W + \Delta W = W + AB \tag{2}$$

where W represents the original frozen weights, and ΔW captures the trainable adaptation. We apply LoRA to the q, k, v, o, gate, up, and down modules of the Vision-LLaMA model, allowing efficient adaptation without requiring full parameter updates.

Supervised Fine-Tuning (SFT)

Supervised Fine-Tuning (SFT) further adapts the model by explicitly teaching it to map dermatological inputs to structured SOAP notes based on weakly supervised ground-truth examples. Unlike pre-training, which loosely guides the model, SFT provides complete input-output pairs, enabling the model to learn structured clinical reasoning patterns.

The SFT training objective minimizes the cross-entropy loss between the predicted SOAP note \hat{y} and the target SOAP note y given the multimodal input x (image and caption):

$$\mathcal{L}_{SFT} = -\sum_{i=1}^{n} y_i \log(\hat{y}_i)$$
 (3)

where \mathcal{L}_{SFT} denotes the supervised fine-tuning loss, n is the number of tokens in the SOAP note, y_i is the ground-truth token at position i, \hat{y}_i is the predicted probability for the i-th token, and x represents the multimodal input consisting of the lesion image and its corresponding clinical caption.

This loss function ensures that the generated notes are not only linguistically coherent but also structurally accurate and clinically reliable.

3.4 Training Setup

The fine-tuning of the Vision-LLaMA 3.2 model is performed using Quantized Low-Rank Adaptation (QLoRA) with a lowrank dimension r=8, a scaling factor $\alpha=16$, and no dropout applied. LoRA modules are inserted into the query, key, value, output, gate, up, and down projections within the model's transformer blocks. Supervised Fine-Tuning (SFT) is conducted with a batch size of 8, using gradient accumulation over 4 steps and 10 warmup steps at the beginning of training. Fine-tuning is performed for 500 epochs with a linear learning rate scheduler, starting from an initial learning rate of 2×10^{-4} . The optimizer used is AdamW with 8-bit precision to improve memory efficiency. Pretraining to generate weakly supervised SOAP notes took approximately 9 hours, while the final fine-tuning process was completed in 1.5 hours on an NVIDIA A100 GPU with 80 GB of VRAM. Mixed-precision training with bfloat16 (bf16) format was employed to optimize memory utilization throughout the training process.

3.5 Inference

At inference time, the fine-tuned Vision-LLaMA model receives a lesion image along with its corresponding clinical features, which were first converted into a clinical caption. The model then generates a structured SOAP note. Because the model has been fine-tuned on weakly supervised yet clinically reliable data, it generalizes effectively to new cases. This enables scalable and structured generation of dermatology SOAP notes, even in environments where expert annotations are limited or unavailable.

4 Evaluation

We evaluate the generated SOAP notes using both quantitative and qualitative methods. For quantitative evaluation we compare the generated notes against expert-annotated ground truth using standard and clinical-domain NLP metrics. Additionally, we introduce two clinical relevance metrics: **Med-ConceptEval** and **Clinical Coherence Score** (**CCS**). The qualitative evaluation is based on the LLM-as-a-Judge approach [Zheng *et al.*, 2023] with Flow-Judge-v0.1, where the model evaluates structure, readability, medical relevance, and consistency with conventional SOAP note standards.

4.1 Quantitative Evaluation

MedConceptEval

To evaluate clinical relevance, we introduce MedConceptEval, a semantic evaluation framework designed to assess the alignment of each SOAP note section with clinically validated concept sets. These concept sets, referred to as descriptor banks, are curated for six major dermatological classes derived from reputable clinical resources [Mayo Clinic Staff, A language model is employed to extract relevant medical concepts and construct disease-specific keyword sets. Each section of the generated SOAP note is encoded using ClinicalBERT, and cosine similarity is computed against the corresponding descriptor bank. For each dermatological class, we calculate both average and maximum similarity scores per section across five representative cases. This method provides a robust, interpretable, and clinically grounded evaluation of the generated SOAP notes, ensuring alignment with disease-specific clinical terminology beyond surface-level keyword matching.

