Clinical Text Summarization using NLP Pretrained Language Models: A Case Study of MIMIC-IV-Notes

Conference Paper · November 2024

CITATIONS
0

READS
321

3 authors, including:

Oluwatomisin Arokodare
Georgia Southern University
6 PUBLICATIONS **3** CITATIONS
SEE PROFILE

Hayden Wimmer
Georgia Southern University
79 PUBLICATIONS **601** CITATIONS
SEE PROFILE

Clinical Text Summarization using NLP Pretrained Language Models: A Case Study of MIMIC-IV-Notes

Oluwatomisin Arokodare
oa03242@georgiasouthern.edu
Georgia Southern University
Statesboro, GA, 30460, USA

Hayden Wimmer
hwimmer@georgiasouthern.edu
Georgia Southern University
Statesboro, GA, 30460, USA

Jie Du
dujie@gvsu.edu
Grand Valley State University
Allendale, MI, 49401, USA

Abstract

As the amount of data available in the health sector continues to grow in the era of information overload, it becomes increasingly crucial than ever to communicate essential information concisely. The vast amount of textual data from electronic health records can overwhelm healthcare professionals, reducing the time they can dedicate to patient care. A key challenge is creating comprehensive medical history summaries during patient admissions which integrate various documents including the history of present illness, discharge condition and medications, and discharge instructions. The need to address this challenge is urgent, as effective summarization of health records can greatly improve patient outcomes, enhance clinical decision-making, and facilitate access to knowledge. This study highlights the utilization of large language models trained to produce concise summaries through machine learning and national language processing algorithms. These models offer a promising avenue for summarizing patients' primary health concerns from daily progress notes, thereby streamlining information in hospital settings concisely and aiding diagnostic processes. In this study, we utilized pre-trained transformer models, including BART, T5, and Pegasus, to summarize patient medical histories. We evaluated the performance of those models using metrics including BLEU, ROUGE, and BERT scores on de-identified clinical notes from MIMIC-IV. Our experimental results show that BART and Pegasus models performed efficiently among the three large language models. The combination of these three models produced the most efficient summaries for clinical notes, given that the summary length generated by the model was shorter than the original medical history text for each medical case.

Keywords: Automatic text summarization, Natural Language Processing, Large language models, MIMIC-IV-clinical notes.

1. INTRODUCTION

Text summarization has advanced since the 1950s to support efficient data processing in response to the growing need. It is becoming especially important in the healthcare industry, where lengthy and specialized medical reports can make it difficult to understand. The significant rise in health sector data presents challenges for healthcare practitioners that impact patient care decisions. When patients are admitted into the hospital, one of the documents written by clinical professionals to conclude their treatment is a medical history summary report. Various medical reports, including history of present illness, brief hospital course, discharge instruction, discharge medication, and general healthcare information, contribute to comprehensive record-keeping. These documents serve as valuable references for future doctors, which helps enhancing their understanding of patients' circumstances during treatment.

However, according to HealthIT.gov (2021) the official website of the United States Health Information Technology Department highlights prevailing challenges in hospitals regarding the digital exchange of health information. Text summarization can extract essential information from complex medical reports without compromising their essence. Text summarization offers a more accessible, concise understanding of information and facilitates better communication between medical experts who generate reports and patients.

In this study we use pre-trained transformer models to summarize patient medical health histories. Our goal is to evaluate their effectiveness in summarizing medical texts and perform a comprehensive analysis on the models to identify the model that can produce succinct, coherent, and precise medical history summaries. This paper is structured into seven sections: Background information on text summarization is provided in Section 2. Related work is highlighted in Section 3. The methodology of the study is outlined in Section 4. The experimental analysis is presented in Section 5. Section 6 highlights the results discussion. Section 7 concludes with a summary of the findings and future work.

2. BACKGROUND

According to Bijal et al. (2017), text summarization is the process of condensing long text into shorter, comprehensible phrases while retaining essential information. It is crucial for managing the vast volume of online content and data including emails, movie reviews, news headlines, student notes, and more. Automation and artificial intelligence have become indispensable since they save time and provide important information so that readers may choose whether to continue or not. It plays a crucial role in regulating the deluge of information and assists users in determining whether to interact with material.

Text summarization techniques are divided into two categories: extractive text summarization and abstractive text summarization (Bhatia & Jaiswal, 2015). In this study we will be considering abstractive text summarization because it reformulates the text from the source text to generate new sentences that express the text's major concepts in a more streamlined and clear manner. Unlike extractive summary, which chooses phrases straight from the source text to generate a summary, abstractive text summarization will rephrase and condense subject matter resulting in summaries that are simpler to read and comprehend while retaining overall meaning (see Figure 1). Gaikwad and Mahender (2016) generated sentences using keywords with the aim of minimizing redundancy and produce accurate summaries. To accomplish this, a comprehensive grasp of the text's context and significance is imperative, along with the ability to rephrase and paraphrase it without compromising its essence. Natural language processing (NLP) interprets and understands the content of a document or text to generate an abstract text summary. Abstract summarization, though capable of generating concise and coherent summaries, typically poses greater challenges due to the requirement for advanced natural language generation techniques.

3. LITERATURE REVIEW

This section discusses the relevant literature used as the basis of our methodology. Specifically, these studies address the use and combination of the abstractive large language models for text summarization.

In the context of the increasing amount of online content, Batra et al. (2020) discussed the importance of text summarization tools. These tools provide concise summaries, which allows readers to decide whether to dig deeper into the content or not. There is a growing need for text summarization to handle complex language as the volume of information on the Internet is increasing and it is increasingly difficult to extract relevant information manually. To assess their effectiveness, popular techniques such as the

Encoder-Decoder Model with Attention, the Pointer Generator, the Pointer Generator with Coverage, UniLM, and BERTSUMABS are analyzed. As part of the study, UniLM is highlighted as one of the models examined using ROGUE metrics. These metrics are applied to the CNN/Daily Mail dataset, and the results are compared with reference summaries. For each model, ROGUE metrics were used to evaluate its results (ROGUE 1, ROGUE 2, ROGUE L), and the scores were compared on CNN/Daily Mail datasets showing overlap and common subsequence statistics.

