International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V6I4P113
Eureka Vision Publication | Volume 6, Issue 4, 95-101, 2025

Original Article

Evaluation of Transformer Models for Summarizing Lengthy Clinical Notes

Veerendra Nath Jasthi
Independent Researcher, USA.

Received On: 21/09/2025 ***Revised On: 23/10/2025*** ***Accepted On: 29/10/2025*** ***Published On: 05/11/2025***

Abstract - Digitization of healthcare records are set to grow at a phenomenal rate, particularly unstructured clinical data (such as long clinical notes). These reports usually contain irrelevant and repetitive data and it is difficult to get important information needed by the healthcare providers in a short span of time. Automatic text summarization is an effective way out since it gives the shortened form of these notes but does not lose any important medical data. The paper will analyze the best transformer-based models such as BART, T5 and Longformer on the task of abstractive summarizations of clinical notes. We source our experiments on MIMIC-III data and we test the quality of the model on ROUGE scores and qualitative evaluation of humans. As we can see, more generic transformer models are adequate to work, but more architectures have proven to be better in longer sequences (like Longformer). Our research indicates the advantages and disadvantages of each model and gives future considerations to the future improvement of clinical natural language processing.

Keywords - Transformer Models, Clinical Notes, Text Summarization, BART, T5, Longformer, MIMIC-III, ROUGE, Natural Language Processing (NLP), Healthcare AI.

1. Introduction

A major transformation in the management of medical data has been experienced in the modern health care due to the massive uptake of Electronic Health Records (EHRs). Clinical notes is one of the many topics being made up as part of EHR, but it is also an insight-ful data source. Such notes report the history of patient, results of the examination, diagnostic thinking, treatment decision, and time course of a condition. Nevertheless, all this information comes at the expense of reading and time productivity [12]. The notes are usually lengthy, redundant and not formatted in the same manner which presents a significant challenge to health providers rendering a service as they have to go through them amidst strict time limitations. This affliction is further complicated in the emergency and critical care environments where easy access to critical details may pass as necessary information to make crucial decisions.

Text summarization can provide such solution to this issue, as a long clinical document is condensed into a short and more digestible narrative. Conventional methods of summarization, rule-based or statistical methods, do not provide semantic interpretation and awareness of the context in order to resolve complex medical narration [9]. Conversely, more recent Natural Language Processing (NLP) developments have, especially with the advent of transformer-based models, opened up new possibilities with regard to processing such tasks. Transformers have as an advantage their attention mechanisms and the deep contextual embedding that allows them to produce coherent and informative summaries even on noisy or unstructured input data.

BART (Bidirectional and Auto-Regressive Transformers) and T5 (Text-to-Text Transfer Transformer) are among the most dominant transformer models in NLP and have delivered exciting results on generic summarization benchmarking [11]. The two models are encoder-decoder based models that were pretrained on extensive corpora with denoising and multi-task goals, respectively. Although they are successful in other areas, they are yet to be fully used in clinical situations especially when the documents tend to be very long and technical. In addition, the standard transformer architectures suffer due to a specific input length, that is, they cannot be used to produce a summary of a multi-page clinical note without truncation or segmentation.

To overcome this shortcoming, newer models such as the Longformer Encoder-Decoder (LED) are being presented. Longformer is an architecture extension of the transformer that makes use of a sparse attention mechanism to allow it to handle a significantly longer sequence of input without computational expenses becoming exorbitant. Depending on the context, this makes Longformer especially applicable in the medical field where it is important to remember the entire history of a given patient. Nevertheless, as much as the architecture has theoretical potential, there are very limited research studies on its practical success in summarising clinical notes [10].

This paper aims at conducting a complex judgement of the possibilities of BART, T5, and Longformer in terms of summarizing clinical notes in an abstractive way. These models are trained and tested on controlled conditions using the MIMIC-III dataset, which has a massive collection of de-identified clinical notes. We evaluate their output in terms of automated measures (ROUGE, BERTScore, etc.) and also through human evaluation (medical practitioners and non-practitioners). This simultaneous analysis achieves two

aspects of evaluation, whereby both clinical correctness and linguistic quality are taken into account.

