Harnessing AI for Primary Care: Creating a Medical Summarizer with Fine-Tuned LLMs for Efficient Patient Outcomes

Eric Shen and **Matthew Zhang**
University of Berkeley MIDS

Abstract

Clinicians dedicate a significant amount of time detailing discharge notes of various patients, yet oftentimes these notes encompass lengthy and complex characteristics making it difficult for others to extract valuable information, especially at larger health institutions with thousands of patients. The capability of generating concise and accurate summaries of clinical records is crucial for enhancing the efficiency of healthcare providers and ensuring comprehensive patient care. In this study, we dive into the adaptation methods of existing fine-tuned LLM architectures and investigate the efficacy of various state-of-the-art text summarization models on the MIMIC IV dataset, which is a de-identified electronic health records database of over 40,000 intensive care unit patients collected at the Beth Israel Deaconess Medical Center from 2008 to 2019; with a focus on discharge summaries and diagnoses. We present our fine-tuned text summarization model that has been built upon Google's T5 model which outperformed our previous model iterations by achieving higher ROUGE scores and more accurate summarization quality. Our study highlights the potential of fine-tuned models in generating high-quality medical summaries, thereby aiding healthcare providers in quick and efficient patient data review.

1 Introduction

Clinical documentation is an essential yet often burdensome aspect of healthcare, requiring meticulous attention to detail and consuming significant amounts of time for clinicians. With the widespread adoption of electronic health record (EHR) systems across U.S. healthcare facilities, documentation workloads have only continued to balloon. A 2016 study found that for every hour clinicians spent with patients, nearly 2 additional hours were spent on EHR documentation (Sinsky et al., 2016). Moreover, Forde-Johnston et al. (2023) found that EHR use diverted nurses' attention away from their patients and reduced open nurse-patient communication.

The absence of a standardized system for compiling and managing clinical notes, coupled with the limited availability of medical staff—who are typically engaged in more critical tasks—leads to two significant challenges: (1) the inconsistency of data originating from diverse sources (e.g., various hospitals, departments, or individual healthcare providers) and (2) the prevalence of abbreviations and typographical errors. These issues, along with the complexities of biomedical text, such as advanced medical jargon and variant spellings, create substantial obstacles for the automated extraction of information from these documents.

In this study, we address this limitation by employing state-of-the-art, pre-trained, open-source text generation models, specifically Google's Pegasus and T5, as well as Meta's BART, to generate coherent and accurate management plans and discharge instructions. By leveraging these advanced models, our approach seeks to alleviate the documentation burden on healthcare providers, allowing them to dedicate more time to quality patient care. We hope the implementation of such models not only optimizes the documentation process, but also contributes to a broader goal of integrating AI to support clinical workflows and improve our healthcare system as a whole.

2 Background/Related Work

Prior machine learning frameworks for extracting meaningful information from unstructured clinician notes have largely centered around specific tasks, such as assigning International Classification of Diseases (ICD) codes, predicting mortality rates, and assessing readmission risks.

Blecker et al. (2016) showed that unstructured clinical notes could be effectively utilized with machine learning models to identify patients with

heart failure. By incorporating unstructured clinical text into their logistic regression classifier model, they significantly enhanced the model's accuracy, demonstrating the valuable insights that can be extracted from EHR notes. Other research has shown that machine learning models can effectively leverage the unstructured clinical narrative for the prediction of clinical outcomes (Jain et al., 2019). With the addition of attention mechanisms to long short-term memory networks (LSTMs), Jain et al. (2019) saw improved performance in predicting clinical outcomes such as mortality and ICU readmission, albeit with attention weights being less interpretable. Additionally, previous work has also demonstrated success in predicting ICD code assignments using clinical notes from the MIMIC-III dataset with deep learning models (Mullenbach et al., 2018; Sadoughi et al., 2018). Mullenbach et al. (2018) enhanced a convolutional model with a per-label attention mechanism, resulting in improved performance and greater interpretability as assessed by experts. Sadoughi et al. (2018) further refined this model by employing multiple convolutions of varying widths followed by max-pooling across channels before applying the attention mechanism.

In the last few years, the rapid advancement of large language models (LLMs) has revolutionized the NLP field, enabling models with superior capabilities for information retrieval and text generation. The development of conversational diagnostic AI, as explored by researchers at Google DeepMind (Tu et al., 2024), underscores the increasing interest and progress in creating AI systems capable of engaging in meaningful interactions with healthcare professionals and patients. Although substantial advancements have been made in medical question answering—exemplified by Google's MedPaLM 2, which achieved human expert-level performance on US Medical Licensing Examination (USMLE) style questions (Singhal et al., 2023)—there has been comparatively less focus on alleviating the workload associated with drafting detailed management plans and discharge summaries.

