SOAP Automated Medical Report Generator

Yassin Saad*, Michael Adel†, Michael Wagdy‡, Sara Aboalyazeed§, Ziad Maher¶
Department of Artificial Intelligence, Nile University, Egypt

***Abstract*—The automation of clinical documentation, particularly the generation of SOAP (Subjective, Objective, Assessment, Plan) notes from patient-provider dialogues, holds significant potential to alleviate the administrative burden on healthcare professionals. This study explores the efficacy of various fine-tuning techniques applied to large language models (LLMs) for SOAP note generation. Utilizing the publicly available `omi-health/medical-dialogue-to-soap-summary` dataset from Hugging Face, we fine-tuned five models: Small T5 with full-parameter tuning, Facebook BART Large with transfer learning, LLaMA 3 2.3B using prefix tuning, Tiny LLaMA with adapter learning, and Microsoft Phi-2 employing Low-Rank Adaptation (LoRA). Additionally, we employed a judge model powered by Grok's LLaMA 3 70B to evaluate the quality of generated SOAP notes based on completeness, correctness, organization, and clinical relevance. Our methodology includes a comparative analysis of these models, assessing their performance using lexical, semantic, and quality-based metrics. The results provide insights into the effectiveness of each fine-tuning approach and the utility of the judge model, contributing to the advancement of automated medical documentation.**

I. Introduction

Clinical documentation is a critical component of healthcare delivery, serving as a record of patient encounters and facilitating communication among healthcare providers. The SOAP note format—comprising Subjective, Objective, Assessment, and Plan sections—is widely adopted for structuring clinical notes. However, manual creation of these notes is time-consuming and contributes to clinician burnout. Advancements in Natural Language Processing (NLP), particularly the development of large language models (LLMs), have opened avenues for automating the generation of clinical notes from patient-provider dialogues.

Recent studies have demonstrated the potential of LLMs in medical summarization tasks. For instance, the MediGen model, a fine-tuned LLaMA3-8B, utilized QLoRA and Parameter-Efficient Fine-Tuning (PEFT) techniques to generate SOAP notes, achieving notable performance metrics and clinical usability [3]. Similarly, prompt engineering strategies have enhanced LLM performance in generating structured medical reports [2].

Building upon these advancements, our study investigates the application of various fine-tuning techniques to different LLM architectures for SOAP note generation. We evaluate five models—Small T5 [5], Facebook BART Large [12], LLaMA 3 2.3B, Tiny LLaMA [11], and Microsoft Phi-2—each fine-tuned using distinct methodologies [6]–[8], [10]. Additionally, we introduce a judge model powered by Grok's LLaMA 3 70B to assess the quality of generated SOAP notes, providing a comprehensive evaluation framework for clinical applicability.

II. Related Work

The automation of clinical note generation has become a focal point in NLP research, driven by the need to reduce administrative burdens in healthcare. Several studies have explored the use of large language models (LLMs) for generating structured clinical summaries, such as SOAP notes, from patient-provider dialogues. These efforts generally focus on three key areas: fine-tuning LLMs for medical summarization, leveraging prompt engineering to improve output quality, and developing robust evaluation frameworks to ensure clinical accuracy and relevance.

Nair et al. (2023) proposed MEDSUM-ENT, a multi-stage summarization framework based on GPT-3. Their approach included medical entity extraction, entity resolution, and structured summarization, achieving high clinical accuracy as validated by human experts. The framework demonstrated the potential of LLMs to generate medically accurate summaries, particularly for structured formats like SOAP notes [1]. Similarly, Van Zandvoort et al. (2023) investigated transformer-based prompt engineering to enhance automated medical reporting. They compared zero-shot, one-shot, and two-shot prompting strategies, finding that two-shot prompting with scope and domain context significantly improved the quality of generated SOAP notes [2].

