CLINICSUM: Utilizing Language Models for Generating Clinical Summaries from Patient-Doctor Conversations

Subash Neupane*, Himanshu Tripathi†, Shaswata Mitra‡, Sean Bozorgzad§,
Sudip Mittal¶, Shahram Rahimi‖, and Amin Amirlatifi**

Dept. of Computer Science and Engineering, Mississippi State University
Potentia Analytics Inc.
Dave C. Swalm School of Chemical Engineering, Mississippi State University
Email: {*sn922, †ht577, ‡sm3843}@msstate.edu, {§sean}@potentiaco.com
{¶mittal, ‖rahimi}@cse.msstate.edu, **amin@che.msstate.edu

Abstract—This paper presents CLINICSUM, a novel framework designed to automatically generate clinical summaries from patient-doctor conversations. It utilizes a two-module architecture: a retrieval-based filtering module that extracts Subjective, Objective, Assessment, and Plan (SOAP) information from conversation transcripts, and an inference module powered by fine-tuned Pre-trained Language Models (PLMs), which leverage the extracted SOAP data to generate abstracted clinical summaries. To fine-tune the PLM, we created a training dataset of consisting 1,473 conversations-summaries pair by consolidating two publicly available datasets, FigShare and MTS-Dialog, with ground truth summaries validated by Subject Matter Experts (SMEs). CLINICSUM's effectiveness is evaluated through both automatic metrics (e.g., ROUGE, BERTScore) and expert human assessments. Results show that CLINICSUM outperforms state-of-the-art PLMs, demonstrating superior precision, recall, and F-1 scores in automatic evaluations and receiving high preference from SMEs in human assessment, making it a robust solution for automated clinical summarization.

Index Terms—Clinical summaries, SOAP, Summarization, PLM, Fine-tuning, RAG

I. INTRODUCTION

The advent of transformer-based models such as OpenAI GPT models [1], Meta LLAMA[2] variants, and Google Gemini[3] has revolutionized Natural Language Processing (NLP) by significantly improving performance across a wide array of tasks. These advancements, driven primarily by transfer learning, have opened up new possibilities for applying these models in specialized domains [1], [2]. One such domain is healthcare, where leveraging Pre-trained Language Models (PLMs) to automatically generate clinical summaries from doctor-patient conversations presents a promising application with substantial benefits for both patients and healthcare providers.

Clinical summaries play a critical role in healthcare by improving patients' understanding of care plans and reducing the risk of misinterpreting medical information. Research indicates that patients forget 40-80% of the medical information
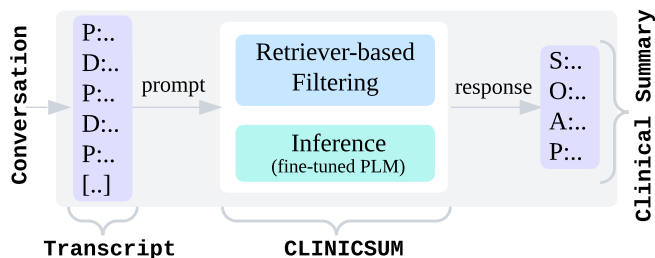
[1] https://platform.openai.com/docs/models
[2] https://llama.meta.com/
[3] https://gemini.google.com

Fig. 1: A graphical overview of the CLINICSUM. *P* denotes the Patient and *D* denotes the Doctor in the conversation transcript. *S, O, A,* and *P* refer to the Subjective, Objective, Assessment, and Plan components of the clinical summary.

provided by healthcare practitioners almost immediately [3] and misconstrue nearly half of what they remember [4]. For healthcare providers, generating these summaries automatically can alleviate the administrative burden of updating Electronic Health Records (EHRs), a task strongly associated with physician burnout [5], [6].

However, the application of PLMs in this context is not without challenges. Since PLMs are generally trained on broad, non-specialized text corpora, they are prone to producing inaccuracies—such as hallucinations [7]—that could have serious consequences for patient care. Addressing these challenges requires more than just deploying PLMs; it necessitates a tailored approach that can accurately capture the nuances of medical conversations while ensuring the reliability of the generated summaries.

In this paper, we present CLINICSUM, a comprehensive framework designed to automatically generate clinical summaries in the Subjective Objective Assessment and Plan (SOAP) format from transcribed patient-doctor conversations. Fig. 1 provides an illustration of CLINICSUM. To tackle the limitations of current PLMs in healthcare, CLINICSUM integrates a retrieval-based filtering module and an inference module, working in tandem to produce accurate and contextually relevant summaries. The retrieval-based filtering

module is responsible for extracting the SOAP components from the given transcript which it achieves by leveraging an ensemble retriever approach, that combines sparse and dense retrieval techniques, to capture both lexical and semantic meanings from the transcribed conversations. By utilizing this dual approach, filtering method ensures that the most relevant information is passed to the inference module, which is fine-tuned to generate clinical summaries.

For fine-tuning a PLM, we collaborated with Potentia Inc.[4], a healthcare software company, to create a new training dataset. This dataset was constructed by combining 1,473 patient-doctor conversations from two publicly available sources, FigShare[5] and MTS-Dialog [8] and generating their corresponding clinical summaries. Subject Matter Experts (SMEs), including doctors and physicians from Potentia Inc., manually reviewed and corrected the summaries to ensure their high quality, providing reliable data for fine-tuning task. The final training data is publicly available through huggingface [6].

Previous research on generating clinical summaries includes Zhang et al. [9], which fine-tuned a BART model to handle long and noisy doctor-patient conversation transcripts, and Giorgi et al. [10], which used fine-tuning and few-shot In-Context Learning (ICL) [11] with GPT-4. While these methods have achieved some success, CLINICSUM introduces a unique combination of retrieval-based filtration and a fine-tuned inference module to generate SOAP format clinical summaries. Unlike the multistage summarization process for long conversations used by Zhang et al. or the focus on ICL by Giorgi et al., our approach refines input data before summarization, leading to superior performance in both automatic and expert human evaluations.

The main contributions of this paper are as follows:

- We demonstrate the feasibility of generating clinical summaries in SOAP format utilizing the transcribed patient-doctor conversations.
- We create a new dataset of clinical summaries corresponding to patient-doctor conversations from the FigShare and MTS-Dialog datasets.
- We built CLINICSUM- a framework that automatically generates clinical summaries.
- We showcase CLINICSUM's proficiency in generating accurate and relevant clinical summaries through both automatic and expert human evaluations.