3 Methods

3.1 Dataset

The complexity of medical terminology and the need for extensive domain-specific knowledge require LLMs to be fine-tuned on specialized clinical datasets, which are often difficult to access due to privacy regulations and the sensitive nature of medical data.

For our project, we leveraged the MIMIC-IV-Note dataset, a comprehensive compilation of 331,794 de-identified discharge summaries from 145,915 patients admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts (Johnson et al., 2023). The MIMIC datasets are highly regarded in the clinical research community, with over 7000 citations in total (Johnson et al., 2016, 2023).

While the initial dataset contained 331,794 summaries, we had to take into consideration the computational cost and efficiency available to us during the duration of this study. Thus, we decided to randomly sample approximately 3% of the original data (10,000 summaries) to develop our model. For our experiments, we then split our cohort into training, validation, and testing splits following an 80/10/10 split.

Each summary in our dataset follows a semi-structured template that includes various sections such as patient demographics, medical history, brief hospital course, physical exam findings, pertinent results, and discharge instructions, with the latter always being the final section. For preprocessing, we cleaned and standardized the text by removing excess whitespace and newline characters. Following a similar format to Xu (2024), we used Regex pattern matching to extract the final "Discharge Instructions" section to use as our labels, with the preceding content being used as our training inputs. However, unlike Xu (2024), who extracted both the "Brief Hospital Course" and "Discharge Instructions" sections to use as separate output labels, we hypothesized that including the Brief Hospital Course (BHC) summary as part of the training input instead would provide valuable information for generating higher quality discharge instructions for a given patient.

3.2 Baseline

For our baseline model, we selected the Pegasus-XSum open source model due to its low-resource summarization capabilities and specialized training as an abstractive summarizer (Zhang et al., 2020). This model's ability to generate succinct and coherent summaries was initially appealing for our goal of efficiently producing detailed discharge summaries from clinical notes. However, given its extreme summarization architecture, we quickly

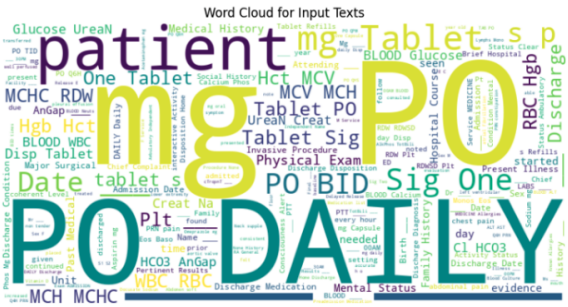

Figure 1: Word Cloud for Input Texts

realized that the brevity of the summaries generated by Pegasus-XSum posed challenges for evaluation using BLEU scores.

The model's tendency to produce very short summaries severely limited the overlap with reference summaries, resulting in BLEU scores of less than 0.001. This result highlighted the both the need for a different evaluation metric and a different model that could produce longer and more detailed summaries, aligning better with the practical needs of clinical documentation.

3.3 Approach and Modeling

Our formal approach uses both Google Research's T5 and Meta's BART models fine-tuned on the MIMIC-IV discharge notes to generate abstractive summaries of the text. Abstractive summarization allows the ability to generate novel sentences that capture the essence of the source text, which we hope to leverage in creating concise and coherent summaries in the medical domain. As mentioned previously, although the dataset contained no reference summaries, we decided for the purposes of evaluation to split the discharge notes into input texts and output texts containing summaries of the patient's condition for training and instructions for evaluation, respectively. Our primary focus shifts to bart-base and bart-large-xsum where the latter is specifically pre-trained on the XSum dataset, which consists of highly abstractive summaries (Lewis et al., 2019). This pre-training allows the model to learn the intricacies of generating concise, coherent, and contextually rich summaries, which is directly applicable to the task of summarizing clinical notes.

In comparison to our specialized Pegasus baseline, we determined that these models would be much more efficient for our task as alternative encoder-decoder architectures in reading input with the encoder, training the decoder as a generator and thus producing a novel summary. Since T5 and BART have balanced architectures and are pre-trained as highly versatile models and robust to noise, they result in more flexible transfer learning capabilities and therefore, have a higher potential overlap for our evaluation metrics between the generated summaries and original text.

In addition to preprocessing our data, each clinical note was tokenized into a sequence of tokens using the respective tokenizers acquired from T5 and BART as a form of suitable input to the transformer models. This process ensures the input text to be appropriately truncated and padded to match the model's expected input length. Moreover, the tokenized input includes attention masks to differentiate between actual tokens and padding tokens, ensuring the model focuses on meaningful content. Both T5 and BART models employ multi-head self-attention mechanisms to capture dependencies between tokens across the entire sequence. This mechanism allows the models to focus on different parts of the input text simultaneously, capturing contextual information crucial for generating coherent summaries (Kanwal and Rizzo, 2022).