The MediGen model, developed by Leong et al. (2024), fine-tuned LLaMA3-8B using QLoRA and PEFT on the ACI-BENCH dataset. This model achieved a ROUGE-1 score of 58.22% and a BERTScore-F1 of 72.1%, with 75% of its generated SOAP notes deemed clinically usable by human evaluators. The study highlighted the effectiveness of parameter-efficient fine-tuning methods in adapting large models for medical documentation tasks [3], [9]. Additionally, the ClinicSum framework by Neupane et al. (2024) employed a two-module architecture combining retrieval-based filtering with inference using fine-tuned pre-trained language models. This approach outperformed traditional methods in generating clinical summaries, emphasizing the importance of integrating retrieval and generation for improved accuracy [4].

Beyond specific applications in healthcare, foundational work on LLMs has informed our approach to model selection and fine-tuning. Raffel et al. (2019) introduced the T5 model, a text-to-text transformer framework that excels in tasks like summarization by treating all NLP tasks as text generation problems. Their work demonstrated the versatility of full-parameter fine-tuning for domain-specific tasks, which we apply to Small T5 in our study [5]. Lewis et al. (2019) developed BART, a denoising autoencoder that combines bidirectional

and autoregressive training, making it particularly effective for sequence-to-sequence tasks like dialogue summarization. BART's architecture influenced our use of Facebook BART Large with transfer learning for SOAP note generation [12]. Zhang et al. (2024) introduced TinyLLaMA, a compact model designed for efficiency on edge devices, which we fine-tuned using adapter learning to explore lightweight solutions for clinical settings [11].

Fine-tuning techniques have also been a critical focus in related work. Farahani et al. (2020) provided a comprehensive review of transfer learning, highlighting its effectiveness in leveraging pre-trained models for specialized tasks like medical text summarization [10]. Liang et al. (2021) proposed prefix tuning, a lightweight fine-tuning method that prepends trainable vectors to model inputs, keeping the base model frozen. This approach inspired our use of prefix tuning for LLaMA 3 2.3B, balancing performance and computational efficiency [6]. He et al. (2021) explored adapter-based tuning, inserting small trainable layers into transformer models, which we applied to Tiny LLaMA to minimize resource demands [7]. Hu et al. (2022) introduced Low-Rank Adaptation (LoRA), which adapts attention layers with low-rank matrices, a method we used for Microsoft Phi-2 to achieve efficient fine-tuning [8]. Han et al. (2024) conducted a survey on parameter-efficient fine-tuning, underscoring the trade-offs between full-parameter tuning and methods like LoRA and adapters, providing a theoretical foundation for our methodology [9].

Evaluation frameworks for medical NLP tasks have also been a growing area of interest. Leong et al. (2024) explored automated quality assessment of clinical summaries using LLMs, often leveraging models like LLaMA for scoring generated outputs based on criteria like completeness and correctness [3]. Barbella and Tortora (2022) analyzed the ROUGE metric for text summarization, noting its effectiveness in capturing lexical overlap but highlighting its limitations in assessing semantic similarity, which motivates our use of BERTScore alongside ROUGE [13]. Zhang et al. (2019) introduced BERTScore, which uses contextual embeddings to evaluate semantic similarity, providing a more nuanced assessment of generated text quality compared to traditional metrics like BLEU [14]. These studies informed our evaluation strategy, combining lexical, semantic, and quality-based metrics to assess SOAP note generation comprehensively.

Our research builds upon these foundations by systematically evaluating fine-tuning techniques across various LLM architectures and introducing a judge model powered by Grok's LLaMA 3 70B to assess the quality of generated SOAP notes, enhancing the reliability and clinical applicability of automated documentation systems.

III. Methodology

A. Dataset

We utilized the omi-health/medical-dialogue-to-soap-summary dataset available on Hugging Face. This dataset contains synthetic medical conversations between clinicians and patients, each paired with a corresponding SOAP note summary. It consists of 10,000 samples, split into 9,250 for training, 500 for validation, and 250 for testing. Each sample includes well-structured dialogue and SOAP-formatted ground truth, making it ideal for training models in clinical note generation tasks.