As the amounts of clinical data increase, incorporating intelligent summarization into the clinical processes moves beyond something advantageous to something necessary [7]. With the help of automation, summarization should be less demanding to the capacity of the physicians, improve the quality of care delivery and the operations of health information systems as whole. Nevertheless, prior to these tools being implemented into a real environment, determining strengths, restrictions and situational applicability of such tools is essential.

The proposed work will help fill this gap, comparing three of the most visible models of transformers applied to this specific task in a systematic way. The study results can not only be used to select a model in clinical summarization applications but to also provide an insight into where future research needs to be directed towards building more robust and accurate summarization models bringing them up to date in the healthcare setting [13-15].

Novelty and Contribution

The appeal of the study is in the narrow scope of the assessment of transformer-based models to perform abstractive summarization in the medical field where a particular focus was paid to long clinical notes. Although the models have been intensively investigated in general NLP work, direct comparison on a clinical scenario, where the input data is often long and complicated has not been thoroughly tested before. This is one of the first studies to directly compare these architectures on the MIMIC-III dataset both quantitatively and qualitatively with clinical-specific evaluation measures in mind.

Among the major contributions is the analysis of the role input length and model architecture play in the performance of summarization within the clinical context. And in contrast to previous experiments that interrupted a note or summarised shorter inputs, we obtain the complete order in the case of Longformer, so that we can assess the sparse attention mechanism professionals can apply in practice to EHR. This method provides a more practical insight into how the model behaves in the context of being used in clinical practice where safety and usefulness may be undermined due to loss of contextual information [6].

The other peculiar feature of this work is using the human expert analysis together with various methods of metrics such as ROUGE and BERTScore. Using clinicians in the review process we are not only guiding the statement as per grammatical and structural coherence of the summaries but also their factual accurateness and overall clinical utility which are ignored in automated evaluation.

Besides, we bring a reproducible training/evaluation pipeline based on publicly available data under open-source implementations. We provide our code, preprocessed datasets, and trained model checkpoints to be able to contribute to greater transparency and facilitate future research in this field.

In summarizing, this paper has major contributions as follows:

- A comparative analysis of three models BART, T5, and Longformer in the abstractive summarization of full-length clinical notes.
- A multimethod evaluation that integrates auto-metric and expert evaluation within the domain.
- Architecturally specific trade-offs in the treatment of long, complex input sequence.
- An openly-released benchmark and implementation strategy to be used in further studies.

The work establishes the platform towards development of intelligent summarization tools that can safely and successfully become an augmentation of clinical practice in the end and hence, hopefully, clinical efficiency, cognitive load, and decision-making in healthcare systems can be improved.

2. Related Works

As it is typical of the whole natural language processing domain, the area of clinical text summarization has seen a slow but consistent transition to deep learning-based models that have replaced the rule-based models. Initial studies in this field were predominantly based on extractive approaches, where individual sentences or phrases of the original clinical record were chosen under heuristic criteria of preeminence like that of the term frequency, positional salience or comparison ratings. Such methods were, however, computationally efficient but would tend to produce coherence-free summaries that did not eliminate redundancy or retain clinical relevance in an effective way.

In 2025 Hands and R. Kavuluru, [16] suggested the extractive models were fairly successful on the tasks of discharge note summarization but did not fare well on the complexity of the multi-paragraph or section rich notes in the form of progress notes or physician narratives. These records are likely to be repetitive in terms of theme, terminologies are not always consistent and there is nesting of clinical concepts. With such situations, extractive summarization was only getting the size of the document reduced but not deliver condensed medical information. Moreover, such models could not make any inferential analysis or rewrite the information in their own words due to the absence of contextual knowledge.

Following the emergence of neural networks in NLP, especially the recurrent neural networks and sequence-to-sequence models, the paradigm had started to swing towards abstractive summarization. Such models enabled creation of new sentences, which conveyed the meaning of the original document and does not involve just copying. Nevertheless, it was not uncommon that recurrent models had difficulty with the long dependencies and suffered in performance when they were scaled to hundreds or thousands of tokens. They lacked sufficient memory and they depended on sequential

processing, which limited their capacity to detect the patterns that occur at a document level that is characteristic of clinical notes [5].