The rest of the paper is organized as follows: Section II discusses the background and related works. Section III describes our task. Section IV provides insight on CLINICSUM's architecture and methodology. In Section V we present our experiments, evaluation, and results. Section VI discusses the limitations and Section VII concludes the paper.

II. BACKGROUND & RELATED WORK

A. Clinical Summaries

Clinical summaries are concise records of patient encounters, detailing medical history, current condition, treatment plans, and progress. Automatically generating these summaries from doctor-patient conversations benefits patients by improving recall and understanding [10] of care plans, and helps doctors by streamlining documentation [12] and reducing administrative workload [13]. One widely used format for clinical summary is SOAP [14]. An example of a clinical summary in SOAP format is presented in Fig. 2(B). The *Subjective* section documents the patient's personal experiences, including the Chief Complaint (CC), History of Present Illness (HPI), and relevant medical history. The *Objective* section records measurable data like vital signs, physical exam findings, and lab results. The *Assessment* section combines subjective and objective information to diagnose the patient's condition, highlighting the problem and differential diagnoses. Lastly, the *Plan* section details the approach for addressing or investigating the problem further. These structured summaries aids in making informed clinical decisions, tracking patient progress, and maintaining continuity of care [1].

B. LLM, RAG and Fine-tuning

Transformer architectures [15] have fueled the advancement of Large Language Models (LLMs) in NLP, thanks to their remarkable parallelization capabilities [16]. Trained on massive internet text datasets and featuring substantial parameter sizes, LLMs exhibit impressive learning abilities. However, LLMs often struggle with factual questions in closed domains, where specialized knowledge is crucial. This difficulty can manifest in factually inaccurate predictions, a phenomenon known as *hallucination* [7]. This limitation may arise from a combination of factors, including a *deficit in domain knowledge*, *reliance on outdated information*, and *forgetting* [17], [18].

To mitigate the knowledge deficiency within PLMs for domain-specific tasks, an additional knowledge ingestion step is required. The two most common approaches currently practiced for external knowledge ingestion are Retrieval Augmented Generation (RAG) and Fine-tuning. The first approach, introduced around mid-2020 by Lewis et al., [19], is designed to enhance the performance of PLMs on knowledge-intensive tasks. This approach involves retrieving relevant information from external knowledge sources based on the input query. The retrieved content is then concatenated with the original query, providing the PLM with enriched context, which leads to more informed and accurate response generation.

The second approach, is to fine-tune PLM. In this approach, a PLM is further trained on a smaller, task-specific dataset to adapt it to a particular application. This process allows the model to leverage the general knowledge it has acquired during pre-training and refine its weights based on the new, more focused data, improving its performance on the target task. As PLM grows in size, updating all parameters during fine-tuning becomes increasingly costly and inefficient, especially with limited computational resources. This challenge has

driven research into Parameter Efficient Fine-Tuning (PEFT) methods that minimize tunable parameters while maintaining performance. Key approaches include adapter-based methods [20], prompt-based techniques [21], LoRA [22], QLoRA [23].

In contrast to RAG and fine-tuning approaches, an alternative approach is ICL. This technique utilizes examples (usually few-shot) embedded within the prompt to guide the model's response generation.

C. Related Works

The field of open-domain dialogue summarization, encompassing the task of summarizing conversations and meetings, remains relatively unexplored. While there have been a limited number of studies dedicated to this area [24], [25], research interest in summarizing dialogue within closed domains, particularly in the medical field, has been gaining momentum in recent years. Specifically, the automatic generation of clinical summaries from doctor-patient conversations has attracted significant attention [6], [12], [26]–[28].

To this day, various methods have been proposed, to generate clinical summaries in SOAP format including extractive and abstractive approaches. For example, Krishna et al. [29] proposed a modular approach combining extractive and abstractive summarization techniques to generate SOAP notes from doctor-patient conversations. Building upon the work of [29], Ramprasad et al., [30] on the other hand, focused on enhancing the faithfulness and consistency of SOAP notes generated by LLMs. Their work introduces section-specific cross-attention parameters in encoder-decoder models to improve the factual accuracy and relevance of generated notes. While Schloss and Konam [12] concentrated on classifying utterances from medical conversations into SOAP sections and speaker roles using a hierarchical encoder-decoder model.

In addition to extractive-abstractive methods, current research in automatic SOAP summary generation often involves fine-tuning PLM which closely aligns with our approach as discussed in Section IV. For example, Zhang et al., [9] fine-tuned a pre-trained BART model to automatically generate summaries from doctor-patient conversations. However, their work was limited to just two specialties, internal medicine and primary care, and the training data included only the History of Present Illness (HPI) section. Similarly, Giorgi et al., [10] explored two approaches first fine-tuning a PLM (Longformer-Encoder-Decoder[7]) and second using few-shot ICL [31]. In contrast, our approach combines retrieval-based filtering with inference using fine-tuned models in a zero-shot setting.

III. TASK FORMULATION

Given a set of patient-doctor conversation transcripts T, where each transcript $t_i \in T$ consists of unstructured conversation. The objective is to generate semi-structured SOAP clinical summaries for each transcript t_i. This involves using a function f, which applies a combination of information

---

[7]https://huggingface.co/docs/transformers/en/model_doc/led

retrieval techniques and a PLM to map each t_i to a semi-structured SOAP format n_i. Specifically, the function can be defined as:

$$n_i = f(t_i) \forall t_i \in T, n_i \in N | i \in \mathbb{N} \tag{1}$$

where, T is the set of Transcripts and N is the set of clinical summaries. These summaries n_i are organized in SOAP format where S is the **Subjective** component that summarizes patient's reported symptoms and experiences. O is the **Objective** component, detailing the observable and measurable clinical findings. A is the **Assessment** component, providing a diagnosis or evaluation based on the information and P is the **Plan** component, outlining the treatment and management strategies discussed in t_i.

IV. ARCHITECTURE & METHODOLOGY

This section presents the architecture of our framework, CLINICSUM, and outlines our methodology. The framework consists of two main modules: *retriever-based filtering* and *inference*, as illustrated in Fig. 2. The following subsections provide a detailed explanation of each module.