B. Models

We experimented with six large language models (LLMs), selected to cover a range of sizes, capabilities, and evaluation roles:

- **Small T5**: A lightweight version of the T5 architecture, known for its effectiveness in text-to-text tasks [5].
- **Facebook BART Large**: A denoising autoencoder for pretraining sequence-to-sequence models, well-suited for generation tasks [12].
- **LLaMA 3 2.3B**: A performant and scalable LLM from Meta, known for efficient fine-tuning compatibility.
- **Tiny LLaMA**: A compact model designed for edge devices, evaluated here for its efficiency in fine-tuning [11].
- **Microsoft Phi-2**: A performant small LLM from Microsoft that emphasizes strong generalization across NLP tasks.
- **Grok's LLaMA 3 70B (Judge Model)**: A larger model used to evaluate the quality of generated SOAP notes based on predefined criteria (completeness, correctness, organization, and clinical relevance).

C. Fine-Tuning Techniques

Full-Parameter Tuning (Small T5): All model weights were updated during training, allowing complete adaptation to the target task [5].

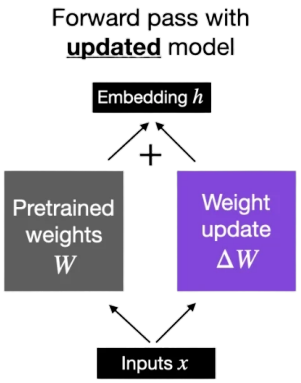
- 

Fig. 1. Illustration of Full-Parameter Tuning where all model parameters are updated.

Transfer Learning (BART Large): The model was fine-tuned with standard cross-entropy loss using the dataset, leveraging pre-trained knowledge [10].

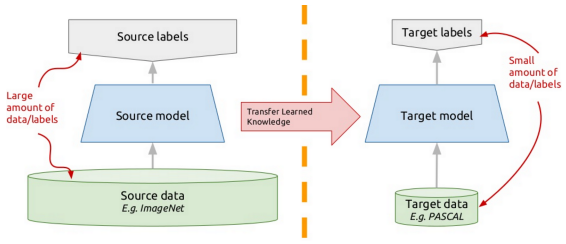


Fig. 2. Transfer Learning approach: Fine-tuning pre-trained weights on target dataset.

Prefix Tuning (LLaMA 3 2.3B): A lightweight method where trainable prefix vectors are prepended to the input at each layer, keeping the base model frozen [6].

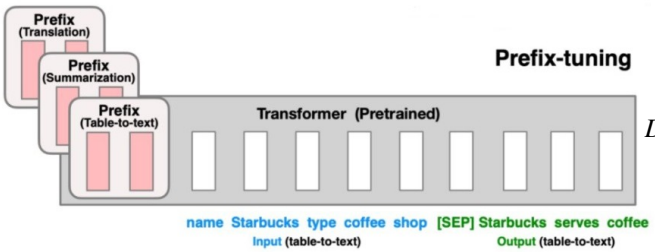


Fig. 3. Prefix Tuning: Adding trainable prefix vectors without updating the main model parameters.

Adapter Learning (Tiny LLaMA): Lightweight bottleneck layers (adapters) were inserted into each transformer block and trained while keeping other parameters frozen [7].

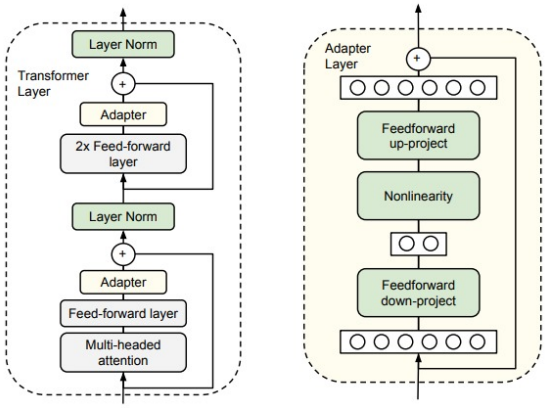


Fig. 4. Adapter Learning: Inserting small trainable layers inside the frozen model.

Low-Rank Adaptation (LoRA) (Phi-2): Introduced trainable low-rank matrices into attention layers to efficiently adapt the model to the task with minimal parameter updates [8].