Tsai et al. (2022) tackled privacy concerns and lack of public datasets in studying outpatient conversations. To address this, they proposed a three-step framework for summarizing outpatient conversations using Transformer-based models and external medical data. The long outpatient conversions are summarized through dialogue segmentation, dialogue summarization, and writing style conversion. The Multilingual T5 (mT5) model was used to summarize longer inputs despite limited training data. The technique yields steady performance in various tasks, as demonstrated by the experimental findings using pre-trained models.

Van Veen et al. (2023) evaluated eight Large Language Models (LLMs) for clinical text summarization from electronic health records (EHR). The experiments included quantitative evaluations and a clinical reader study to assess LLM performance and potential improvements in healthcare workflows. Adaptation methods are highly important in the study, which shows that even one in-context example significantly improves performance. When sufficient in-context examples are provided, proprietary models GPT-3.5 and GPT-4 consistently outperform open-source models. In all metrics across datasets, Sequ2seq models (FLAN-T5, FLAN-UL2) outperform autoregressive models (Llama-2, Vicuna), with GPT-4 achieving the highest performance on all metrics. FLAN-T5 excels in syntactical metrics. The results show that LLM-generated summaries often surpass human experts in completeness, correctness, and conciseness.

H. Zhang et al. (2019) introduced a neural network framework with an encoder-decoder architecture for summarizing multiple sentences in a document. The two-stage encoder-decoder framework combines BERT to encoding input sequences and Transformer-based decoding to predicting words sequentially. The model uses pre-trained contextualized language models to enhance performance without manual features, maximizing the likelihood of generating accurate summaries. The model demonstrates improved performance on CNN/Daily Mail and New York Times datasets, achieving state-of-the-art performance on CNN/Daily Mail with a score of 33.33 on ROUGE-1, ROUGE-2, and ROUGE-L, and a 5.6% relative improvement in ROUGE-1 on the New York Times dataset. The NYT50 corpus generates longer summaries than CNN/Daily Mail, and the model captures long-term dependencies effectively. The model performs better across diverse data distributions than other methods, with significant improvements observed in ROUGE-1 and 0.51 in ROUGE-2.

Yang et al. (2022) highlighted the significance of NLP powered by clinical language models and focused on utilizing artificial intelligence (AI) for processing digital health records. They presented GatorTron, a huge clinical transformer model based on a corpus of more than 90 billion words from UF Health, PubMed, the website Wikipedia, and MIMIC III. When tested on five clinical Natural Language Processing (NLP) tasks, GatorTron trained with different parameter sizes consistently outperformed previous clinical and biological transformers. The findings suggest that increasing the number of models and training data can greatly enhance medical AI system performance, which may have consequences for the provision of healthcare.

The advancements in neural network technologies and the availability of large amounts of data are responsible for the rise of summarization models in information technology (Kryściński et al., 2019). The methods used nowadays include hybrid extractive-abstractive models, multi-task training, copying mechanisms, attention mechanisms, and reinforcement learning. Despite these developments, benchmarks such as the CNN/Daily Mail news corpus have not advanced as much as they formerly did. Uninformative assessment processes and uncurated datasets in research setups are to blame for this stagnation. A more reliable setup for text summarization, with special emphasis on analyzing datasets, assessment measures, and model outputs is needed. Large-scale summarizing model assessment is labor-intensive whether done manually or semi-automatically. This has prompted the creation of automatic measures like the ROUGE package, which evaluates the degree of lexical similarity between prospective and reference breakdowns.

4. METHODOLOGY

In our study, we conducted experiments on distinct datasets MIMIC-IV-Note to assess the performance of various abstractive Large Language Models (LLMs) in the context of text summarization. These datasets serve as the foundation for our evaluation and comparison of the summaries generated by large language model. Figure 2 appendix Item captures the framework of our methodology. We employed pre-trained transformer models, including BART, T5, and Pegasus, which were used to summarize patient medical histories from the dataset. We evaluated the performance of those models using standard metrics presented in the LLM literature which includes BLEU, ROUGE, and BERT with the equations for their respective calculations detailed in the experimental results. Our experimental results show that BART and Pegasus models performed efficiently among the three large language models. The combination of these three models produced the most efficient summaries.

Dataset

The dataset used in this research is MIMIC-IV-Note: de-identified free-text clinical notes. According to Johnson et al. (2023), the dataset contains 357,289 discharge summaries and 2,471,881 radiology reports from 161,403 patients admitted to the Beth Israel Deaconess Medical Center in Boston, MA, United States. The dataset belongs to the Medical Information Mart for Intensive Care (MIMIC), and it has protected health information removed in accordance with HIPAA Safe Harbor provisions. The dataset consists of unstructured text data, and it is intended to stimulate research in clinical natural language processing and related areas, providing context to the clinical data within the MIMIC-IV database (Johnson et al., 2018). It includes a diverse set of clinical notes, which include a wide range of medical information such as patient present illness histories, discharge conditions and instructions, diagnoses, discharge medication, and treatment plans.

Preprocessing of the MIMIC-IV-Clinical Note Text Dataset

The dataset contains information about patient discharge for hospitalizations. These are long form narratives which describe the reason for a patient's admission to the hospital, their hospital course, their health history, and any relevant discharge instructions. In this study, we focused on the medical health history of patients.

According to Johnson et al. (2018), the steps in the preprocessing involve:

1) eliminating empty and/or duplicate clinical notes, converting all text to UTF-8 encoding, and removing any invalid UTF-8 sequences,
2) standardizing special characters,
3) tokenization-dividing the medical text into smaller units like words or phrases,
4) performing normalization - we ensured text uniformity by converting it to lowercase and removing unnecessary spaces and expanding contractions,
5) performing lemmatization to reduce the words to their base form to handle variations,
6) assigning grammatical categories to words and grouping them based on their grammatical structure,
7) removing details that are irrelevant to the analysis or could potentially identify the patient, and masking any identifiers that could link the data back to a specific patient,
8) ensuring consistent tokenization so that the same entities are consistently anonymized throughout the dataset, and
9) ensuring that the de-identified data comply with relevant privacy regulations such as Health Insurance Portability and Accountability Act (HIPAA) while retaining the useful information for the analysis.

Large Language Models

Abstractive approach allows for enhanced comprehension and coherence. This is particularly important in clinical contexts where the summaries need to be easily understandable by medical professionals. Therefore, we focus on the abstractive models in this study. We aim to evaluate and assess the efficacy of three finely tuned state-of-the-art abstractive text summarization models: BART, T5, and Pegasus, each of which was trained on our dataset.