3.3.1 Fine-Tuning

Fine-tuning plays a crucial role in adapting the pre-trained models to our summarization task. We fine-tuned both T5 and BART models on our sample of the MIMIC-IV using consistent configurations and hyperparameters of maximum sequence length 512, batch size 2, epochs 2 and keeping the remaining parameters the same as pre-training. This process adjusts the weights of the pre-trained models based on the MIMIC-IV dataset, optimizing their performance for clinical summarization (Kanwal and Rizzo, 2022).

Both T5 and BART models employ multi-head self-attention mechanisms to capture dependencies between tokens across the entire sequence. This mechanism allows the models to focus on different parts of the input text simultaneously, capturing contextual information crucial for generating coherent summaries. The T5 model uses self-attention to encode the input sequence and cross-attention to generate the output summary. The attention mechanism helps the model to effectively understand and transform the input text into a summary. BART utilizes a similar attention mechanism with bidirectional encoding and autoregressive decoding, ensuring that the generated summaries are both contextually accurate and fluent.

3.3.2 Evaluation

Evaluation in summarization tasks has proved to be ambiguous as there is a lack of consistency in an objectively indicative metric. To evaluate the performance of the fine-tuned models, we used content-based evaluation, looking at a combination of ROUGE scores, which measures the overlap of n-grams between the generated summaries and the reference summaries, as well as model loss; it considers both precision and recall, hence being more comprehensive than BLEU. As this is a more complex and specialized task that requires a strong accuracy in summarizing medical records, we decided to use a combination of assessment techniques to determine the effectiveness of our model. The first two are the ROUGE-1 and ROUGE-2 scores; which capture the overlap of n-grams (1 or 2) between the generated summary and the reference summary. These two metrics particularly can help us understand if key medical concepts and terms are being captured in the generated summaries. ROUGE-L helps measure the longest common subsequence between the candidate and reference summaries. With this, we are able to grasp sentence-level similarity by looking at any medical terminology or specifically, instructions that occur in the same order. Finally, we want to look at evaluation loss performance to confirm a robust model that can generate medical summaries on unseen data. The evaluation provided insights into the models' abilities to capture the essence of the clinical notes and produce meaningful summaries.

4 Results

We have presented results comparing our baseline abstractive summarization models of un-trained Pegasus. First, we note that Pegasus-Xsum is inherently simplistic in nature and is specialized for extreme summarization tasks. While we evaluate the BLEU score to be almost negligible being <1, we also note that the produced summaries are much shorter in length and fail to capture much of the information that is presented in the discharge notes. Looking at the ROUGE-N scores of 0.06 and 0.008, we can conclude that the model is hardly able to measure the overlap of n-grams between generated and reference summaries.

Next, we investigate the performance of T5. Although T5 has a sequence-to-sequence architecture and has been pre-trained to handle a variety of different tasks, for our purposes we will focus on its summarization capabilities; this is defined by a 'summarization:' prefix as part of the input prior to training. We wanted to observe the results as a model that was originally pre-trained for versatility in NLP problems. The t5-base model produced ROUGE-1 and ROUGE-2 scores less than our bart-base model which will be discussed in the following section. However, we were able to still see significant improvements from our original baseline model with an evaluation loss of 1.11.

Lastly, we decided to learn towards Meta's BART model for fine-tuning consisting of the bart-base model and bart-large-xsum trained on the XSUM corpus. The proposed architecture demonstrates significant improvements to baseline approaches as well as visual improvements on the generated summaries. As mentioned in the *Evaluation* section abovem ROUGE-1 and ROUGE-2 measure the overlap of n-grams between the generated and reference summaries. With bart-large-xsum, we produced a large ROUGE-1 score of 0.42 and ROUGE-2 score of 0.17; which indicates that 42% and 17% of the unigrams and bigrams in the reference summaries are present in the generated summaries, respectively. We find this to be a valuable metric as clinical discharge notes contain crucial keywords that need to be accurately included and can significantly impact the understanding of discharge instructions. This is an improvement to the previous iteration of our fine-tuned BART model, bart-base, with ROUGE-1 and ROUGE-2 scores of 0.38 and 0.14, respectively. In addition, the ROUGE-L score of 0.28 improves upon the same model with a ROUGE-L score of 0.25. In this study, ROUGE-L is valued for its largest common subsequence of between the summaries to reflect the overall structure and coherence; this ensures that the notes remain understandable and important information is retained between the doctor and patient. As for the model loss, we also see significant improvement between the two models on evaluation loss. Our model produced an evaluation loss of 0.57, which is lower than both the training and validation losses of 0.84 and 0.62, respectively. We conclude that the model is able to perform on unseen data, indicating generalization capabilities.