A. Retriever-based Filtering

The first module in our framework systematically processes doctor-patient conversation transcripts t_i to extract the SOAP elements for clinical summary n_i using a retrieval prompt (query) Q_R. For example, our retrieval prompt is *"Extract subjective, objective, assessment, and plan details from a given transcript"*. For extraction this module utilizes the following three sub-components:

*1) **Splitting**:* Splitting is the process of converting entire transcript t_i into a set of individual sentences/chunks c_i. To split t_i into c_i, we apply a sentence split regular expression. Hence, the splitting ($t_i \rightarrow c_i$) can be formulated as:

$$\{c_i : c_i \in C\} = split(t_j) \forall t_j \in T | i, j \in \mathbb{N}, i \geq j \tag{2}$$

*2) **Indexing**:* Given the set of sentences (C) in t_i, where c_i represents a sentence in the transcript, indexing is the process of projecting c_i into vector space (E) through an embedding model $\xi(.)$, where, e_i is the vector embedding of c_i and we store this obtained vector embedding into vector storage.

$$\{e_i : e_i \in E\} = \xi(c_i) \forall c_i \in C | i \in \mathbb{N} \tag{3}$$

*3) **Retrieval**:* The retrieval process uses an ensemble method that combines sparse retriever (for example BM25 [32]) and dense retriever (for example DPR [33] or our previous work [34]), assigning different weights to each (W_{Sparse}, and W_{Dense}), then ranks it using ranking algorithm. In our use case, we implemented Reciprocal Rank Fusion (RRF) [35] which combines rankings from multiple sources by computing reciprocal rank scores.

A sparse retriever searches for documents (c_i) similar to Q_R based on exact token matches in the sparse vector space

(A) Retrieval Prompt — Retriever — Retrieved Vectors

Embedded Chunks

[0.8,-1.2,-2.5...] [1.4,-1.1,-1.5...]
[0.4,0.1,0.9...] [1.3,3.2,0.4,...]

[0.8,-1.2,-2.5...] [1.4,-1.1,-1.5...] [0.8,-1.2,-2.5...]

Embedding Model

[0.8,-1.2,-] [1.4,-1.1,-] Filtered Document

Chunks

Text Reconstruction

Sentence Splitter

Combined Text — Context

Transcribed Conversation

Retriever-based filtering

Summary (SOAP)

Prompt / Instruction

Generator

PLM

Finetuned

Inference

(B) Patient-Doctor Conversation

D: Good morning! How can I help you today?
P: Hi, I've been experiencing shortness of breath for the past week and it's been bothering me. I thought I should come in and get it checked out.
D: I see. Shortness of breath can be concerning. Can you tell me more about your symptoms? How severe is the shortness of breath?
P: It's been moderate. I notice it especially when I exercise or engage in physical activities.
D: Okay. Have you noticed any other symptoms along with the shortness of breath? For example, do you hear any wheezing or have a cough?
P: Yes, I do hear wheezing when I breathe and sometimes I have a cough as well.
D: Thank you for sharing that information. Have you experienced similar symptoms in the past?
P: Yes, I had childhood asthma, but it hasn't bothered me much in recent years.
D: I see. It's important to consider your medical history. Does anyone in your family have a history of asthma or any other respiratory conditions?
P: Yes, my mother also had asthma.
D: That's helpful to know. Now, let's talk about any allergies you may have. Are you allergic to anything, such as pollen?
[...]

(C) Clinical Summary

Subjective:
- Symptoms: Shortness of breath, wheezing, cough
- Severity: Moderate
- Duration: Past week
- Associated symptoms: Wheezing, cough
- Relevant medical history: Childhood asthma, no recent issues
- Family history: Mother had asthma
- Allergies: Allergic to pollen
- Other concerns: Patient [..]

Objective:
- Physical examination findings: Wheezing on auscultation [..]

Assessment:
- Likely diagnosis: Asthma exacerbation
- Differential diagnosis: Other conditions that may present with similar symptoms
- Clinical impression: Patient has a history of asthma and is experiencing symptoms consistent with an asthma exacerbation [..]

Plan:
- Spirometry test tomorrow
- Blood tests tomorrow
- Consider long-acting bronchodilator for daily management
- Discuss proper inhaler technique and asthma management strategies [...]
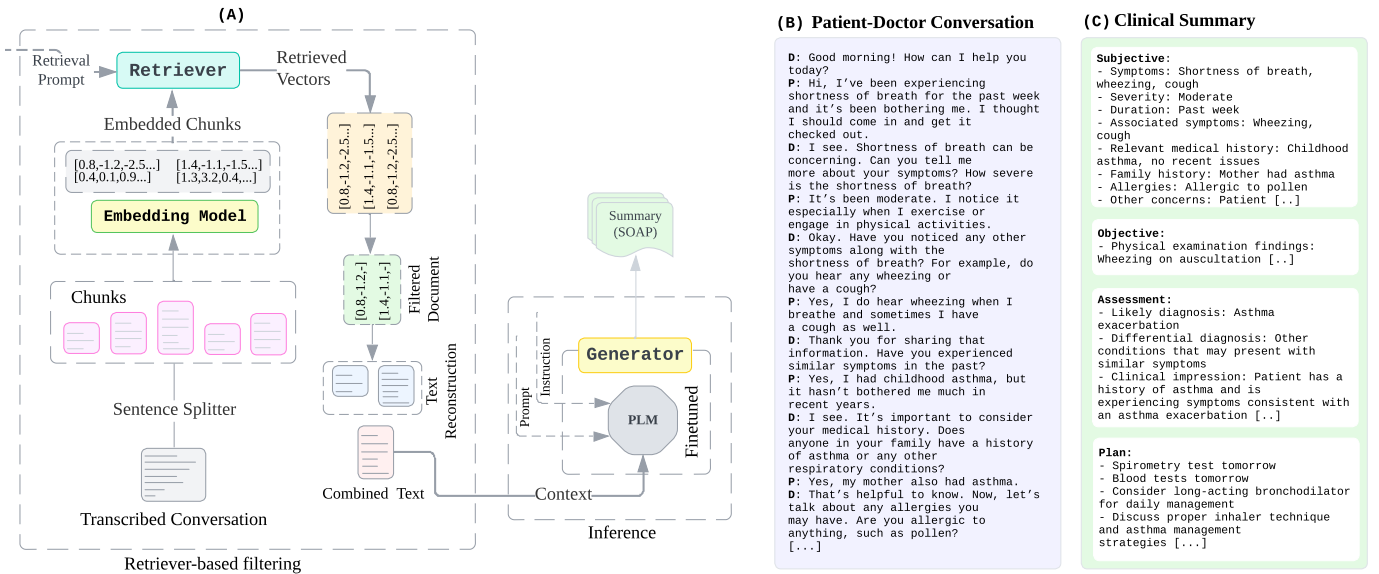
Fig. 2: **A** is graphical illustration of the CLINICSUM architecture. It comprises two modules: *retrieved-based filtering* and *inference*. **B** represents patient-doctor conversation, and **C** represents generated clinical summary. [...] (used for brevity) indicates that there is more textual information.