In Table 1, the Assessment and Plan sections consistently achieved higher average similarity scores compared to Subjective and Objective sections, indicating stronger alignment with medically relevant concepts. Conditions like Melanoma and Nevus demonstrated particularly high alignment, with maximum similarity values exceeding 0.90 in the Plan section, suggesting that generated notes closely matched expert medical descriptors. However, slightly lower scores were observed for SCC and BCC, highlighting potential areas for improvement in capturing subtle clinical features for these cases.

Statistical Significance Analysis: We conducted a Two-Way ANOVA to evaluate the effects of SOAP note sections and dermatological conditions on the average MedConceptE-val similarity scores. The results indicated a statistically significant main effect of the SOAP section (F(3,15))

Condition	Section	Avg Similarity	Max Similarity	
	Subjective	0.7768	0.8746	
Seborrheic Keratosis	Objective	0.7952	0.8648	
(SEK)	Assessment	0.8168	0.8680	
	Plan	0.7764	0.8310	
	Subjective	0.7786	0.8468	
Navyo (NEV)	Objective	0.7786	0.8598	
Nevus (NEV)	Assessment	0.8006	0.8708	
	Plan	0.8626	0.8976	
	Subjective	0.7790	0.8624	
Malamama (MEL)	Objective	0.7952	0.8676	
Melanoma (MEL)	Assessment	0.8234	0.8770	
	Plan	0.8526	0.9036	
	Subjective	0.7354	0.8182	
Actinic Keratosis (ACK)	Objective	0.7844	0.8554	
Actilic Keratosis (ACK)	Assessment	0.8092	0.8458	
	Plan	0.8400	0.8864	
	Subjective	0.7846	0.8360	
Squamous Cell	Objective	0.7754	0.8360	
Carcinoma (SCC)	Assessment	0.7802	0.8398	
	Plan	0.7854	0.8596	
	Subjective	0.7740	0.8464	
Basal Cell Carcinoma	Objective	0.7658	0.8182	
(BCC)	Assessment	0.7882	0.8220	
	Plan	0.7738	0.8212	

Table 1: **MedConceptEval:** Semantic similarity between SOAP sections and curated clinical concept sets across six dermatologic conditions.

4.31, p=0.022), demonstrating that semantic alignment varied across the Subjective, Objective, Assessment, and Plan sections. However, the effect of lesion type was not significant (F(5,15)=1.44, p=0.268), suggesting that the model's semantic consistency was maintained across different skin conditions. Here, $F(d_1,d_2)$ represents the F-statistic with d_1 and d_2 degrees of freedom for between-group and within-group variance, respectively. These results underscore the greater influence of SOAP note structure over disease type in shaping semantic alignment with clinical concepts.

Clinical Coherence Score (CCS)

We introduce the Clinical Coherence Score (CCS), a metric that evaluates the semantic alignment between the caption and the structured SOAP note sections. To compute this score, we utilize ClinicalBERT based contextual embeddings, which are specifically designed to capture clinical terminology and relationships. Unlike traditional lexical overlap metrics (e.g., ROUGE, BLEU), CCS provides a domain-specific signal of how faithfully the model retains and reflects the original clinical information in each section of the output. It does so by embedding both the caption and each SOAP section into a shared semantic space using ClinicalBERT and computing their cosine similarity.

We have compared the CCS for both our generated notes and dermatologists written notes as shown in table 2 which indicates that the LLM-generated SOAP notes exhibit consistently higher semantic alignment with the captions compared to the dermatologist provided handwritten notes. These scores suggests that our model is more skillful at capturing the terminology and phrasing present in the caption, likely due to its training in a shared embedding space with similar

data. While, clinician written notes may introduce complex language or clinical reasoning not explicitly present in the caption, resulting in slightly lower CCS despite being more accurate.

This opens up an opportunity in research that reveals that LLM, still does not fully replicate the depth and complexity of hand-written notes. Bridging this gap could enable future models to match or even complement the quality, consistency, and clinical reasoning present in expert written documentation.