Using a combination of the ROUGE-1, ROUGE-2, ROUGE-L and evaluation loss metrics, we are able to provide a comprehensive evaluation of the summarization model. The discharge notes in the

	ROUGE-1	ROUGE-2	ROUGE-L	EVALUATION LOSS
PEGASUS-XSUM (BASELINE)	0.06	0.01	0.04	7.46
T5-BASE	0.31	0.10	0.19	1.11
BART-BASE	0.38	0.14	0.25	0.75
BART-LARGE-XSUM	0.42	0.17	0.28	0.57

Figure 2: Quantitative Evaluation of Our Model Results Using Relevant Metrics

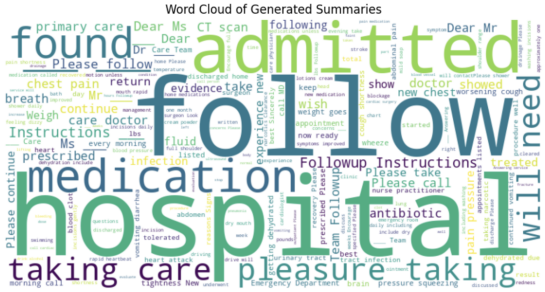

Figure 3: Word cloud of generated summaries from the fine-tuned BART model

MIMIC-IV dataset remain complex and contain critical information that needs to be accurately conveyed in summaries. These metrics communicate the performance of our model and ensure informative, accurate and easy to understand summaries for the patients to read and digest. Overall, the attention mechanism poses great abstraction power for the summarization task. The BART-large-xsum model shows strong potential for summarizing MIMIC-IV clinical discharge notes with a high ROUGE-1 score and good generalization capabilities, as indicated by the evaluation loss. However, we note the model's limitations and ability to approve in areas of training and validation, ROUGE-2 score and potential pre-training.

5 Conclusion

In our research, we explore the fine-tuning of transformer-based text models—namely Pegasus, Google's T5, and Meta's BART—for the abstractive summarization of clinical discharge notes from the MIMIC-IV dataset. Our findings demonstrate that these models are effective in generating coherent and concise summaries, with significant potential for enhancement through more extensive fine-tuning and specialized pre-training in clinical contexts. The integration of attention mechanisms has improved model interpretability, offering insights into the areas of focus within the generated summaries. Among the models tested, BART-large-xsum achieved the highest quality summaries, evidenced by superior ROUGE scores compared to the original text. This approach not only maintains the original context better but also provides more accurate summaries, which is critical for the medical field.

However, there are limitations to our current approach. Given our computational constraints, we were only able to sample a small subset of the MIMIC-IV corpus, which contains a wealth of additional medical information. Additionally, our study does not encompass all clinical document types, making it hard to generalize our results. For example, our progress notes task relies on ICU notes from a single medical center, which may differ in structure and length from non-ICU notes or those from other centers.

Future work should focus on expanding our evaluation to other summarization tasks, incorporating more comprehensive and varied datasets, and exploring methods to enhance model context length. By addressing these limitations, we can advance the effectiveness of clinical summarization models, reduce the documentation burden, improve technical comprehension, and ultimately enhance patient care.

References

Saul Blecker, Stuart D. Katz, Leora I. Horwitz, et al. 2016. Comparison of approaches for heart failure case identification from electronic health record data. *JAMA Cardiology*, 1(9):1014–1020.

Claire Forde-Johnston, Deborah Butcher, and Helen Aveyard. 2023. An integrative review exploring the impact of electronic health records (ehr) on the quality of nurse-patient interactions and communication. *Journal of Advanced Nursing*, 79(1):48–67.

Sarthak Jain, Ramin Mohammadi, and Byron C. Wallace. 2019. An analysis of attention over clinical notes for predictive tasks.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. Mimic-iv-note: Deidentified free-text clinical notes (version 2.2).

Alistair Johnson, Tom Pollard, Lu Shen, et al. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Neel Kanwal and Giuseppe Rizzo. 2022. Attention-based clinical note summarization. Accessed on 28 February 2022.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text.

Najmeh Sadoughi, Greg P. Finley, James Fone, Vignesh Murali, Maxim Korenevski, Slava Baryshnikov, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. 2018. Medical code prediction with multi-view convolution and description-regularized label-dependent attention.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models.

Christine Sinsky, Linnea Colligan, Ling Li, Mirela Prgomet, Susan Reynolds, Lori Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties. *Annals of Internal Medicine*, 165(11):753–760. Epub 2016 Sep 6.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. 2024. Towards conversational diagnostic ai.

Jing Xu. 2024. Discharge me: Bionlp acl'24 shared task on streamlining discharge documentation (version 1.3).

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.