(C), usually employing traditional keyword-based methods or indexing techniques. We employ BM25 as our sparse retriever (R_{Sparse}) that can be represented as:

$$\{c_{i\_Sparse} : c_{i\_Sparse} \in C | 1 \leq i \leq k\} = R_{Sparse}(C, Q_R) \quad (4)$$

where, k is the number of chunks with highest term frequency.

A dense retriever searches for documents relevant to Q_R based on the exact or approximate neighbor similarity of embedded vectors (e_i) in a continuous embedding vector space (E), using dense representations. For retrieval, we also embed the retrieval prompt Q_R such that:

$$e_{Q_R} = \xi(Q_R) \quad (5)$$

Dense retriever (R_{Dense}) can be represented as:

$$\{e_i : e_i \in E, 1 \leq k\} = R_{Dense}(E, e_{Q_R}) \quad (6)$$

The similarity function ($Sim(.)$) can be cosine, dot-product, or euclidean. The top k relevant embedding (e_i) are then decoded ($c_i \leftarrow e_i$) to corresponding sentences (c_i). where the similarity function in R_{Dense} is as follows:

$$Sim(e_i, Q_R) = \frac{e_{Q_R} \cdot e_{c_i}}{\|e_{Q_R}\| \cdot \|e_{c_i}\|} \quad (7)$$

In order to obtain the corresponding sentence/chunk (c_i) from embedding e_i, we apply an inverse embedding or decoding function ($\overline{\xi(\cdot)}$).

$$\{c_{i\_Dense} : c_{i\_Dense} \in C | 1 \leq i \leq k\} = \{\overline{\xi(e_i)} \forall e_i \in E\} \quad (8)$$

Next, we combine both c_{i_Sparse} and c_{i_Dense} to obtain the final set of embedded documents, before re-ranking them:

$$c_{i\_Retrieved} = c_{i\_Dense} \cup c_{i\_Sparse} \quad (9)$$

The cardinality of $c_{i\_Retrieved}$ (say p) will be less or equal to the total number of chunks retrieved using sparse and dense retriever ($k + k = 2k | k \in \mathbb{N}$) i.e., $p \leq |c_{i\_Sparse}| + |c_{i\_Dense}| \leq \mathbb{N}$. Once combined, we apply a ranking method to reorder the documents. This is done by using RRF algorithm. The algorithm works by calculating rank score (r_i) for the corresponding retrieved chunk ($c_{i\_Retrieved}$). If a document appears in both c_{i_Sparse} and c_{i_Dense} with different rankings, we sum the reciprocals of each rank from both retrievers. Typically $S(c_{i\_Sparse}), S(c_{i\_Dense}) \in [0, 1]$, this summed reciprocal score can exceed 1. This combined score is used for final ranking with retriever weights ($W_{Sparse}, W_{Dense} | W \in [0, 1]; W_{Sparse} + W_{Dense} = 1$), with higher scores indicating greater relevance.

$$r_i = W_{Sparse} \times S(c_{i\_Sparse}) + W_{Dense} \times S(c_{i\_Dense}) \quad (10)$$

Then, we sort the obtained rank score with respect to Q_R in descending order using the following equation, where, λ is a constant to avoid division by 0. Finally, top k chunks are retrieved.

$$sort(Q_R, c_{i\_Retrieved}) = \sum_{i=1}^{\mathbb{N}} \frac{1}{\lambda + r_i(Q_R, c_{i\_Retrieved})} \quad (11)$$

$$\{c_{i\_Sorted} : c_{i\_Sorted} \in C | 1 \leq i \leq k\} = sort(Q_R, c_{i\_Retrieved}) \quad (12)$$

The final decoded output will include only those chunks that contain the subjective, objective, assessment, or plan components from the given transcript. By passing the transcripts through the retriever-based filtering, we effectively reduce the tokens that do not correspond to the generation of SOAP notes, thereby abstracting unnecessary information before sending it for inference. This reduction in tokens not only helps CLINICSUM avoid the token overflow problem but also mitigates the inference model from hallucination.

B. Inference

The inference module receives the patient context, derived from the final retrieved concatenated chunks c_i, along with an *instruction* , and a *prompt* ($Q_{\mathcal{PLM}_{FT}}$) as shown in Fig. 2. The instruction guides the language model in performing its task. In our case, we utilize Alpaca prompt, as shown in Fig. 3 as our instruction. On the other hand, prompt directs a fine-tuned PLM to produce clinical summaries in a zero-shot setting. The fine-tuned generator processes the prompt, patient context, and instructions to generate a comprehensive clinical SOAP summary. In the following subsections we describe our fine-tuning approach and then detail summary generation.

1) Fine-tuning: In this work, we leverage PEFT [20] approach to fine-tune a PLM for clinical summary generation. PEFT enables efficient fine-tuning with minimal resources and costs. Specifically, we adopt Low Rank Adaptation (LoRA) [36] method and load pre-trained models onto a GPU as quantized 4-bit weights. Our motivation for this approach is two-folds: first, to explore the **feasibility** of training a PLM, such as LLAMA-3, on a single consumer GPU with 24GB of memory (e.g., Nvidia 4090), and second, to assess the **effectiveness** of fine-tuned PLMs with 4-bit precision in accurately generating clinical summaries. Additionally, PEFT helps prevent *catastrophic forgetting* [18] after the model has been trained [37]. We use the Alpaca prompt [38] for both fine-tuning and inference tasks, as illustrated in Fig. 3. The training is conducted using Supervised Fine-Tuning (SFT). More information on training dataset is provided in Section V-A.


```
Instruction: """Below is an instruction that describes a task,
paired with an input that provides further context. Write a
response that appropriately completes the request.

### Instruction:
{}

### Input:
{}

### Response:
{}"""
```

Fig. 3: An example of an Alpaca prompt.