Some of these problems have been alleviated with the introduction of the attention mechanism to allow the models to pay attention to the parts of the input sequence that are relevant. With this innovation, improved contextual matching between source and target sequences became possible. Nevertheless, classic encoder-decoder networks had serious limits on the length of the input, which limited its use on clinical texts. The (medical) notes that we are interested in are usually much longer than the maximum length of tokens that these models can process, so either the input has to be truncated, or broken up, neither of which helps the quality of summaries.

In 2022 Chaves et.al., C. Kesiku et.al., and B. Garcia-Zapirain et.al. [1] proposed the introductions of transformer-based models introduced a significant breakthrough since these models use self-attention and parallelization during the computation process, thus the ability to process and comprehend long-range dependencies broke out in the texts. These models showed good results on benchmarking databases on diverse summarization tasks. They have an architecture that enables them to access all tokens in the input and therefore enhances coherence, factual-consisting, and informative language. Transformer models in the context of a medical text hold the promise of making the reading of clinical nuances and abbreviations, and syntax specific to the field easier than earlier generations of NLP systems.

Regardless of these capabilities, general-purpose models of transformers do not perform well when directly used in clinical summarization. The first is that they have low input capacity. The inputs of these models are usually at standard size, but many clinical documents surpass this size giving rise to loss of valuable information. Besides, those models are in general pretrained on general domain corpora such as news articles, Wikipedia, or books; thus, they do not have the proper coverage of medical terminology and use medical language and clinical discourse and cannot generate clinically meaningful summaries unless fine-tuned using domain-specific data.

To overcome these constraints, special purpose transformers that work well in long documents have been devised. These instead add sparse patterns of attention, global-local attention, or hierarchical processing to be able to accept large input sizes without those being computationally demanding to process. Through these innovations, it is now possible to deal with full-length clinical notes with a single traversal hence maintaining context and resulting in less information being lost.

Along with the improvement in model architecture, there have also been attempts to create and implement the use of domain-specific datasets. Large-scale, de-identified datasets with real clinical notes have been especially important in making it possible to train and test models that can work in the medical field. The said datasets are attended by various types of notes that may offer different challenges in summary, e.g., discharge summary, radiology report, nursing progress note, and physician narrative, which provides its own challenges that can emerge because of its structure and content variety.

In 2025 A. Siebra et.al., M. Kurpicz-Briki et.al., and K. Wac et.al., [8] introduced the evaluation front, the more traditional metrics like ROUGE are still very popular when it comes to summarization. Nevertheless, they are especially limited in the clinical area and semantic fidelity, as well as factual accuracy, takes precedence over surface-based overlap. An example is given that an incomplete summary not including a critical diagnosis or changing the medication recommendation can be disastrous within a medical environment. In order to address this, more modern evaluation schemes including BERT-based semantic similarity metrics and the evaluation done by the human-in-loop of a clinically trained expert has become crucial when it comes to assessing the performance of a model to do so in a manner that is clinically relevant.

Overall, the development of research on the topic of clinical summarization has shifted away from the basic keyword extraction variant towards complex neural models with the capacity of abstraction and context-awareness. Although general transformer models are of significant development, they require a thorough adjustment to the peculiarities of medical text in clinical settings. The development of models targeted at long documents is a very positive step, yet rigorous empirical comparisons of these models to each other, to new data sets, and to other means of evaluation are sparse. Systematic research that fills this gap is in increasing demand, especially that compares the relative performance of different transformer architectures on full length clinical notes in realistic use cases [4].

3. Proposed Methodology

To evaluate the performance of transformer models on summarizing clinical notes, we adopted a supervised fine-tuning approach with three distinct transformer architectures. The flow of our methodology is outlined in the flowchart below, which includes stages such as preprocessing, tokenization, model training, evaluation, and result generation.