Statistical Significance Analysis: To examine the impact of different note type and SOAP structure on semantic alignment, we performed a Two-Way ANOVA on CCS values across three lesion cases. The analysis revealed a statistically significant main effect of note type (F(1,19) = 57.53, p <0.001), indicating that generated notes exhibited substantially higher semantic alignment with the caption than ground truth notes. In contrast, the effect of SOAP section was not significant (F(3,19) = 1.47, p = 0.254), suggesting that the variation in alignment did not differ meaningfully across the Subjective, Objective, Assessment, and Plan sections. These results shows the model generated notes are consistently coherent outputs across all sections of the SOAP note, surpassing human-written references in terms of semantic alignment with the input caption. Given that the reference notes were written by a domain expert, this finding suggests a promising avenue for future research, where LLM generated clinical documentation could eventually match or even complement the quality and consistency of expert-written notes.

SOAP	Gei	nerated no	otes	Ground Truth notes				
	Case 1	Case 2	Case 3	Case 1	Case 2	Case 3		
Subjective	0.9308	0.9306	0.9318	0.3335	0.7184	0.7475		
Objective	0.9168	0.8891	0.9334	0.5448	0.4235	0.7239		
Assessment	0.9178	0.8933	0.9208	0.4926	0.7266	0.7299		
Plan	0.8795	0.8842	0.8981	0.3899	0.5371	0.3368		
Average	0.9112	0.8993	0.9210	0.4402	0.6014	0.6501		

Table 2: Clinical Coherence Score: Semantic alignment between the caption and each SOAP section for both generated and ground truth SOAP notes across three lesion cases.

Furthermore, we obtained expert-annotated SOAP notes for three different lesion images from a board-certified dermatologist. Each lesion image, along with its corresponding caption, was provided to the dermatologist to generate structured SOAP notes. These expert-written notes were treated as ground truth (reference) and are compared with our generated notes (candidate) to evaluate their alignment using a range of lexical and semantic metrics, including BLEU [Post, 2018], ROUGE [Lin, 2004], METEOR [Lewis *et al.*, 2020], CHRF++ [Popović, 2017], BERT Score [Zhang *et al.*, 2019], and ClinicalBERT Score [Huang *et al.*, 2019].

ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measures n-gram overlap between generated and reference notes. We report ROUGE-1 (unigrams), ROUGE-2 (bigrams), and ROUGE-L (longest common subsequence), capturing both lexical recall and structural consistency.

METEOR

METEOR (Metric for Evaluation of Translation with Explicit ORdering) incorporates synonymy and word stemming to better handle clinical paraphrasing. It is more tolerant to surface-level variations and is especially useful when evaluating partially reworded notes.

CHRF++

CHRF++ computes F-scores over character n-grams, offering robustness to minor lexical differences, such as pluralization, typos, or morphological variations (e.g., "carcinoma" vs. "carcinomas"). It is particularly well-suited for clinical domains where such variations are common.

BERT Score

BERT Score uses contextual embeddings from a pretrained language model (typically BERT-base) to compute semantic similarity between reference and generated notes. It evaluates whether generated tokens align semantically with reference tokens beyond exact matches.

Clinical BERT Score

To capture clinically grounded semantics, we use BERT with Clinical BERT, a model pretrained on large-scale clinical notes (e.g., MIMIC-III). This allows for more accurate evaluation of clinical relevance, capturing medical synonyms, abbreviations, and contextual language common in SOAP documentation.

In the Table 3, skin-SOAP demonstrates consistently strong performance, especially on METEOR and CHRF++, indicating high fluency and surface-level coherence. While GPT-40 achieved slightly higher ROUGE-L in Cases 1 and 3, our method outperformed it in ROUGE-2 and METEOR, which better reflect phrasal and semantic overlap. Notably, our ClinicalBERT F1 scores are either the highest or on par with the best-performing model in each case, underscoring our model's superior alignment with clinical concepts. Here, the difference between the BERT and ClinicalBERT scores further highlights how clinically tuned models capture semantic relevance more effectively than general-purpose transformer models. Overall, our results are comparable with the other models and these results confirm the effectiveness of our weakly supervised multimodal approach for producing high-quality and clinically grounded SOAP notes.