2) Summary Generation: We utilize the output of the first module—the context, i.e., the decoded relevant chunks containing subjective, objective, assessment, and plan information from a given transcript—along with an instruction

and a prompt as input to the inference module to generate a clinical summary. These input are concatenated and passed together to a fine-tuned PLM for summary generation, where $\mathcal{PLM}_{FT}$ is a fine-tuned PLM that understand how to generate a clinical summaries, $Q_{\mathcal{PLM}_{FT}}$ is a prompt (query), $inst$ denotes instruction and [.,.,.] stands for concatenation.

$$
\begin{aligned}
Summary = \mathcal{PLM}_{FT}([context, prompt, inst]) = \\
\mathcal{PLM}_{FT}([c_{i\_Sorted}, Q_{\mathcal{PLM}_{FT}}, inst])
\end{aligned} \tag{13}
$$

Inference is conducted in a zero-shot setting using a fine-tuned PLM. The fine-tuning process equips the model to generalize effectively, allowing it to generate accurate clinical summaries even for new, unseen patient conversations without requiring additional few-shot examples. An example of the final clinical summary for a specific patient-doctor conversation is provided in Fig. 2 (B) and (C) respectively.

V. EXPERIMENT & EVALUATION

A. Dataset Description and Preparation

In this research, we utilize two different datasets for the fine-tuning task. The first dataset is the Figshare dataset, which contains 272 patient-doctor conversations. These conversations span five medical specialties: *Cardiovascular, Gastrointestinal, Musculoskeletal, Dermatological*, and *Respiratory*. Table I provides an example from this dataset. The second dataset we use is the MTS-dialog dataset [8], which contains 1,701 patient-doctor conversations. These conversations are centered around *General Medicine, Orthopedic, Dermatology, Neurology*, and *Allergy/Immunology*. From this dataset, we selected a subset of 1,201 clean conversations for our study. We then combined them, resulting in a total of 1,473 conversations. Additional statistics, including the total number of sentences, words, characters, unique vocabulary, and tokens for the conversations in the combined dataset are presented in Table II.

TABLE I: Example of a doctor-patient conversation from the FigShare dataset.

Patient-Doctor Conversation
D: What brought you in today?
P: Sure, I'm I'm just having a lot of chest pain and and so I thought I should get it checked out.
D: OK, before we start, could you remind me of your gender and age?
P: Sure 39, I'm a male.
D: OK, and so when did this chest pain start?
P: It started last night, but it's becoming sharper.
D: ...
P:...
Medical Speciality: Cardiovascular

1) Ground-truth Generation: To fine-tune our models, we require ground truth data, which neither of these datasets provided. The MTS-dialog dataset contains very brief summaries, averaging less than three sentences, which are insufficient for our task. The Figshare dataset includes only conversations, with no summaries available. One key contribution of this paper is the creation of ground truth summaries for these

TABLE II: Statistics of the FigShare and MTS-Dialog datasets.

Patient-Doctor Conversation (Figshare)					
Metric	Sentences	Words	Char	Vocab	Tokens
Count	37910	369552	1478738	98535	472384
Mean	139.37	1358.64	5436.53	362.26	1736.70
Max	255	2401	9636	589	3102
Min	74	808	3229	254	1020
Patient-Doctor Conversation (MTS-Dialog)					
Count	15839	118558	500393	72225	152232
Mean	13.18	98.71	416.64	60.13	126.75
Max	167	1474	6823	457	2038
Min	1	1	8	1	3

conversations. To accomplish this, we collaborated closely with Potentia Analytics, a healthcare-focused data analytics and information technology company. We initially generated clinical summaries for all 1,473 conversations using the *GPT-4-O-Mini* model (managed through API calls) in a zero-shot setting. Subject Matter Experts (SMEs), specifically medical doctors from Potentia Analytics, then manually evaluated and verified the factual correctness and contextual relevance of these summaries. Based on their feedback, we rectified any inconsistencies, discrepancies, or inaccuracies pertaining the summaries. We then created a final training dataset of 1,473 conversation-summary pairs, which we then utilized for our fine-tuning task. This dataset is publicly accessible on HuggingFace. Table III provides additional statistics for the ground-truth summaries, broken down into subjective, objective, assessment, and plan components, detailing the number of sentences, words, characters, vocabulary, and tokens.

B. Evaluation

Due to the strict privacy concerns and Health Insurance Portability and Accountability Act (HIPAA) regulations around patient data, coupled with its inaccessibility, we opted to simulate patient-doctor conversations. For this, we partnered with the Department of Theatre & Film at Mississippi State University to create 20 simulated conversations. These conversations were staged as role-playing scenarios, where theater arts students simulated realistic interactions between patients and doctors. The conversations were recorded in WAV format and subsequently processed using Automatic Speech Recognition (ASR) techniques, specifically utilizing the Whisper-large[8] model. The average length of these role-played conversations is approximately 9 minutes. Additional statistics on these conversations are provided in Table IV. We then utilized these simulated conversations as our evaluation dataset to assess the robustness of CLINICSUM for the clinical summaries generation task. The evaluation was conducted through both automatic and manual methods.

In the following subsections, we discuss the results of our assessment.

1) Automatic Evaluation: In this paper, we consider two types of metrics: *lexical-based* and *text-embedding-based*, to evaluate the clinical summaries generated by CLINICSUM. For the lexical-based metric, we choose Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [39], which primarily focuses on lexical overlaps between generated summaries and the ground truth but does not capture the semantic meaning of the summaries. Considering the limitation of ROUGE, we also employ text-embedding-based metrics, such as BertScore [40]. It uses pre-trained contextual embeddings from a BERT-based model to evaluate the semantic similarity between the ground truth and generated summaries by computing cosine similarity. Specifically, we utilize the *deberta-xlarge-mnli*[9] model in our experiments.

Table V presents the clinical summarization results. We compared the performance of the state-of-the-art GPT-based proprietary PLMs with our approach (combination of retriever-based filtering and fine-tuning) equipped with open source models in generating clinical summaries from patient-doctor conversation in zero-shot settings. Notably, CLINICSUM, particularly when paired with LLAMA-3, outperformed GPT-based models in both ROUGE and BERTScore metrics.

In terms of ROUGE-1, which measures unigram overlap, *LLAMA 3-8B* achieved the highest precision of 0.72 and F1-score of 0.70. A high precision score indicates that a large proportion of the words generated by models (unigrams) are also found in the ground-truth summary, whereas a high F-1 score reflects a model's overall effectiveness in producing a summary that is accurate and covers the ground truth well. In contrast, the *GPT-4-Turbo* model scored the lowest with an F-1 score of 0.58 and precision of 0.50. This trend persisted in ROUGE-2, where *LLAMA 3-8B* led with an F1-score of 0.48 and precision of 0.50, significantly outperforming *GPT-4-Turbo*, which only achieved an F-1 score of 0.29 and precision of 0.24. Similarly, for ROUGE-L, which assesses the longest common sub-sequence between generated and reference texts, *LLAMA 3-8B* excelled with an F1-score of 0.55 and precision of 0.48, while *GPT-4-Turbo* lagged behind with an F1-score of 0.36 and precision of 0.31. BERTScore, which evaluates the semantic similarity between generated and ground-turth summaries, further corroborated these findings. *LLAMA 3-8B* stood out with the highest F-1 of 0.84 and precision on 0.87, reflecting its strong alignment in meaning with the ground truth summary. Conversely, *GPT-4-Turbo* recorded the lowest F-1 of 0.73, indicating its relative difficulty in generating semantically accurate summaries. The second best performing model with our approach is *Gemma-2-9B* with impressive F-1 score of 0.82, and precision of 0.82.