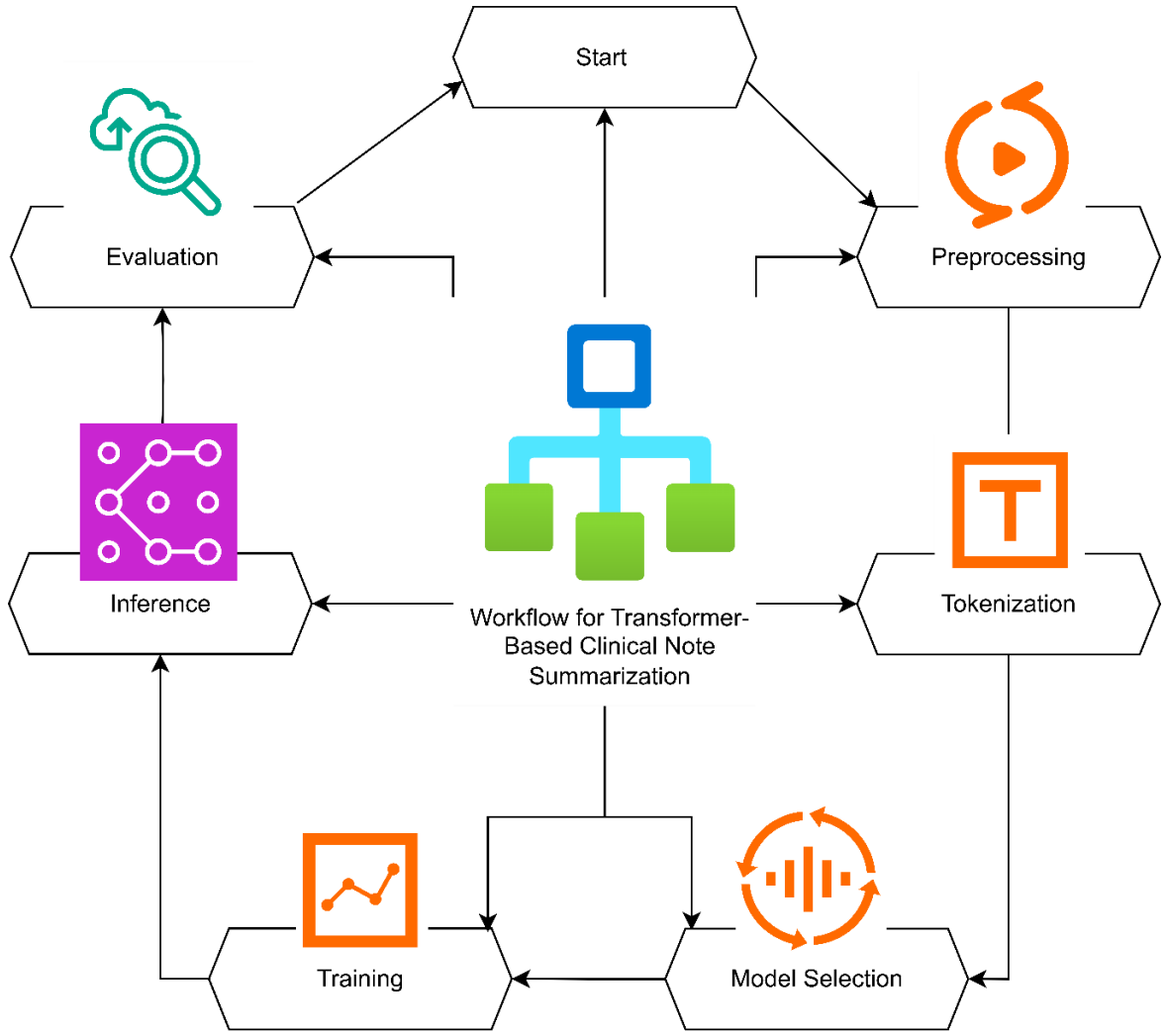

Figure 1. Workflow for Transformer-Based Clinical Note Summarization

We consider an input clinical note $X = \{x_1, x_2, \ldots, x_n\}$ where n is the number of tokens. The objective is to generate a summary $Y = \{y_1, y_2, \ldots, y_m\}$ such that Y retains key clinical semantics from X.