A detailed overview of the three LLMs employed for the patient history clinical text summarization is presented in this section. These models represent advanced, large-scale NLP models that can understand and generate human-like language using complex machine learning techniques and extensive training on text data.

Bidirectional and Auto-Regressive Transformers (BART) Model

The BART model is a type of transformer-based neural network architecture introduced by Facebook AI Research and is designed mainly for text generation, summarization, and translation

tasks. In 2019, the BART model combines two popular architectures elements: BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) models, enabling it to be fine-tuned on small, supervised datasets for domain-specific tasks (Devlin et al., 2018). It generates autoregressive sequences with an autoregressive decoder and records contextual information from each side of a sequence using a bidirectional encoder. BART model is more effective than BERT and GPT-1, with 140 million parameters, because of its special mix of autoregressive generation and bidirectional context awareness. The encoder-decoder mechanism that makes up BART's architecture is used to mask or remove input tokens during preprocessing, which results in an inaccurate representation of the sequence (Arokodare & Wimmer, 2023). Subsequently, the corrupted form of the sentence is rebuilt, the corrupted input is mapped to a latent representation by the encoder, and the original phrase is generated by the decoder using this representation.

Text-to-Text Transfer Transformer (T5) Model

As introduced by Google AI researchers in 2019, the Text-to-Text Transfer Transformer (T5) model is a transformer-based neural network architecture. All tasks are framed as converting one textual input into another textual output, which is called a "text-to-text" approach. It ensures accuracy across tasks by learning to translate input and by minimizing a loss function. According to Roberts et al. (2019), transfer learning is a potent technique designed for a variety of natural language processing tasks which involves pre-training a model on a data-rich task before fine-tuning it for downstream tasks. T5 is trained on a range of tasks and datasets using a unified text-to-text architecture. By providing appropriate input-output pairs during training, it can handle a broad variety of tasks and obtain state-of-the-art results throughout a wide range of language comprehension tasks, such as translation, summarization, question answering, text classification etc. T5 has encoder and decoder layers and is comparable to those of BERT and GPT. It has high reliability and adaptability and has been utilized in many benchmarks and applications.

Pegasus Model

In 2020, Google AI researchers developed the transformer-based neural network model, PEGASUS. Its purpose is to produce precise and succinct summaries of lengthy papers or articles. It has been tailored for abstract text summarization. PEGASUS expects masked sentences from an input document using a pre-training task known as "gap-sentence generation" and a self-attention mechanism. This enables the model to comprehend sentence interactions and provide logical summaries depending on the context from which they originate. According to Delangue (2016), PEGASUS creates summaries by rewriting the original text in a way that keeps consistency while collecting the essential details. PEGASUS is a sequence-to-sequence model with a similar encoder-decoder architecture to BART. It is pre-trained using Masked Language Modeling (MLM) and Gap Sentence Generation (GSG), both of which use a causal mask to hide future words. MLM randomly replaces encoder input tokens, while GSG replaces entire sentences with a mask, like a regular auto-regressive transformer decoder.

Abstractive Combined Model

In the context of clinical text summarization, a combined model is a language model that combines several features of multiple pre-trained abstractive models to provide comprehensive summaries of clinical documents. The goal is to harness each model's distinct unique strengths and capabilities as an abstractive text summarization tool and to provide summaries that are more precise, thorough, and more accurate.

Insights into Model Selection

In this study we considered some rationales like architecture, performance, and adaptability for managing the complexity of clinical text which are crucial when choosing models for clinical text summarization. However, each model (BART, T5, and Pegasus) are effective clinical text summarization models, each with its own strengths and challenges.

- **BART model** is designed to handle a variety of NLP tasks and to combine the strengths of both bidirectional and autoregressive models. The architecture of this model requires significant computational power and memory. It is useful for summarizing a variety of noisy, unstructured text and messy clinical notes.

- **T5 model**'s text-to-text format enhances the accuracy of clinical reports, despite being extremely CPU-intensive. Its consistent methodology results in excellent quality summaries spanning

different tasks.

- **Pegasus model** is a pre-trained model that demonstrates its efficacy in managing the complicated and relevant structure of clinical text through the generation of clear and pertinent summaries. Pegasus, like BART and T5, requires significant computing capacity.

We selected the **combination of the 3 large language models** to summarize the clinical text because clinical data often contains complex medical terminology and detailed patient information.

- The combined summarization model will provide a summary that captures the underlying medical information & context more effectively.
- The combined abstractive model summarization synthesizes key points, reducing redundancy and emphasizing relevant information. This ensures clinical summaries are concise and focused on critical aspects of a patient's health and treatment, which is essential for efficient clinical decision-making.
- The combination of these three models allows us to expand the overall performance and quality of our summaries provided by the models.

5. EXPERIMENTAL ANALYSIS

Hardware and Environment Setting

The experiment and testing procedures presented in this paper were conducted using a Dell Inspiron 14 mounted with the Windows 11 operating system and the processor is Intel® Core i7-1255U CPU @1.70 GHz. The MIMIC-IV dataset was imported into the Python Google collab notebook, a platform optimized for swift Python coding. Given the dataset's substantial size, we opted for GPU runtime to enhance processing efficiency. For each large language model variant, essential libraries such as AutoTokenizer and pipeline dependencies from transformers were installed from the Hugging Face community package. These packages, offered by the Hugging Face community, provide tools for building, training, and deploying open-source machine learning models and ensuring accuracy and effectiveness throughout the summarizing procedure (see Appendix Item Figure 3). The model parameters used include "facebook/bart-base," "t5-base," and "google/pegasus-large" pre-trained models for BART, T5, and Pegasus, respectively. The necessary auto tokenizer and pipelines dependencies from transformers for BART model, BartForConditionalGeneration and BartTokenizer were utilized, while T5 model employed T5ForConditionalGeneration and T5 Tokenizer. Pegasus variants employed Pegasus Tokenizer and PegasusForConditionalGeneration. The selection of pre-trained models was informed by memory constraints in Google Collab.

Evaluation Metrics/ Performance

In this section, evaluation metrics employed in the text summarization experimentation are analyzed. These metrics serve to assess the quality and effectiveness of the generated summaries, leveraging a suite of widely recognized and accepted evaluation criteria of different LLMs.