4.2 Qualitative Evaluation

LLM-as-a-Judge

We conducted a qualitative evaluation using an LLM-as-Judge framework, as obtaining human verifiers is difficult, expensive. Specifically, we employed **Flow-Judge-v0.1** to assess each generated SOAP note across four criteria. According to the HuggingFace Judge Arena: Benchmarking LLMs as Evaluators, Flow-Judge (3.8B) is an open-source model developed by Flow AI, achieves the highest ELO rating of 1335, outperforming larger proprietary models such as GPT-40 (1320), Claude 3 Opus (1268), and Meta Llama 3.1 405B (1267). In our study, Flow-Judge rated each note on a 5-point Likert scale (1 = Poor, 5 = Excellent) based on the following evaluation criteria:

Metrics	Case 1			Case 2			Case 3					
	skin-SOAP	GPT-40	Janus Pro	Claude	skin-SOAP	GPT-40	Janus Pro	Claude	skin-SOAP	GPT-40	Janus Pro	Claude
ROUGE-1	0.396	0.324	0.348	0.294	0.4183	0.399	0.397	0.460	0.3999	0.4360	0.351	0.480
ROUGE-2	0.0827	0.0326	0.1228	0.0281	0.125	0.0966	0.1122	0.0671	0.0939	0.0909	0.1088	0.0805
ROUGE-L	0.1748	0.1943	0.2347	0.1538	0.2614	0.2214	0.2299	0.2266	0.2181	0.2556	0.2229	0.2472
METEOR	0.2221	0.1745	0.1804	0.1952	0.2495	0.1692	0.1728	0.2202	0.2276	0.2242	0.1968	0.2370
CHRF++	44.91	44.90	37.25	38.92	43.78	42.39	41.64	42.30	47.515	45.71	41.75	47.33
BERT (F1)	0.1223	0.0619	0.1144	0.0117	0.0974	0.1487	0.1460	0.1481	0.0770	0.2100	0.0994	0.2105
Clinical BERT (F1)	0.7750	0.7409	0.7348	0.6974	0.7609	0.7723	0.7550	0.7528	0.7890	0.8098	0.7573	0.7846

Table 3: Evaluation of generated SOAP notes across lexical (ROUGE, METEOR), character-level (CHRF++), and semantic F1-score (BERT, Clinical BERT) metrics for three cases. Models: **Our Approach (skin-SOAP)**, GPT-40, Janus Pro Deepseek, Claude 3.7 Sonnet.