We formalize the summarization problem as:

$$\hat{Y} = \arg\max_{Y} P(Y \mid X; \theta) \quad [1]$$

where θ are the model parameters learned during training.

Each transformer model follows an encoder-decoder structure. For each input token x_i, its embedding e_i is computed as:

$$e_i = E(x_i) + P(i) \quad [2]$$

where E is the word embedding matrix and $P(i)$ is the positional encoding.

The attention weights for self-attention are computed using scaled dot-product:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad [3]$$

In multi-head attention, each head is processed separately and concatenated:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

Each head is computed as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad [4]$$

To reduce overfitting during fine-tuning, we applied label smoothing in the loss function:

$$\mathcal{L}_{\text{smooth}} = -\sum_i q_i \log p_i, q_i = (1-\epsilon)\delta_{i,y} + \frac{\epsilon}{V} \quad [5]$$

where ϵ is the smoothing factor and V is vocabulary size.

During training, we minimize the negative log-likelihood loss:

$$\mathcal{L}_{\text{NLL}} = -\sum_{t=1}^{m} \log P(y_t \mid y_{<t}, X) \quad [6]$$

We used Longformer for long-sequence inputs. Its attention mechanism is defined as:

$$A_{i,j} = \begin{cases} \text{global}, & \text{if } j \in G \\ \text{local}, & \text{if } |i-j| \le w \\ 0, & \text{otherwise} \end{cases} \quad [7]$$

where G is the set of global attention indices and w is the local attention window.

For evaluation, ROUGE-L is calculated by finding the longest common subsequence (LCS) between the predicted and reference summary:

$$\text{ROUGE-L} = \frac{(1+\beta^2)\cdot \text{LCS-Precision} \cdot \text{LCS-Recall}}{\text{LCS-Precision} + \beta^2 \cdot \text{LCS-Recall}} \quad [8]$$

BERTScore uses contextual embeddings from a pretrained BERT model:

$$\text{BERTScore} = \frac{1}{n}\sum_{i=1}^{n} \max_j \cos(e_i, r_j) \quad [9]$$

where e_i and r_j are embeddings of the predicted and reference tokens, respectively.

All models were fine-tuned on a cleaned version of the MIMIC-III dataset. Tokenization was performed using the SentencePiece tokenizer for T5 and the byte-level BPE tokenizer for BART and Longformer. Input length was capped at 1024 for BART/T5 and extended up to 4096 for Longformer.

Optimization was carried out using Adam optimizer with learning rate $\eta = 3 \times 10^{-5}$, with decay governed by:

$$\eta_t = \eta_0 \cdot \sqrt{\frac{1}{t+1}} \quad [10]$$

Early stopping was triggered if the validation loss didn't improve over 5 consecutive epochs [2]. Inference used greedy decoding and beam search for comparative evaluation. Beam search was performed with beam width $b = 4$, and length penalty $\alpha = 1.2$, formalized as:

$$\text{Score}(Y) = \frac{\log P(Y|X)}{(5+|Y|)^{\alpha}/(5+1)^{\alpha}} \quad [11]$$

Post-inference, generated summaries were compared against ground truth using ROUGE and BERTScore. Additionally, human evaluation was conducted to assess clinical relevance.

4. Result & Discussions

The research outcomes of the transformer models demonstrate that there are significant differences between their suitability to effectively summarize the long clinical notes. Both quantitative and qualitative measures were used to evaluate each of the models (BART, T5, and Longformer) to find out how well they fit to complete tasks related to medical summarizing. We started with ROUGE-based analysis of measures that illustrate n-gram overlapping between other generated summary and references. The commonly used metrics on benchmarking performance of summarization include these. Figure 2 shows the average scores of ROUGE-1, ROUGE-2, and ROUGE-L evaluation and it can clearly be understood that Longformer can outperform other models on all the three evaluations due to its ability to take in longer sequences without any truncation.