2) Expert Human Evaluation: Expert human evaluation plays a critical role in assessing the quality of generated summaries, especially as automatic metrics like ROUGE and BERTScore, though useful, may not always align with expert judgment [10], [41]. Recognizing these limitations, we incorporated human evaluation in this study. Given the expensive

---

[8]https://huggingface.co/openai/whisper-large

[9]https://huggingface.co/microsoft/deberta-xlarge-mnli

TABLE III: Statistics of ground-truth clinical summaries from 1,473 patient-doctor conversations.

| | Clinical Summaries Statistics | | | | | | | | | |
| Metric | Subjective | | Objective | | Assessment | | Plan | | Vocab/Token | |
	Sentences	Words	Sentences	Words	Sentences	Words	Sentences	Words	Vocab	Tokens
Count	48775	44927	1957	12330	1912	11494	2017	19396	77384	226246
Max	22	323	9	110	7	110	6	80	198	500
Mean	3.25	63.43	1.328	8.37	1.298	10.02	1.371	13.17	18.85	36.47

TABLE IV: Evaluation dataset statistics from staged patient-doctor conversations.

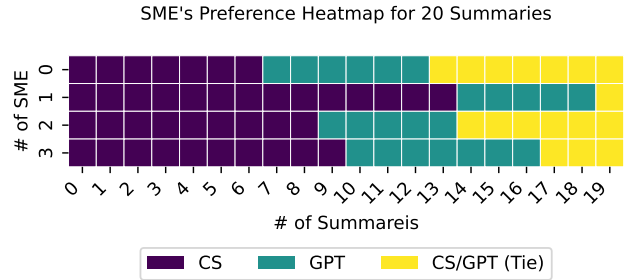| | Staged Conversation Statistics | | | | |
Metric	Sentences	Words	Char	Vocab	Tokens
Count	1997	20359	99518	6965	25690
Mean	99.85	1017.95	4975.9	348.25	1284.5
Max	146	1573	7886	448	1995
Min	61	551	2629	222	708

Fig. 4: Heatmap illustrating the preferences between summaries generated by CLINICSUM and GPT, along with ties indicating equal preference between the two.

nature of human evaluation, we assembled a panel of four SMEs, including medical resident doctors and physicians, to compare the summaries generated by CLINICSUM with those from GPT-based models with prompting techniques. Specifically, we focused on the best-performing models from both systems for zero-shot summarization. Using the same set of 20 conversations and summaries as in the quantitative analysis ensured a fair comparison.

Following the evaluation strategy outlined by Giorgi et al. [10], the SMEs were provided with ground-truth data, summaries from LLAMA-3-8B (best performing model) in our framework, and summaries from GPT-4-O-Mini (best performing among gpt models). Summaries were anonymized and labeled as 'A' and 'B' with the SMEs instructed to choose their preferred version or select both if there is a tie. While our evaluation strategy is similar to [10], there are some key differences. They relied on three criteria: *critical, non-critical*, and *irrelevant information* from previous research by Savkov et al. [42] to guide SME preferences. Additionally, they included the ground truth in their evaluation, while we focused solely on comparing the summaries generated by CLINICSUM and those from GPT-based models. In contrast, we introduced a fourth criterion: *factual correctness* (must capture all key factual information). Based on this, we redefine a good summary as one that is *"factually accurate, includes all critical information, some non-critical information, and contains minimal irrelevant details"*. The results of this evaluation are shown in Table VI, and we visualize the agreement between SMEs using heatmaps, as illustrated in Fig. 4.

Overall, the summaries generated by CLINICSUM are strongly preferred over summaries generated by the GPT-based models further validating the high performance reported by the automatic evaluation metrics. In addition, we assessed Inter-Rater Reliability (IRR) among the four SMEs using two statistical measures: Fleiss' Kappa (κ) [43] and Krippendorff's Alpha (α). The results, shown in Table VII, indicate moderate

agreement (0.41 to 0.60) for both κ and α. This suggests that while the SMEs were not perfectly aligned in their preferences, they demonstrated a fair level of consensus. In our opinion, this variability arises from differences in the SMEs' experience levels and subjective interpretations, which align with similarly low agreement scores reported in previous research [1], [9], [10].
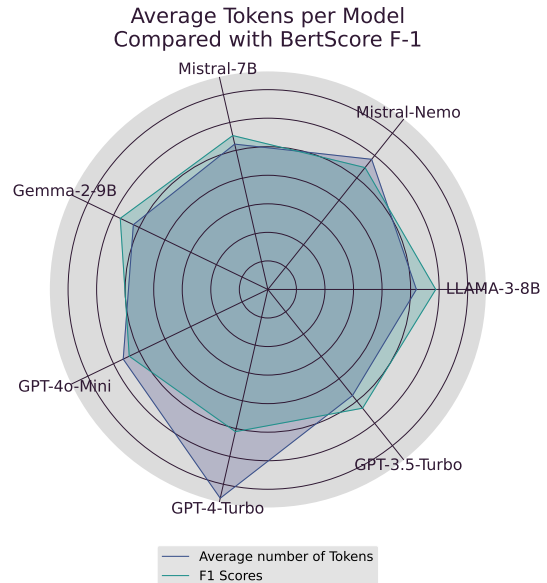

Fig. 5: Radar chart illustrating how different models compare in terms of two key metrics: the average number of tokens and F-1 scores of BertScore.

Findings with respect to # number of Tokens: In this study, we further investigated whether there is a quantitative

TABLE V: Comparison of GPT models using zero-shot prompting and CLINICSUM for generating clinical summaries, evaluated with lexical-based (ROUGE) and embedding-based (BERTScore) metrics.