BLEU Score (Bilingual evaluation understudy)

According to Papineni et al. (2002), BLEU is a metric for assessing the quality of text translated by a machine from one natural language to another. The algorithm uses n-grams found in human-translated sentences. It measures the similarity and the precision of the model's output compared to a reference summary.

The geometric mean of modified precision scores in a test corpus is calculated, multiplied by an exponential brevity penalty factor, and then used to compute the brevity penalty BP and weighted by the BP. The formular is shown below:

$$P_n = \frac{Count\ of\ n-grams\ in\ translation\ that\ appear\ in\ refernce}{Count\ of\ n-grams\ in\ generated\ translation} \quad (1)$$

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c \leq r \end{cases} \quad (2)$$

$$\text{BLEU Score} = BP \times \exp\left(\sum_{n=1}^{N} WnlogPn\right) \quad (3)$$

Where:

- c is the length of the generated translation that appears in reference
- r is the length of effective reference corpus length.
- Wn is the weight assigned to the precision of n-grams.
- Pn is the precision of n-grams.

The BLEU score ranges from 0 to 1. Higher BLEU scores indicate better overlap between the generated summary and the reference summary, indicating better quality translation. Lower BLEU scores imply less precision or accuracy in the model's output compared to the reference summary, indicating lower quality translation.

ROUGE Score (Recall-Oriented Understudy for Gisting Evaluation)

According to Lin (2004), ROUGE Score assesses the overlap of n-grams (sequences of words) between the generated summary and reference summaries. It considers metrics such as ROUGE-N (unigrams, bigrams, etc.) and ROUGE-L (longest common subsequence) to evaluate content overlap.

ROUGE-N refers to the overlap of n-grams between the system and reference summaries.

$$ROUGE - N = \frac{Total\ number\ of\ n-grams\ in\ reference\ summary}{Number\ of\ overlapping\ n-grams} \quad (4)$$

ROUGE-1 is the term used to describe how the framework and reference summaries overlap in terms of unigrams, or words.

$$ROUGE - 1 = \frac{Total\ number\ of\ unigrams\ in\ reference\ summary}{Number\ of\ overlapping\ unigrams} \quad (5)$$

ROUGE-2 refers to the overlap of bigrams between the system and reference summaries.

$$ROUGE - 2 = \frac{Total\ number\ of\ bigrams\ in\ reference\ summary}{Number\ of\ overlapping\ bigrams} \quad (6)$$

ROUGE-L refers to statistics based on the Longest Common Subsequence. To automatically identify the longest co-occurring in sequence of n-grams, the longest common subsequence issue takes sentence-level structural similarity into consideration.

$$ROUGE - L = \frac{Total\ number\ of\ words\ in\ reference\ summary}{Length\ of\ longest\ common\ subsequence} \quad (7)$$

The ROUGE metrics indicate that scores between 0 and 1. Higher ROUGE scores indicate better recall of important content from the reference while low ROUGE indicates that the generated summary may not accurately capture or recall important content from the reference summary.

BERT Score (Bidirectional Encoder Representations from Transformers)

A BERT Score measures the similarity between the model's representation of the summary and the reference using pre-trained contextual embeddings (T. Zhang et al., 2019). The formula computes the cosine similarity between the generated and reference phrases' contextual embeddings (Hanna & Bojar, 2021):

$$\text{BERT Score} = \sum_{i=1}^{L} \mathbf{CS}(\boldsymbol{xi}, \boldsymbol{y}) / \boldsymbol{L} \quad (8)$$

$$\text{BERT Score} = \sum_{i=1}^{N} \mathbf{F1}(\boldsymbol{CS}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{i})) / \boldsymbol{N} \quad (9)$$

Where:

- CS (xi, y) is the cosine similarity between a generated sentence x and its entire reference sentence y based on the contextual embeddings for the i-th token in each.
- L is the length of the generated sentence.
- N is the number of layers of BERT used for scoring.
- F1 is the harmonic mean function.

Higher BERT scores indicate a closer semantic match between the generated and reference summaries. Low BERT Scores suggest that the generated summary may not closely match the meaning or content of the reference summary.

6. DISCUSSION

The experiments focused on summarizing three test samples, each corresponding to a clinical note, from the MIMIC-IV-de-identified dataset. Performance evaluation includes metrics like BLEU, ROUGE, and BERT scores. The experimental results demonstrate the summarization capabilities of three large language models across diverse patient clinical notes and the evaluation metrics used in these experiments are examined to gauge the quality and efficacy of the generated summaries. Evaluation performance of the three LLMs for text summarization is detailed in Appendix Tables 1 - 3. The experimental results reveal distinct strengths among the three LLMs evaluated. The analysis showcases these models' impressive capacity to summarize complex medical reports into more concise forms.

In the first clinical note, the table performance scores show the comparison between the individual models and the combined model. The BART model demonstrates superior precision, achieving the highest BLEU score (0.012046), similar ROUGE-1 and ROUGE-L scores (0.304348), and the highest ROUGE-2 score (0.299270) across the 3 models. This suggests strong alignment between the generated and referenced summaries in terms of unigram overlap, longest common sequence, bigram overlap, and BERT score.

In the second and third clinical notes, the table performance scores show the individual model scores, the Pegasus model outperforms the other models in evaluation metrics including BLEU score, ROUGE score, and BERT scores. This indicates superior precision, recall, and F1 score in summary generation and suggests strong alignment between the generated and referenced summaries regarding unigram overlap, longest common sequence, and bigram overlap.

However, from the tables of evaluation performance, the combination of the 3 models shows significantly higher scores across all metrics in each clinical case sample indicating better overall performance and outperforming the individual models. The combined models indicate that combining the features from the three models will lead to better performance in generating a concise summary of clinical notes.

A comparison between the original clinical note and the summaries generated by the three LLMs and the combined model is shown in Appendix Tables 4 - 6, which echoes the evaluation performance reported in Appendix Tables 1 – 3. The summary generated by the combined model seems to contain more comprehensive and essential information from the original text source compared to three individual LLMs. The experimental findings and the generated text summaries underscore the remarkable proficiency of the three large language models when combined.