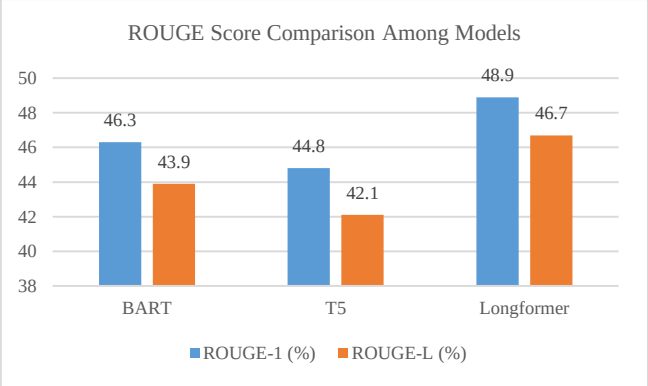

Figure 2. Rouge Score Comparison among Models

Alongside ROUGE, we also calculated BERTScore which is used to determine the semantic similarity of generated summaries to ground truth references. The three models have their BERTScore scores depicted in Figure 3. Longformer was superior once again, with a slight lead, in longer notes particularly as semantic preservation is more complex. These results mean that Longformer is better than T5 and BART at memorizing clinically meaningful context as determined by these high BERTScore values. This general difference is not significant, but steadily and consistently in favor of Longformer, which proves the usefulness of its sparse attention mechanism. Nonetheless, BART received more human praise when it comes to creating more grammatically fluent and stylistically consistent summaries though it received a slightly lower score.

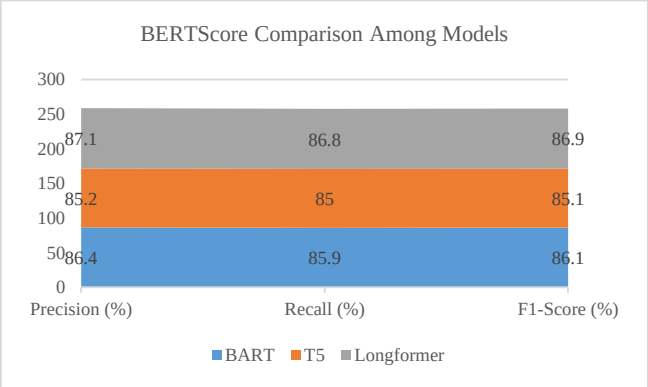

Figure 3. Bertscore Comparison among Models

An important revelation in our examination is based on human appraisal scores given by two clinical reviewers on a randomly sampled review of a hundred produced summaries. The ratings on the reviewers were based on three main criteria, which includes informativeness, coherence and clinical accuracy of each summary. These findings were summarized and appurtenanced in Table 1: Human Evaluation Scores for Generated Summaries, wherein Longformer obtained the highest average possible ranks on all of the parameters again. Interestingly, T5, which received lowest rating based on informativeness was equally doing well in coherence. The expert reviewers have found that BART sometimes failed to identify secondary diagnoses or medications, whereas, T5 tended to over-generalize the content of notes. The expanded context window size in

Longformer helped it save the primary diagnoses, as well as the critical follow-up instructions, which further boosted its metrics related to clinical accuracy.

Table 1. Human Evaluation Scores for Generated Summaries

Model	Informativeness (out of 5)	Coherence (out of 5)	Clinical Accuracy (out of 5)
BART	4.1	4.3	3.9
T5	4.0	4.2	4.0
Longformer	4.4	4.5	4.3

In further experiments, we have evaluated the performance of the model on the three classes of notes, namely discharge summaries, progress notes, and radiology reports. The variation in the performance is depicted in Figure 4 comparison of model ROUGE-L scores on the note types. Discharge summaries were the best performing model, probably because discharge summaries are more ordered and have common templates. The most difficult problem was progress notes, where once again Longformer was more resistant because of its ability to have a larger token window and greater understanding of temporally stratified information. This kind of comparison is essential to figure out the behavior of these models within real-clinical contexts where the variability of note format can be vast in detail and length.