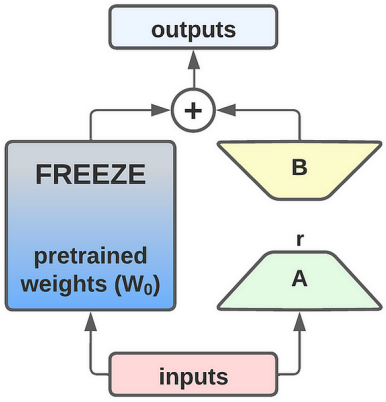


Fig. 5. LoRA Fine-Tuning: Injecting low-rank matrices for efficient parameter adaptation.

D. System Architecture

The system consists of four main stages:

1) **Preprocessing**: Medical dialogues are preprocessed to extract relevant information.
2) **Fine-Tuning**: Various LLMs are fine-tuned using different techniques to generate SOAP notes.
3) **Postprocessing**: Generated outputs are structured into proper SOAP note format.
4) **Evaluation (Judge Model)**: Grok's LLaMA 3 70B evaluates the quality of generated SOAP notes based on completeness (0.25 weight), correctness (0.35 weight), organization (0.20 weight), and clinical relevance (0.20 weight), producing a weighted score from 0 to 10.

E. Evaluation Metrics

We evaluated the models using two sets of metrics:

- **Lexical and Semantic Metrics (Generation Models)**:
  - **ROUGE-1**: Measures unigram (word-level) overlap.
  - **ROUGE-2**: Measures bigram (two-word sequence) overlap.
  - **ROUGE-Lsum**: Computes the longest common subsequence.
  - **BERTScore-F1**: Assesses semantic similarity using BERT embeddings.
- **Judge Model Score (Grok's LLaMA 3 70B)**: A weighted score (0–10) based on completeness (0.25), correctness (0.35), organization (0.20), and clinical relevance (0.20).

IV. Results and Discussion

We present the performance of the five generation models using both lexical/semantic metrics and judge model scores.

Fig. 6. General architecture of the SOAP note generation system, including preprocessing, fine-tuning, postprocessing, and evaluation stages.

TABLE I
COMPARISON OF MODEL PERFORMANCE ON LEXICAL AND SEMANTIC METRICS

Model	ROUGE-1	ROUGE-2	ROUGE-Lsum	BERTScore-F1
Small T5	0.4724	0.2698	0.4003	0.8709
BART Large	0.5394	0.3064	0.4569	0.8940
LLaMA 3 2.3B	0.4804	0.2451	0.4304	0.8680
Tiny LLaMA	0.3490	0.1070	0.4320	0.5147
Microsoft Phi-2	0.3280	0.1020	0.3980	0.5037

TABLE II
AVERAGE JUDGE MODEL SCORES FOR GENERATED SOAP NOTES

Model	Judge Score (0–10)
Small T5	5.1
BART Large	6.1
LLaMA 3 2.3B	7.9475
Tiny LLaMA	1.2925
Microsoft Phi-2	1.1785

A. Lexical and Semantic Metrics

Table I shows the performance comparison using ROUGE-1, ROUGE-2, ROUGE-Lsum, and BERTScore-F1 metrics.

B. Judge Model Scores

Table II presents the average scores assigned by Grok's LLaMA 3 70B judge model, evaluating the quality of generated SOAP notes.

The judge model results indicate that LLaMA 3 2.3B with prefix tuning achieved the highest quality score (7.9475), reflecting its strong performance across completeness, correctness, organization, and clinical relevance. BART Large also performed well (6.1), benefiting from transfer learning. Tiny LLaMA and Microsoft Phi-2 scored the lowest (1.2925 and 1.1785, respectively), likely due to their compact sizes and limited capacities for complex clinical text generation.