7. CONCLUSIONS AND FUTURE WORK

This study evaluates the performance of three widely used abstractive large language models (BART, T5, and Pegasus) and a combination of these models for clinical text summarization. The experimental results highlight the distinct strengths of each model, with the combined model emerging as the most effective approach for summarizing clinical notes. Limitations include a limited dataset. Furthermore, there are additional LLMs which need to be included in the evaluation. Additionally, in future work we aim to introduce the gold standard of human evaluation.

The combined model consistently outperforms the individual models across BLEU, ROUGE, and BERT metrics, demonstrating its superior ability to produce high-quality and robust summaries of clinical information. As detailed in Appendix Tables 1-3, integrating BART, T5, and Pegasus leverages the unique strengths of each model, resulting in a more comprehensive and accurate summarization.

The Implications of these experimental results are significant for clinical practice includes:

- **Reduced Cognitive Load**: Healthcare professionals can spend less time interacting with patients and providing urgent treatment when the cognitive strain of reading through extensive clinical notes is lessened. In demanding settings like critical care facilities and emergency departments, this is extremely advantageous.

- **Consistency and Accuracy:** The combined model ensures precise and consistent summaries by integrating many evaluation metrics. Such consistency is particularly important in clinical settings, where discrepancies or missing information in data can have serious implications.

- **Improved Efficiency:** With the combined model, healthcare professionals can quickly grasp essential information and review extensive medical records in less time and with greater efficiency. Precise summaries of medical text enable physicians to swiftly assimilate important details, leading to improved treatment of patients.

- **Improved Decision-Making Process:** To improve diagnostic and medical care decisions, the combined model ensures that medical practitioners have access to top-notch and high-quality summaries that highlight essential health information and clinical observations.

It is expected that the combined large language models for clinical text summarization have the potential to transform the delivery of medical services by improving patient experiences and the utilization of resources, especially in health care settings where accuracy is critical. The combination of different models in clinical text summarization provides for both present and potential scalability, making it a viable option for the health sector as medical data becomes substantially more complex and voluminous.

In addition to summarizing medical histories, the ability to extend this proposed approach of combining the 3 large language model for text summarization to other fields highlights its broader applicability and potential to transform practices across diverse fields such as (educational content summarization, technical documentation, news article summarization etc.)

The combination of these models (BART, T5, and Pegasus) into a broad language model ensures that clinical practitioners have access to relevant and comprehensive summaries, ultimately resulting to increase in productivity decisions, more effective and informed practices in the healthcare.

Further research efforts are essential to improve the combined summaries generated from various text summarization models. Assessing their effectiveness across a variety of datasets from various healthcare settings sheds light on how to enhance clinical note summarization and improve patient healthcare information outcome. Also, fine-tuning the model could further enhance its performance, as the synergy of its components often yields superior results compared to individual elements.

8. REFERENCES

Arokodare, O., & Wimmer, H. (2023). Large Language Models for Phishing and Spam Detection: A BERT Approach. *1st Annual Fall National Conference on Creativity, Innovation, and Technology Conference (NCCIT)*, USA.

Batra, P., Chaudhary, S., Bhatt, K., Varshney, S., & Verma, S. (2020). A Review: Abstractive Text Summarization Techniques using NLP. 2020 *International Conference on Advances in Computing, Communication & Materials (ICACCM)*, Dehradun, India, 2020, pp. 23-28, doi: 10.1109/ICACCM50413.2020.9213079.

Bhatia, N., & Jaiswal, A. (2015). Trends in extractive and abstractive techniques in text summarization. *International Journal of Computer Applications*, 117(6), 21-24. DOI:10.5120/20559-2947.

Bijal, D., Nikita, P., & Sanket, S. (2017). A review paper on text summarization for Indian languages. *IJSRD-International Journal for Scientific Research & Development*, 5(7), 744-745.

Delangue, C. (2016). Hugging face. https://huggingface.co/docs/transformers/model_doc/pegasus.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805.

Gaikwad, D. K., & Mahender, C. N. (2016). A review paper on text summarization. *International Journal of Advanced Research in Computer and Communication Engineering*, 5(3), 154-160. DOI:10.17148/IJARCCE.2016.5340.

Hanna, M., & Bojar, O. (2021). A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

HealthIT.gov. (2021). Challenges to Electronic Public Health Reporting among Non-Federal Acute Care Hospitals During the COVID-19 Pandemic. https://www.healthit.gov/data/data-briefs/electronic-public-health-reporting-among-non-federal-acute-care-hospitals-during.

Johnson, A., Pollard, T., Horng, S., Celi, L., & Mark, R. (2023). MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2). PhysioNet. https://doi.org/10.13026/1n74-ne17.

Johnson, A. E., Stone, D. J., Celi, L. A., & Pollard, T. J. (2018). The MIMIC Code Repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1), 32-39. DOI: 10.1093/jamia/ocx084.

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., & Socher, R. (2019). *Neural text summarization: A critical evaluation*. arXiv preprint arXiv:1908.08960. https://doi.org/10.48550/arXiv.1908.08960.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages* 311-318. https://doi.org/10.3115/1073083.1073135.

Roberts, A., Raffel, C., Lee, K., Matena, M., Shazeer, N., Liu, P. J., Narang, S., Li, W., & Zhou, Y. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. Google, Tech. Rep. https://doi.org/10.48550/arXiv.1910.10683.

Tsai, H.-Y., Huang, H.-H., Chang, C.-J., Tsai, J.-S., & Chen, H.-H. (2022). Patient History Summarization on Outpatient Conversation. 2022 IEEE/WIC/ACM *International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Niagara Falls, ON, Canada, 2022, pp. 364-370, doi: 10.1109/WI-IAT55865.2022.00060.

ISSN: 2473-4901
v10 n6122

Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E. P., & Seehofnerova, A. (2023). Clinical text summarization: Adapting large language models can outperform human experts. *Research Square*. doi: 10.21203/rs.3.rs-3483777/v1.

Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Costa, A. B., & Flores, M. G. (2022). A large language model for electronic health records. *NPJ Digital Medicine*, 5(1), 194. https://doi.org/10.1038/s41746-022-00742-2.

Zhang, H., Xu, J., & Wang, J. (2019). Pretraining-based natural language generation for text summarization. arXiv preprint arXiv:1902.09243. https://doi.org/10.48550/arXiv.1902.09243.