Model 1: Our Approach	
I. Structure: 5/5	
	COAD structure with class constation and appropriate content under each coeffici
	OAP structure with clear separation and appropriate content under each section.
2. Readability: 5/5	
	to follow. The note progresses from the patient's presentation and history to the clinical examination, assessment, and treatment plan. There are no instances of excessive verbosity or awkward
	nvey information efficiently. Overall, the SOAP note meets the highest standards of clarity, conciseness, and readability for a medical professional.
3. Completeness: 5/5	
	overing all important clinical points without any significant omissions. It provides a complete and medically accurate representation of the lesion description and context. Overall, the SOAP note me
	lly complete and covers all important clinical points.
1. Medical Relevance: 5/5	
•	igh level of clinical accuracy and appropriateness for the given input description of the skin lesion.
Total score: 20/20	
Model 2: GPT 4o	
I. Structure: 4/5	
	follows the SOAP format, but the minor issues with the history and location details prevent it from being a perfect score.
2. Readability: 5/5	
	nbiguity, presenting information in a straightforward manner that would be easily understood by a medical professional. The use of bullet points for the Plan section enhances readability without
	istrates excellent clarity, conciseness, and readability for a medical professional, meeting the highest standards of the provided rubric.
3. Completeness: 4/5	
Feedback: The note doesn't explicitly mention the "e	elevated borders" and "irregular pigmentation" as described in the input. Despite these minor omissions, the note is still quite detailed and covers most of the important clinical points. It provides a
overview of the patient's condition and outlines a cle	ar plan of action.
1. Medical Relevance: 5/5	
Feedback: Overall, the SOAP note is comprehensive,	, medically relevant, and plausible based on the input description.
Total score: 18/20	
Model 3: Claude 3.7 Sonnet	
I. Structure: 5/5	
eedback: The output does not contain any unneces	sary information and each section is clearly defined and relevant to the task. Therefore, the output meets the criteria for a perfect SOAP structure with clear separation
2. Readability: 5/5	
eedback: The note flows naturally, with each section	n logically leading to the next. There is no excessive verbosity or awkward phrasing. The assessment and plan sections are particularly well-written, providing clear next steps for diagnosis and
reatment. Overall, the note meets the highest stand	dard of being very easy to read and flows naturally, as per the scoring rubric.
3. Completeness: 4/5	
eedback: The note could have included more specif	ic details about the irregular pigmentation. Despite these minor omissions, the note is largely complete and covers the important clinical points.
Medical Relevance: 5/5	
eedback: There are no major inaccuracies or irrelev	vant statements in the SOAP note. The information provided is relevant, plausible, and appropriate for the given clinical description. Therefore, the output meets the highest standard of the scoring
rubric.	
Total score: 19/20	
Model 4: Deepseek Janus Pro	
L. Structure: 5/5	
eedback: The output demonstrates a perfect under	standing and application of the SOAP structure, with clear separation and appropriate content under each section. There are no formatting issues or missing elements.
2. Readability: 5/5	
	n logically building upon the previous one. It avoids excessive verbosity or awkward phrasing, making it easy for a medical professional to quickly grasp the key information and next steps. Overall
	nciseness, and readability for a medical professional.
3. Completeness: 4/5	,
	siled assessment of other potential symptoms or concerns the patient might have. Overall, the SOAP note is quite thorough and covers most of the important clinical points, with only minor omission
Medical Relevance: 5/5	and descent of the potential symptoms of contents are potent ingre-interest source in the in-quite information of the important chinesis with only limited of the source in the inspectation of the important chinesis with the potential symptoms of contents are potential inspectations.
	medically relevant, and appropriate for the described skin lesion. There are no major inaccuracies or irrelevant statements, and the clinical content is highly plausible.

Figure 3: Comparison of Flow-Judge Feedback Across Four Language Models

- Structure: Does the clinical note correctly follow the structured SOAP format, with distinct and appropriate content under each section (S: Chief Complaint and Medical History; O: Examination findings and Observations; A: Investigations, Diagnosis, and Summary; P: Treatment Plan and Patient Education)?
- **Readability:** Is the language of the clinical note clear, concise, and readable for a medical professional without excessive complexity or ambiguity?
- Completeness: Does the clinical note cover all the key details described in the input lesion description and context, and address all aspects of the clinical scenario, ensuring that no critical details are overlooked?
- Medical Relevance: Is the clinical content of the SOAP note medically relevant, plausible, and appropriate given the input description of the skin lesion?

This setup enabled a structured, blinded review of the generated outputs, allowing us to capture both technical and

clinical quality dimensions beyond traditional quantitative scores. The total scores for each model were shown in Fig 3, our approach achieved a perfect score of 20/20, with feedback highlighting its clear organization, clinical accuracy, and comprehensive coverage of all relevant details. In contrast, GPT-40 and the other baselines scored between 18/20 and 19/20. While their outputs were well-structured and readable, minor issues such as omissions of specific symptoms or less detailed assessments slightly impacted their completeness scores. Overall, Flow-Judge evaluations confirm that our framework delivers higher clinical fidelity and completeness than SOTA models, highlighting key strengths and improvement areas.

5 Example of Generated SOAP Notes

Figure 4 shows two representative SOAP notes generated by our model using a lesion image and its caption. These examples demonstrate the framework's ability to produce clinically structured and coherent documentation aligned with the

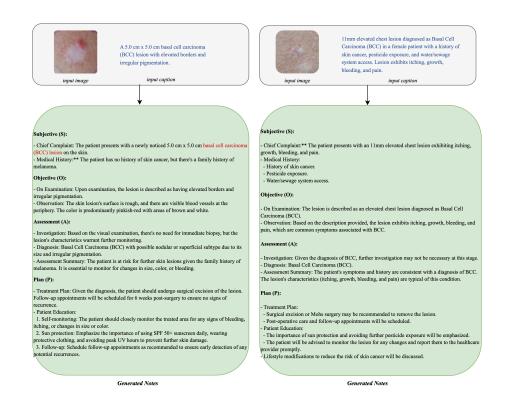


Figure 4: Examples (a) and (b) show structured SOAP notes generated by our proposed multimodal framework using an input lesion image and its corresponding caption.