V. CONCLUSION

This study demonstrates the potential of fine-tuned large language models to automate SOAP note generation, addressing the administrative burden in healthcare. By evaluating five models—Small T5, Facebook BART Large, LLaMA 3 2.3B, Tiny LLaMA, and Microsoft Phi-2—using various fine-tuning techniques, and introducing a judge model (Grok's LLaMA 3 70B) for quality assessment, we provide a comprehensive analysis of their effectiveness. The judge model offers a practical evaluation framework, with LLaMA 3 2.3B achieving the highest quality score (7.9475). Full-parameter tuning and transfer learning offer robust adaptation, while parameter-efficient methods like LoRA and adapter learning provide computational efficiency [8], [9]. These findings contribute to NLP in healthcare, offering scalable solutions for automated documentation.

VI. FUTURE WORK

Future research can enhance the system by expanding the dataset to include diverse medical specialties and real-world dialogues, incorporating multimodal inputs (e.g., audio or visual data) [4], and exploring advanced prompt engineering techniques like few-shot learning [2]. Integrating real-time clinical decision support systems and conducting clinical validation studies with healthcare professionals are also promising directions [3]. Additionally, enhancing the judge model by incorporating feedback from healthcare professionals to refine evaluation criteria could further improve its reliability.

REFERENCES

[1] V. Nair, E. Schumacher, and A. Kannan, "Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models," *arXiv preprint arXiv:2305.05982*, 2023, doi: 0.48550/arxiv.2305.05982.

[2] D. Van Zandvoort, L. Wiersema, T. Huibers, S. Van Der Sandra, and S. Brinkkemper, "Enhancing summarization performance through transformer-based prompt engineering in automated medical reporting," *arXiv preprint arXiv:2311.13274*, 2023, doi: 0.48550/arxiv.2311.13274.

[3] H. Y. Leong, Y. F. Gao, J. Shuai, Y. Zhang, and U. Pamuksuz, "Efficient fine-tuning of large language models for automated medical documentation," *arXiv preprint arXiv:2401.02385*, 2024, doi: 0.13140/rg.2.2.26884.74881.

[4] S. Neupane, H. Tripathi, S. Mitra, S. Bozorgzad, S. Mittal, S. Rahimi, and A. Amirlatifi, "CLINICSUM: Utilizing language models for generating clinical summaries from patient-doctor conversations," *arXiv preprint arXiv:2412.04254*, 2024, doi: 0.48550/arxiv.2412.04254.

[5] C. Raffel *et al.*, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *arXiv preprint arXiv:1910.10683*, 2019, doi: 0.48550/arxiv.1910.10683.

[6] X. L. Li and P. Liang, “Prefix-tuning: Optimizing continuous prompts for generation,” *arXiv preprint arXiv:2101.00190*, 2021, doi: 0.48550/arxiv.2101.00190.

[7] R. He *et al.*, “On the effectiveness of adapter-based tuning for pretrained language model adaptation,” *arXiv preprint arXiv:2106.03164*, 2021, doi: 0.48550/arxiv.2106.03164.

[8] E. J. Hu *et al.*, “LoRA: Low-rank adaptation of large language models,” *ICLR*, 2022, doi: 0.48550/arxiv.2106.09685.

[9] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, “Parameter-efficient fine-tuning for large models: A comprehensive survey,” *arXiv preprint arXiv:2403.14608*, 2024, doi: 0.48550/arxiv.2403.14608.

[10] A. Farahani, B. Pourshojae, K. Rasheed, and H. R. Arabnia, “A concise review of transfer learning,” in *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2020, pp. 344–351, doi: 0.1109/csci51800.2020.00065.

[11] P. Zhang, G. Zeng, T. Wang, and W. Lu, “TinyLlama: An open-source small language model,” *arXiv preprint arXiv:2401.02385*, 2024, doi: 0.48550/arxiv.2401.02385.

[12] M. Lewis *et al.*, “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” *arXiv preprint arXiv:1910.13461*, 2019, doi: 0.48550/arxiv.1910.13461.

[13] M. Barbella and G. Tortora, “Rouge metric evaluation for text summarization techniques,” *SSRN Electron. J.*, 2022, doi: 0.2139/ssrn.4120317.

[14] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating text generation with BERT,” *arXiv preprint arXiv:1904.09675*, 2019, doi: 0.48550/arxiv.1904.09675.