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., & Artzi, Y. (2019). *BERTScore: Evaluating Text Generation with BERT*. ArXiv, abs/1904.09675. https://doi.org/10.48550/arXiv.1904.09675.

ISSN: 2473-4901
v10 n6122

APPENDIX

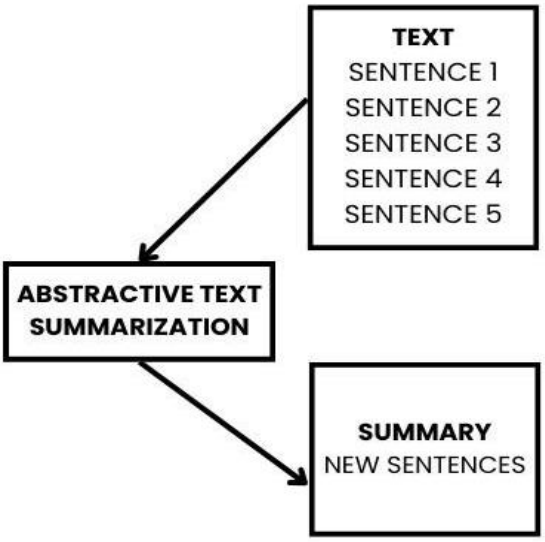

Figure 1 Abstractive Text Summarization Workflow

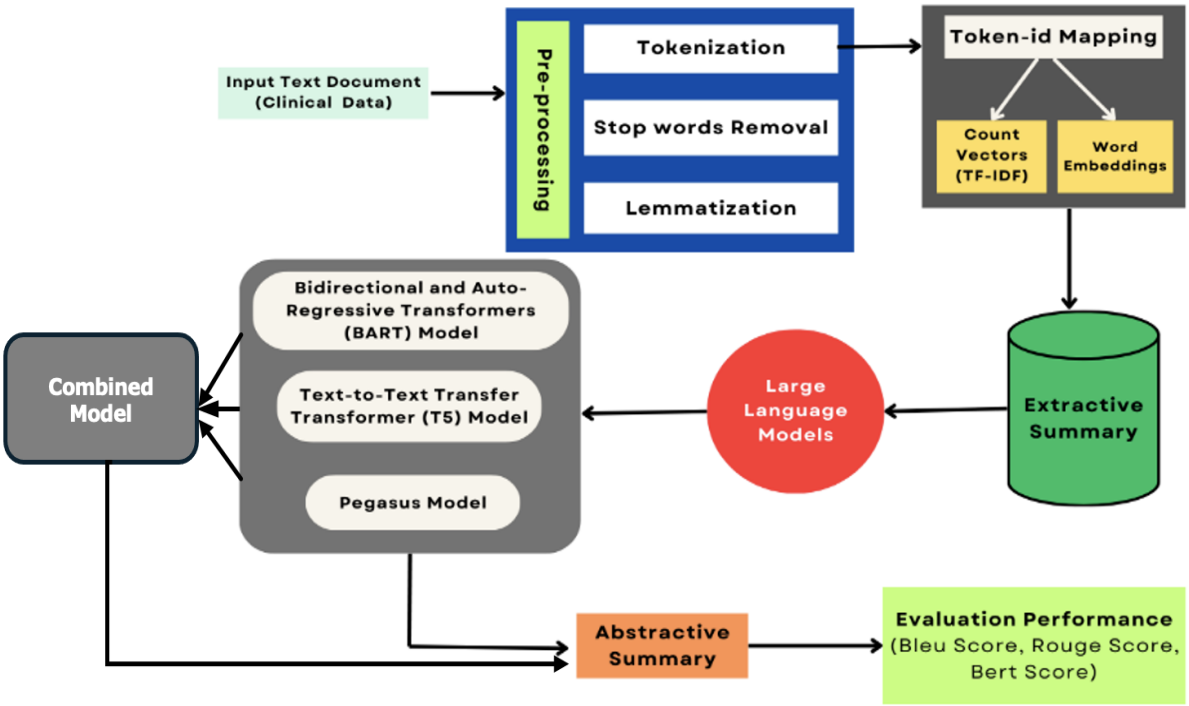

Figure 2 Architecture Diagram for Clinical Text Summarization with Large Language Model

ISSN: 2473-4901
v10 n6122

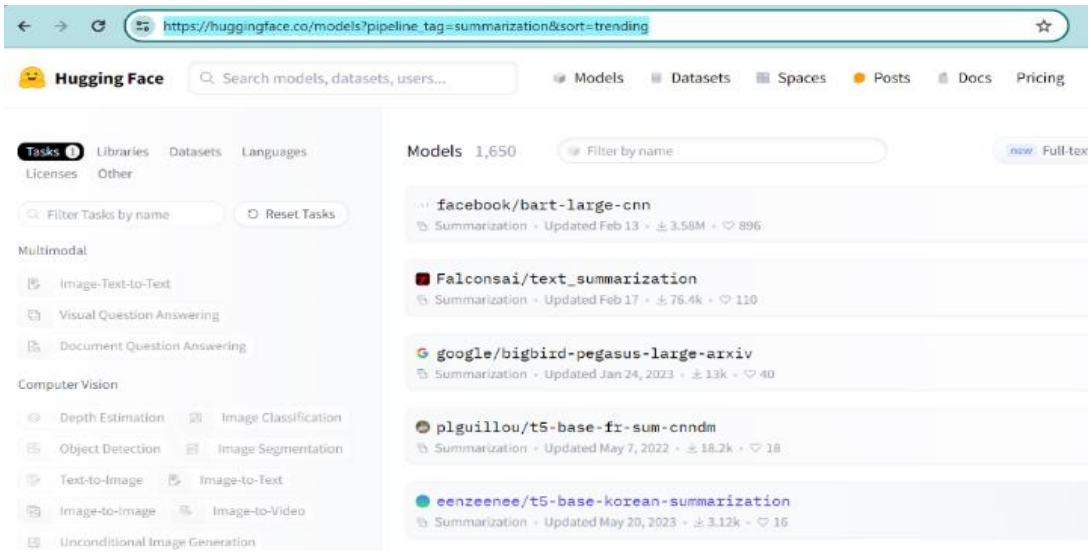

Figure 3 Hugging Face Model Parameters Snapshot for Text Summarization

		Clinical Note 1				
		Scores				
		BLEU	Rouge1	Rouge2	RougeL	BERT(Precision/Recall/F1)
Model	T5	0.002011	0.247191	0.241509	0.247191	0.926807 ,0.813933, 0.866710
	Bart	0.012046	0.304348	0.299270	0.304348	0.918589,0.850790, 0.883391
	Pegasus	0.000545	0.193050	0.186770	0.193050	0.949601,0.812216,0.875552
	Combined Model	0.197573	0.473054	0.445783	0.375449	0.963512,0.838817, 0.896851