Model	Rouge-1			Rouge-2			Rouge-L			BertScore		
	P	R	F-1	P	R	F-1	P	R	F-1	P	R	F-1
GPT-4-Turbo	0.50	0.72	0.58	0.24	0.35	0.29	0.31	0.45	0.36	0.73	0.74	0.73
GPT-4-0-Mini	0.64	0.66	0.64	0.38	0.39	0.38	0.45	0.46	0.45	0.76	0.78	0.77
GPT-3.5-Turbo	0.64	0.59	0.61	0.35	0.32	0.33	0.42	0.39	0.40	0.74	0.79	0.76
CLINICSUM LLAMA-3-8B	0.72	0.69	0.70	0.50	0.48	0.48	0.57	0.54	0.55	0.87	0.82	0.84
CLINICSUM Mistral-Nemo-12B	0.67	0.72	0.68	0.44	0.48	0.45	0.50	0.54	0.51	0.76	0.82	0.78
CLINICSUM Mistral-7B	0.70	0.69	0.68	0.46	0.46	0.45	0.51	0.50	0.49	0.75	0.83	0.79
CLINICSUM Gemma-2-9B	0.69	0.67	0.67	0.44	0.43	0.43	0.50	0.48	0.49	0.82	0.83	0.82

TABLE VI: An overview of human evalaution. Four SME's evaluated summaries generated by CLINICSUM (CS) and GPT-O-Mini (GPT). For each case, the SMEs selected their preferred summary. The win rate represents the percentage of cases where a summary was preferred, with ties excluded from the calculation.

SME	Preferred		Ties	Win rate %	
	CS	GPT	CS/ GPT	CS	GPT
1	7	6	7	0.54	0.46
2	14	5	1	0.74	0.26
3	9	5	6	0.64	0.36
4	9	8	3	0.53	0.47
Total	39	24	17	0.61	0.39

TABLE VII: IRR metrics, including Fleiss' Kappa (κ) and Krippendorff's Alpha (α), demonstrate moderate agreement among the four SMEs.

Inter-Rater Reliability (IRR)		
Overall Agreement	Fleiss Kappa (κ)	Krippendorff's Alpha (α)
	0.43746	0.44450

correlation between the best and worst performing models in terms of F-1 score and their average token count. Fig. 5 presents a radar chart comparing the average tokens per model with their corresponding BERTScore F-1 scores. Token counts for both GPT-based models and open-source PLMs were computed using the *BAAI/bge-large-en-v1.5* model from Hugging Face. A key finding from this analysis is the correlation between token count variability and accuracy: LLAMA-3-8B (avg token count: 260) and Gemma-2-9B (average token count: 262), which closely matches the ground truth (avg token 268), achieves the highest BERTScore, while GPT-4-Turbo (avg token count: 375), with significantly high token usage likely hallucinating, performs the worst. Mistral-Nemo-12B (average token count:292) and GPT-4-0-Mini (average token count: 281) also show increased token generation in some cases but have a tendency to generate fewer tokens, while Mistral-7B (average token count: 259) and GPT-3.5-Turbo (average token count: 237) generally produce fewer tokens, potentially missing critical details. Despite these variations, LLAMA-3-8B and Gemma-2-9B maintain a balanced approach of using and generating tokens that are strictly

coherent with the context provided to them by retriever-based filtering, producing summaries that are aligned with the ground truth.

This suggests that models generating summaries with token counts closer to the ground truth tend to produce more accurate outputs. In contrast, greater variability in token count, as seen in GPT-4-Turbo and GPT-3.5-Turbo, may lead to less accurate summaries. This conclusion is further supported by expert human assessment, where SMEs showed a clear preference for the summaries generated by LLAMA-3-8B over those produced by GPT-4-Turbo.

VI. LIMITATIONS AND DISCUSSION

Despite CLINICSUM's encouraging results in generating good clinical summaries, several limitations should be noted. First, the model's performance is highly reliant on the quality and diversity of the training data used in fine-tuning task. The fine-tuning dataset, comprising 1,473 conversations from the FigShare and MTS-Dialog datasets with clinical summaries generated by a PLM and validated by SMEs, is limited in scope, as it only encompasses a narrow range of medical specialties.

Another limitation is the use of simulated patient-doctor conversations in the evaluation phase. While these simulated conversations are useful for HIPAA compliance, they may fail to capture the complexities and variability of real-world clinical interactions. Consequently, the generated summaries may not perform as well in real-world clinical settings, where patient communication is less structured and more nuanced.

Furthermore, while the framework uses retrieval-based filtering to improve factual accuracy, the risk of hallucinations persists, especially when summarizing conversations with ambiguous or incomplete information. Ensuring factual accuracy is critical in medical settings, and additional validation mechanisms may be necessary to mitigate these risks. Moreover, while we added a criterion of factual correctness to the human evaluation process, the moderate IRR scores indicate that subjective interpretation among SMEs can still result in inconsistencies in summary evaluations. Another important aspect is Biases, PLMs trained on vast amounts of text data may inadvertently capture and reproduce biases present in the data. For example, it may over-prioritize common condition such

as "upper respiratory infections" when interpreting symptoms, potentially overlooking rarer but more serious conditions.

Lastly, the CLINICSUM's dependence on fine-tuning PLM necessitates substantial computational resources, potentially constraining its scalability in low-resource clinical settings. The current implementation is tailored to operate on a consumer GPU with 24GB of memory, which may not be readily available to all healthcare institutions. Due to limited computational resources, we focused our efforts on models with $\leq$ 12 Billion parameters. Fine-tuning a bigger PLM for example model with 30B, 70B or higher parameter may yield even better results.

VII. Conclusion & Future Work

In this paper, we demonstrated the feasibility of automatically generating clinical summaries directly from patient-doctor conversations using a framework with two module architecture referred to as CLINICSUM. The first module, retriever-based filtering, acts as an extractive component, identifying relevant portions of the transcript that contain subjective, objective, assessment, and plan information. The advantage of this approach is that it not only filters out unnecessary information from the transcripts but also reduces the risk of hallucination by passing only the relevant chunks to the second module. The second module, inference, utilizes the filtered information as context and uses a fine-tuned PLM to generate clinical summaries through abstraction.

We created a high-quality fine-tuning training dataset consisting of 1,473 conversation-summary pairs and used it to fine-tune four open-source PLMs with $\leq$ 12B parameters. Surprisingly, when combined with our framework for inference, these fine-tuned open-source PLMs substantially outperformed state-of-the-art GPT models in both automatic and expert human evaluations. Expert human assessments by SMEs confirmed that the summaries generated by CLINICSUM were more preferable than those produced by GPT models using prompting. We believe our results are encouraging, and CLINICSUM offers a promising solution for automating clinical summarization.