SOAP format. In both cases, the model integrates visual and textual inputs to generate comprehensive notes.

Figure 4(a) illustrates a structural error, where the diagnosis of BCC is incorrectly placed in the Chief Complaint section instead of the Assessment section. In contrast, Figure 4(b) shows a correctly formatted note, with symptoms captured under Chief Complaint and diagnosis placed appropriately in Assessment. These examples highlight the model's potential to learn documentation structure, while also revealing areas for improvement.

6 Conclusion

In this work, we presented, skin-SOAP, a weakly supervised multimodal framework for generating clinically structured SOAP notes from limited dermatologic inputs. By leveraging generative language model generated captions, retrieval-augmented knowledge integration, and fine-tuning a Vision-LLaMA model with weak supervision, this study reduces dependence on large-scale expert annotations while maintaining strong clinical relevance and structural coherence. Through both qualitative and quantitative evaluations, including our proposed metrics such as MedConceptEval and Clinical Coherence Score (CCS), we demonstrated that our method produces high-quality, clinically meaningful notes, advancing the development of scalable and reliable clinical documentation systems. Ultimately, our framework has the potential to accelerate dermatology clinical workflows, reduce time-to

treatment, and improve overall patient care.

7 Limitations and Future Work

While our proposed framework shows strong promise for structured SOAP notes generation, it has few limitations. The quality of the generated notes remains dependent on the accuracy of the retrieved domain-specific knowledge, which may introduce biases or propagate incomplete information. Additionally, we utilized only a single dataset, as variations in metadata across different sources posed challenges for standardization. Our evaluation was further constrained by a small set of expert-annotated samples, restricting large-scale validation. Like most generative models, our approach may hallucinate when faced with ambiguous or unfamiliar inputs. Although retrieval-augmented generation helps reduce this risk by grounding outputs in clinical knowledge, additional precautions are needed for real-world deployment. In Future work, we will focus on expanding to more diverse datasets, and incorporating human-in-the-loop refinement strategies. Furthermore, developing evaluation benchmarks that capture the progression of clinical reasoning across multiple encounters and support decision making could significantly improve the utility of automated SOAP note generation in real-world healthcare settings.

References

- [Amenyo *et al.*, 2025] Solomon Amenyo, Maura R Grossman, Daniel G Brown, and Brendan Wylie-Toal. Assessment of ai-generated pediatric rehabilitation soap-note quality. *arXiv preprint arXiv:2503.15526*, 2025.
- [American Cancer Society, 2024] American Cancer Society. Donate to support the fight against cancer. https://donate.cancer.org/, 2024.
- [Biswas and Talukdar, 2024] Anjanava Biswas and Wrick Talukdar. Intelligent clinical documentation: Harnessing generative ai for patient-centric clinical note generation. arXiv preprint arXiv:2405.18346, 2024.
- [Brown et al., 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.
- [Chen and Hirschberg, 2024] Yu-Wen Chen and Julia Hirschberg. Exploring robustness in doctor-patient conversation summarization: An analysis of out-of-domain soap notes. *arXiv preprint arXiv:2406.02826*, 2024.
- [Cheng et al., 2025] Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. Caparena: Benchmarking and analyzing detailed image captioning in the llm era. arXiv preprint arXiv:2503.12329, 2025.
- [Dettmers *et al.*, 2023] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.
- [He et al., 2024] Fei He, Kai Liu, Zhiyuan Yang, Yibo Chen, Richard D Hammer, Dong Xu, and Mihail Popescu. pathclip: Detection of genes and gene relations from biological pathway figures through image-text contrastive learning. *IEEE Journal of Biomedical and Health Informatics*, 2024.
- [Huang *et al.*, 2019] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.
- [Jung et al., 2024] HyoJe Jung, Yunha Kim, Heejung Choi, Hyeram Seo, Minkyoung Kim, JiYe Han, Gaeun Kee, Seohyun Park, Soyoung Ko, Byeolhee Kim, et al. Enhancing clinical efficiency through llm: Discharge note generation for cardiac patients. arXiv preprint arXiv:2404.05144, 2024.
- [Leong *et al.*, 2024] Hui Yi Leong, Yi Fan Gao, Ji Shuai, Yang Zhang, and Uktu Pamuksuz. Efficient fine-tuning of large language models for automated medical documentation. *arXiv* preprint arXiv:2409.09324, 2024.
- [Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for