Table 1: Clinical Note Test Sample 1

		Clinical Note 2				
		Scores				
		BLEU	Rouge1	Rouge2	RougeL	BERT(Precision/Recall/F1)
Model	T5	0.011963	0.324706	0.307329	0.320000	0.947187, 0.821805, 0.880052
	Bart	0.024612	0.366133	0.358621	0.361556	0.946826,0.826174, 0.882395
	Pegasus	0.062197	0.421286	0.405345	0.421286	0.956568,0.835085, 0.881708
	Combined Model	0.362768	0.575707	0.467446	0.465890	0.962171, 0.846523, 0.892728

Table 2: Clinical Note Test Sample 2

ISSN: 2473-4901
v10 n6122

		Clinical Note 3				
		Scores				
		BLEU	Rouge1	Rouge2	RougeL	BERT(Precision/Recall/F1)
Model	T5	0.000002	0.131261	0.117851	0.131261	0.920835,0.800128,0.856248
	Bart	0.000007	0.147260	0.140893	0.113014	0.932428,0.803353,0.863091
	Pegasus	0.000506	0.202322	0.189684	0.202322	0.909739,0.817642,0.861235
	Combined Model	0.058681	0.415205	0.384164	0.339181	0.941585,0.863127,0.900651

Table 3: Clinical Note Test Sample 3

Actual Medical Health History- Case 1	Generated Summary by the Model
"HCV cirrhosis c/b ascites, hiv on ART, h/o IVDU, COPD, bioplar, PTSD, presented from OSH ED with worsening abd distension over past week. Pt reports self-discontinuing lasix and spirnolactone weeks ago, because she feels like they dont do anything and that she doesnt want to put more chemicals in her. She does not follow Na-restricted diets. In the past week, she notes that she has been having worsening abd distension and discomfort. She denies edema, or SOB, or orthopnea. She denies f/c/n/v, d/c, dysuria. She had food poisoning a week ago from eating stale cake n/v 20 min after food ingestion, which resolved the same day. She denies other recent illness or sick contacts. She notes that she has been noticing gum bleeding while brushing her teeth in recent weeks. she denies easy bruising, melena, BRBPR, hemetesis, hemoptysis, or hematuria.Because of her abd pain, she went to OSH ED and was transferred to for further care. Per ED report, pt has brief period of confusion - she did not recall the ultrasound or bloodwork at osh. She denies recent drug use or alcohol use. She denies feeling confused, but reports that she is forgetful at times. In the ED, initial vitals were 98.4 70 106/63 16 97%RA Labs notable for ALT/AST/AP , Tbili1.6, WBC 5K, platelet 77, INR 1.6"	**BART Model** "HCV cirrhosis c/b ascites, hiv on ART, h/o IVDU, COPD, bioplar, PTSD, presented from OSH ED with worsening abd distension over past week. Pt reports self-discontinuing lasix and spirnolactone weeks ago, because she feels like they dont do anything." **T5 Model** "in the past week, she has been having worsening abd distension and discomfort. she denies edema, or SOB, or orthopnea. she had food poisoning a week ago from eating stale cake n/v." **Pegasus Model** "HCV cirrhosis c/b ascites, hiv on ART, h/o IVDU, COPD, bioplar, PTSD, presented from OSH ED with worsening abd distension over past week" **Combined Model (BART, T5 and Pegasus Model)** "Pt with HCV cirrhosis, HIV on ART, COPD, bipolar disorder, and PTSD presented with worsening abdominal distension, discontinue taking lasix and spironolactone because she feels like they dont do anything. in the past week, she has worsened abdominal discomfort, denies edema, SOB, or orthopnea. she had food poisoning from eating stale cake."

Table 4: Comparison between Clinical Text Sample 1 and Summaries Generated by the Four Models

ISSN: 2473-4901
v10 n6122

Actual Medical Health History- Case 2	Generated Summary by the Model
"This is a woman with a history of ETOH abuse who presents with abdominal distention, back pain, fever, and elevated white count from Liver Clinic. Ms. was recently admitted to this hospital about 1 week ago for treatment of ascites and work-up of alcoholic hepatitis. At that time she had a diagnostic and therapeutic paracentesis and was treated for a UTI. She wasdischarged home and instructed to follow-up in Liver Clinic in 1 week. On day of presentation to liver clinic, patient complained of worsening abdominal pain and low-grade fevers at home. Her labwork was also significant for an elevated white count. As such, Ms. was admitted for work-up of fever and white count, and for treatment of recurrent ascites. with recently diagnosed alcoholic hepatitis, persistent ascites, and persistent fevers and leukocytosis which have been atributed to her hepatitis who presented to today with worsening abdominal distention, pain, and persistent fever. She denies chills but did have sweats the night prior to admission. She has tried to be strictly compliant with her low socium diet and fluid restriction, and denies any increased fluid or sodium intake. She reports sobriety from alcohol since . At she was febrile and tender to palpation, so she was referred to the ED.In the ED initial vital signs were 99.0 113/72 132 16 99% on RA. Her temp increased to 100.4 and her pulse came down to the 100s with Ativan. She received morphine 4mg IV x 4 for pain, tylenol PO x1 for fever, ondansetron 4mg IV x2 for nausea, and lorazeman 0.5mg IV x1 for anxiety. She underwent a diagnostic paracentesis but the samples were initially lost. She was treated with ceftriaxone 2g IV x1 for possible SBP. She was admitted to Medicine for further management. Fortunately, her samples were found after she arrived on the floor. On the floor her mood is labile. She is at times tearful and at times pleasant. She does seem uncomfortable. She is not confused or obviously encephalopathic. She denies cough, dysuria, diarrhea, or rash. She does endorse decreased UOP for the past few days."	**BART Model** "Ms. was recently admitted to this hospital about 1 week ago for treatment of ascites and work-up of alcoholic hepatitis. At that time she had a diagnostic and therapeutic paracentesis and was treated for a UTI. She was discharged home and instructed to follow-up in Liver Clinic in 1 week. On day of presentation to liver clinic, patient complained of worsening abdominal pain and low-grade fevers at home. Her labwork was also significant for an elevated white count." **T5 Model** "a woman with a history of ETOH abuse presents with abdominal distention, back pain, fever, and elevated white count. she was recently admitted to this hospital about 1 week ago for treatment of ascites and work-up of alcoholic hepatitis. on the day of presentation to liver clinic, patient complained of worsening abdominal pain and low-grade fevers at home. her labwork was also significant for an elevated white count." **Pegasus Model** "This is a woman with a history of ETOH abuse who presents with abdominal distention, back pain, fever, and elevated white count from Liver Clinic. was recently admitted to this hospital about 1 week ago for treatment of ascites and work-up of alcoholic hepatitis. On day of presentation to liver clinic, patient complained of worsening abdominal pain and low-grade fevers at home. with recently diagnosed alcoholic hepatitis, persistent ascites, and persistent fevers and leukocytosis which have been atributed to her hepatitis who presented to today with worsening abdominal distention, pain, and persistent fever." **Combined Model (BART, T5 and Pegasus Model)** "a woman with a history of ETOH abuse presents with abdominal distention, back pain, fever, and elevated white count. She was recently admitted to this hospital about 1 week ago for ascites treatment and alcoholic hepatitis work-up. On presentation day at the liver clinic, she reported worsening abdominal pain and low-grade fevers at home, alongside elevated white count in lab work results".