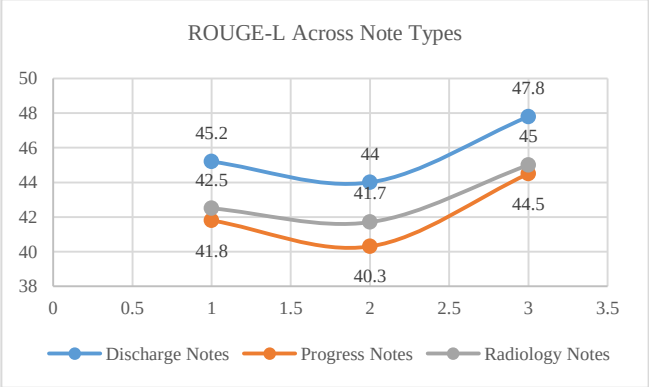

Figure 4. Rouge-L across Note Types

The other aspect of measuring was by monitoring training performance and program resource requirements. T5, with the shortest training time and the lowest architecture size, was also least robust to noisy inputs. Longformer needed the most GPUs memory and training time, though it was worth it in eloquence of accuracy and in depth of summarization. BART found itself in the middling position with an excellent compromise between fluency and computational speed. A comparative overview of these points is provided in Table 2: Computational and Training Comparison of Transformer Models, which can assist in future deployments, depending on the availability of resources and application limitations.

Table 2. Computational and Training Comparison Of Transformer Models

Model	Max Input Tokens	Training Time (per epoch)	GPU Memory Usage	Summary Fluency
BART	1024	25 min	10 GB	High
T5	512	20 min	8 GB	Medium
Longformer	4096	40 min	16 GB	High

According to user feedback and clinical evaluation, the other intriguing tendency was constantly observed: BART tended to generate readable summaries yet in some cases, too compressed ones, lacking important information. T5 generated facts correct summaries which were sometimes too abstract. Longformer generated denser and longer summaries containing nearly all applicable medical concepts, but at the expense of a mechanical-sounding output, especially when compared to Longformer ST. The differences can also be valuable when selecting a model on clinical setting: e.g., BART can be nominated to face patients in their summaries, whereas Longformer is better in physician use-cases where it is essential to have access to all details [3].

In general, the findings support the effectiveness of transformer models as a means of summary of long clinical notes; nevertheless, its architecture-restricted characteristics greatly influence its performance and the extent of its use. All three figures, 2,3 and 4, stress the multi-metric analysis that justifies the choice of Longformer as the best-compatible model to treat detailed medical summarization tasks. Table 1 and Table 2 are used to contrast the qualitative human dimension on the one hand as well as the practical deployment issues on the other hand. These results indicate that additional fine-tuning and workflow integration may drastically decrease cognitive load on the medical staff whereas still keeping patients safe and documentation consistent through the use of models such as Longformer.

5. Conclusion

The transformer models have proven to have an outstanding potential in summarizing long clinical notes, and Longformer has proven to be the best model capable of doing the same because of its massive capacity in accepting very long input sequences. BART and T5 have a problem with the input length, but they are still capable of creating coherent and fluent summaries. We find that application-specific transformers and architecture tuning result in much better summarization quality in the clinical domain even when the general-purpose transformers are useful.

Future research may examine integrating extractive and abstractive elements in a hybrid model, include medical knowledge graphs in checking facts and the degree to which the model is robust across the systems of various hospitals. With the further development of healthcare AI, it will be essential to incorporate trustworthy summarization solutions in the EHR systems that would assist with clinical decision-making and patient outcomes.

References

[1] Chaves, C. Kesiku, and B. Garcia-Zapirain, "Automatic text Summarization of Biomedical text data: A Systematic review," *Information*, vol. 13, no. 8, p. 393, Aug. 2022, doi: 10.3390/info13080393.

[2] L. Agilandeeswari, A. Dagar, A. Deepthi, and R. Arangasakthivel, "Automatic Text Summarization for Medical Dataset-An Analysis," in *Lecture notes in networks and systems*, 2024, pp. 336–352. doi: 10.1007/978-3-031-64813-7_35.