- knowledge-intensive nlp tasks. Advances in neural information processing systems, 33:9459–9474, 2020.
- [Li et al., 2024] Yizhan Li, Sifan Wu, Christopher Smith, Thomas Lo, and Bang Liu. Improving clinical note generation from complex doctor-patient conversation. arXiv preprint arXiv:2408.14568, 2024.
- [Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- [Mayo Clinic Staff, 2023] Mayo Clinic Staff. Melanoma symptoms and causes, 2023. Accessed: 2025-04-30.
- [National Cancer Institute, 2024] National Cancer Institute. Comprehensive cancer information. https://www.cancer.gov/, 2024.
- [NHS, 2024] NHS. Melanoma skin cancer symptoms. https://www.nhs.uk/conditions/melanoma-skin-cancer/ symptoms/, 2024.
- [Pacheco et al., 2020] Andre GC Pacheco, Gustavo R Lima, Amanda S Salomao, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio CR Alves Jr, José GM Esgario, Alana C Simora, Pedro BC Castro, et al. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in brief*, 32:106221, 2020.
- [Popović, 2017] Maja Popović. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618, 2017.
- [Post, 2018] Matt Post. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*, 2018.
- [Ramprasad *et al.*, 2023] Sanjana Ramprasad, Elisa Ferracane, and Sai P Selvaraj. Generating more faithful and consistent soap notes using attribute-specific parameters. In *Machine Learning for Healthcare Conference*, pages 631–649. PMLR, 2023.
- [Rogers et al., 2015] Howard W. Rogers, Martin A. Weinstock, Steven R. Feldman, and Brett M. Coldiron. Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the us population, 2012. *JAMA Dermatology*, 151(10):1081–1086, 2015.
- [Schloss and Konam, 2020] Benjamin Schloss and Sandeep Konam. Towards an automated soap note: classifying utterances from medical conversations. In *Machine Learning for Healthcare Conference*, pages 610–631. PMLR, 2020.
- [Singh et al., 2023] Gagandeep Singh, Yue Pan, Jesus Andres-Ferrer, Miguel Del-Agua, Frank Diehl, Joel Pinto, and Paul Vozila. Large scale sequence-to-sequence models for clinical note generation from patient-doctor conversations. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 138–143, 2023.
- [Singhal *et al.*, 2023] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

- [Skin Cancer Specialists, 2024] Skin Cancer Specialists. Skin cancer diagnosis and treatment. https://www.stxskincancer.com, 2024.
- [Van Veen et al., 2023] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. Clinical text summarization: adapting large language models can outperform human experts. *Research square*, pages rs–3, 2023.
- [Wei *et al.*, 2024] Maria L Wei, Mikio Tada, Alexandra So, and Rodrigo Torres. Artificial intelligence and skin cancer. *Frontiers in medicine*, 11:1331895, 2024.
- [Yan et al., 2024] Lawrence KQ Yan, Qian Niu, Ming Li, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Benji Peng, Ziqian Bi, Pohsun Feng, Keyu Chen, et al. Large language model benchmarks in medical tasks. arXiv preprint arXiv:2410.21348, 2024.
- [Yang et al., 2023] Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263, 2023.
- [Zhang *et al.*, 2019] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- [Zheng et al., 2023] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging Ilm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36:46595– 46623, 2023.
- [Zhou et al., 2024] Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, Shawn Afvari, et al. Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4. *Nature Communications*, 15(1):5649, 2024.