Future work will focus on expanding the both training and validation dataset, improving framework's scalability, and exploring real-world applications in diverse clinical settings. We also intend to further investigate methods to further reduce hallucination and potential biases of PLMs.

Acknowledgement

References

[1] S. Neupane, S. Mitra, S. Mittal, N. A. Golilarz, S. Rahimi, and A. Amirlatifi, "Medinsight: A multi-source context augmentation framework for generating patient-centric medical responses using large language models," *arXiv preprint arXiv:2403.08607*, 2024.

[2] P. Zhao, H. Zhang, Q. Yu, Z. Wang, Y. Geng, F. Fu, L. Yang, W. Zhang, and B. Cui, "Retrieval-augmented generation for ai-generated content: A survey," *arXiv preprint arXiv:2402.19473*, 2024.

[3] L. C. Mcguire, "Remembering what the doctor said: organization and adults' memory for medical information," *Experimental aging research*, vol. 22, no. 4, pp. 403–428, 1996.

[4] J. L. Anderson, S. Dodman, M. Kopelman, and A. Fleming, "Patient information recall in a rheumatology clinic," *Rheumatology*, vol. 18, no. 1, pp. 18–18, 1979.

[5] S. Kumar, "Burnout and doctors: prevalence, prevention and intervention," in *Healthcare*, vol. 4, no. 3. MDPI, 2016, p. 37.

[6] T. Knoll, F. Moramarco, A. P. Korfiatis, R. Young, C. Ruffini, M. Perera, C. Perstl, E. Reiter, A. Belz, and A. Savkov, "User-driven research of medical note generation software," *arXiv preprint arXiv:2205.02549*, 2022.

[7] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023.

[8] A. Ben Abacha, W.-w. Yim, Y. Fan, and T. Lin, "An empirical study of clinical note generation from doctor-patient encounters," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2291–2302. [Online]. Available: https://aclanthology.org/2023.eacl-main.168

[9] L. Zhang, R. Negrinho, A. Ghosh, V. Jagannathan, H. R. Hassanzadeh, T. Schaaf, and M. R. Gormley, "Leveraging pretrained models for automatic summarization of doctor-patient conversations," *arXiv preprint arXiv:2109.12174*, 2021.

[10] J. Giorgi, A. Toma, R. Xie, S. S. Chen, K. R. An, G. X. Zheng, and B. Wang, "Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models," *arXiv preprint arXiv:2305.02220*, 2023.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[12] B. Schloss and S. Konam, "Towards an automated soap note: classifying utterances from medical conversations," in *Machine Learning for Healthcare Conference*. PMLR, 2020, pp. 610–631.

[13] C. Sinsky, L. Colligan, L. Li, M. Prgomet, S. Reynolds, L. Goeders, J. Westbrook, M. Tutty, and G. Blike, "Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties," *Annals of internal medicine*, vol. 165, no. 11, pp. 753–760, 2016.

[14] L. Weed, "The problem oriented record as a basic tool in medical education, patient care and clinical research." *Annals of clinical research*, vol. 3, no. 3, pp. 131–134, 1971.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–40, 2023.

[17] C. Wang, X. Liu, Y. Yue, X. Tang, T. Zhang, C. Jiayang, Y. Yao, W. Gao, X. Hu, Z. Qi *et al.*, "Survey on factuality in large language models: Knowledge, retrieval and domain-specificity," *arXiv preprint arXiv:2310.07521*, 2023.

[18] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[19] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[20] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.

[21] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.

[22] M. Valipour, M. Rezagholizadeh, I. Kobyzev, and A. Ghodsi, "Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation," *arXiv preprint arXiv:2210.07558*, 2022.

[23] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[24] C.-W. Goo and Y.-N. Chen, "Abstractive dialogue summarization with sentence-gated modeling optimized by dialogue acts," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 735–742.

[25] M. Li, L. Zhang, H. Ji, and R. J. Radke, "Keep meeting summaries on topic: Abstractive multi-modal meeting summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2190–2196.

[26] Y. Zhang, D. Y. Ding, T. Qian, C. D. Manning, and C. P. Langlotz, "Learning to summarize radiology findings," *arXiv preprint arXiv:1809.04698*, 2018.

[27] G. Finley, E. Edwards, A. Robinson, M. Brenndoerfer, N. Sadoughi, J. Fone, N. Axtmann, M. Miller, and D. Suendermann-Oeft, "An automated medical scribe for documenting clinical encounters," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2018, pp. 11–15.

[28] S. Enarvi, M. Amoia, M. D.-A. Teba, B. Delaney, F. Diehl, S. Hahn, K. Harris, L. McGrath, Y. Pan, J. Pinto *et al.*, "Generating medical reports from patient-doctor conversations using sequence-to-sequence models," in *Proceedings of the first workshop on natural language processing for medical conversations*, 2020, pp. 22–30.

[29] K. Krishna, S. Khosla, J. P. Bigham, and Z. C. Lipton, "Generating soap notes from doctor-patient conversations using modular summarization techniques," *arXiv preprint arXiv:2005.01795*, 2020.

[30] S. Ramprasad, E. Ferracane, and S. P. Selvaraj, "Generating more faithful and consistent soap notes using attribute-specific parameters," in *Machine Learning for Healthcare Conference*. PMLR, 2023, pp. 631–649.

[31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[32] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford *et al.*, "Okapi at trec-3," *Nist Special Publication Sp*, vol. 109, p. 109, 1995.

[33] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," *arXiv preprint arXiv:2004.04906*, 2020.

[34] H. Tripathi, "Experimental approach toward training and analysing siamese deep neural network for sentence with no repeated expressions," in *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2023, pp. 1–5.

[35] G. V. Cormack, C. L. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms condorcet and individual rank learning methods," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 758–759.

[36] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[37] Meta, "How to fine-tune: Focus on effective datasets," https://ai.meta.com/blog/how-to-fine-tune-llms-peft-dataset-curation/, 2024.

[38] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Stanford alpaca: An instruction-following llama model," https://github.com/tatsu-lab/stanford_alpaca, 2023.

[39] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[40] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[41] D. Deutsch, R. Dror, and D. Roth, "Re-examining system-level correlations of automatic summarization evaluation metrics," *arXiv preprint arXiv:2204.10216*, 2022.

[42] A. Savkov, F. Moramarco, A. P. Korfiatis, M. Perera, A. Belz, and E. Reiter, "Consultation checklists: Standardising the human evaluation of medical note generation," *arXiv preprint arXiv:2211.09455*, 2022.

[43] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.