[3] S. Madan, M. Lentzen, J. Brandt, D. Rueckert, M. Hofmann-Apitius, and H. Fröhlich, "Transformer models in biomedicine," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, Jul. 2024, doi: 10.1186/s12911-024-02600-5.

[4] K. Denecke, R. May, and O. Rivera-Romero, "Transformer Models in Healthcare: A survey and thematic analysis of potentials, shortcomings and risks," *Journal of Medical Systems*, vol. 48, no. 1, Feb. 2024, doi: 10.1007/s10916-024-02043-5.

[5] R. Mehta, N. Mehta, V. Purohit, I. Saha, and P. Mishra, "Text Summarization for Research Papers using Transformers," *2022 IEEE 7th International Conference for Convergence in Technology (I2CT)*, Apr. 2024, doi: 10.1109/i2ct61223.2024.10543503.

[6] J. Whitton and A. Hunter, "Automated tabulation of clinical trial results: A joint entity and relation extraction approach with transformer-based language representations," *Artificial Intelligence in Medicine*, vol. 144, p. 102661, Sep. 2023, doi: 10.1016/j.artmed.2023.102661.

[7] Y. Dash, A. Kumar, S. S. Chauhan, A. V. Singh, A. Ray, and A. Abraham, "Advances in Medical text Summarization: Comparative performance analysis of PEGASUS and T5," *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–5, Jun. 2024, doi: 10.1109/icccnt61001.2024.10724845.

[8] A. Siebra, M. Kurpicz-Briki, and K. Wac, "Transformers in health: a systematic review on architectures for longitudinal data analysis," *Artificial Intelligence Review*, vol. 57, no. 2, Feb. 2024, doi: 10.1007/s10462-023-10677-z.

[9] R. Haruna, A. Obiniyi, M. Abdulkarim, and A. A. Afolorunsho, "Automatic Summarization of Scientific documents using Transformer Architectures: A review," *2022 5th Information Technology for Education and Development (ITED)*, pp. 1–6, Nov. 2022, doi: 10.1109/ited56637.2022.10051602.

[10] L. B. Elvas, A. Almeida, and J. C. Ferreira, "Natural language processing in medical text processing: A scoping literature review.," *PubMed*, vol. 204, p. 106049, Jul. 2025, doi: 10.1016/j.ijmedinf.2025.106049.

[11] Y. Gao, D. Dligach, T. Miller, M. M. Churpek, O. Uzuner, and M. Afshar, "Progress Note Understanding — Assessment and Plan Reasoning: Overview of the 2022 N2C2 Track 3 shared task," *Journal of Biomedical Informatics*, vol. 142, p. 104346, Apr. 2023, doi: 10.1016/j.jbi.2023.104346.

[12] T. Lai, "Interpretable Medical Imagery Diagnosis with Self-Attentive Transformers: A Review of Explainable AI for Health Care," *BioMedInformatics*, vol. 4, no. 1, pp. 113–126, Jan. 2024, doi: 10.3390/biomedinformatics4010008.

[13] S. Sharma, S. Srivastava, P. Verma, A. Verma, and S. N. Chaurasia, "A comprehensive analysis of Indian legal documents summarization techniques," *SN Computer Science*, vol. 4, no. 5, Aug. 2023, doi: 10.1007/s42979-023-01983-y.

[14] E. Kotei and R. Thirunavukarasu, "A Systematic Review of Transformer-Based Pre-Trained Language Models through Self-Supervised Learning," *Information*, vol. 14, no. 3, p. 187, Mar. 2023, doi: 10.3390/info14030187.

[15] Z. A. Nazi and W. Peng, "Large language models in healthcare and medical domain: A review," *Informatics*, vol. 11, no. 3, p. 57, Aug. 2024, doi: 10.3390/informatics11030057.

[16] Hands and R. Kavuluru, "A survey of NLP methods for oncology in the past decade with a focus on cancer registry applications," *Artificial Intelligence Review*, vol. 58, no. 10, Jul. 2025, doi: 10.1007/s10462-025-11316-5.