Table 5: Comparison between Clinical Text Sample 2 and Summaries Generated by the Four Models

Actual Medical Health History- Case 3	Generated Summary by the Model
"yo female with history of Afib on Xarelto, COPD, HTN, PAD who presents for abnormal labs. She noted dark, tarry, stool on and presented to PCP , where H/H was noted to be 8.8/28.6 from prior 11..6 baseline Hct about 38. She has also been experiencing bright red blood with wiping, she believes from her hemorrhoids. PCP called pt who agreed to come to ED. She had colonoscopy in with showed a benign polyp, internal hemorrhoids, and diverticulosis. Her last BM was , was reportedly regular. She currently complains of increased exertional fatigue and has been feeling more SOB than her baseline. Over the last 6 months she has noticed she becomes increasingly out of breath, walking or climbing stairs. She becomes SOB after 6 stairs or less than 1 block, requiring her to stop, and at times use albuterol inhaler. She used to use her only use her inhaler times per day, now she uses it over four times a day and nebulizers twice a day. F with pmhx of COPD nighttime O2, htn, afib who presents with dyspnea, currently being treated for COPD and admitted for Afib with RVR. The patient went to the ED on and was diagnosed with a COPD flare. She was discharged with a prednisone taper currently on 60mg and azithromycin. This AM she initially felt well, then developed dyspnea at rest, worsening with exertion. Her inhalers improved her SOB. She felt that these symptoms were consistent with her COPD. She saw her PCP today in clinic where she was found to be in Afib w/ RVR, rate around 110-120. She has a history of afib. He referred her to the ED for persistent SOB and afib with RVR. She states she been compliant with nebs and steroid/azithro regimen. She denies any edema, orthopnea. She denies recent travel, surgeries. She had an episode of chest tightness this AM that felt like her COPD flares. Denies fevers or coughing or production of sputum, hemomptysis. year old female with history of COPD on home O2, HTN, Afib admitted with dyspnea and cough. Pt states inc dyspnea since this am, also one episode of retrosternal chest pressure lasting 2minuts on way to ED. No cp currently. on home O2. no fevers/chills or abd sx. Patient was recently admitted from with COPD flare and afib with RVR. She could not receive azithromycin due to concern for QTc prolongation and so was treated with ceftriaxone/cefpodoxime. She was treated with 60mg PO prednisone and discharged with a prednisone taper of 10 mg decrease q3d until at 10 mg, then stay at 10 mg until pulm follow up. She was also counseled to do pulmonary rehab and follow up with Dr. . She was discharged on 2L supplemental O2 to be worn at all times. He theophylline was decreased from 300 mg BID to mg BID due to her afib with RVR.She was also seen in the ED on and due to dyspnea which was felt to be a continuation of her COPD flare in the setting of patient not taking her home medications. She was given nebulizers and improved. She was DCed home with for assistance with medications. She declined pulmonary rehab facility disposition."	**BART Model** "The patient went to the ED on and was diagnosed with a COPD flare. She had colonoscopy in with showed a benign polyp, internal hemorrhoids, and diverticulosis. She currently complains of increased exertional fatigue and has been feeling more SOB than her baseline." **T5 Model** "yo female with history of afib on Xarelto, COPD, HTN, PAD. she has been experiencing bright red blood with wiping, believes from her hemorrhoids. over the last 6 months she has noticed she becomes increasingly out of breath." **Pegasus Model** "She became SOB after 6 stairs or less than 1 block, requiring her to stop, and at times use albuterol inhaler. F with pmhx of COPD nighttime O2, htn, afib who presents with dyspnea, currently being treated for COPD and admitted for Afib with RVR. year old female with history of COPD on home O2, HTN, Afib admitted with dyspnea and cough." **Combined Model (BART, T5 and Pegasus Model)** "yo female with history of afib on Xarelto, COPD, HTN, PAD. she has been experiencing bright red blood with wiping, believes from her hemorrhoids. over the last 6 months she has noticed she becomes increasingly out of breath.The patient went to the ED on and was diagnosed with a COPD flare. She had colonoscopy in with showed a benign polyp, internal hemorrhoids, and diverticulosis. She currently complains of increased exertional fatigue and has been feeling more SOB than her baseline.She became SOB after 6 stairs or less than 1 block, requiring her to stop, and at times use albuterol inhaler. F with pmhx of COPD nighttime O2, htn, afib who presents with dyspnea, currently being treated for COPD and admitted for Afib with RVR. year old female with history of COPD on home O2, HTN, Afib admitted with dyspnea and cough."

Table 6: Comparison between Clinical Text Sample 3 and Summaries Generated by the Four Models