

A Two-Stage DistillBERT-BART Hybrid Pipeline for Efficient, Reliable, Stable Text Summarization

Oluwatomilola Motunrayo Moradeyo*, Ibiyinka Temilola Ayorinde**

*(Department of Computer Science, Adeseun Ogundoyin Polytechnic, Eruwa, Oyo State, tokenny2003@yahoo.com)

** (Department of Software Engineering, University of Ibadan, Ibadan, Nigeria, temiayorinde@yahoo.com)

Abstract:

This paper introduces a two-stage DistillBERT-BART hybrid pipeline designed to address the critical balance between factual accuracy, logical coherence, and readability flow in an automated text summarization. While a single direct summarization path produces either complete fact checked or highly readable summaries, its deployment is frequently challenged by disjointed outputs, high inference latency, and a tendency to generate hallucinations. To alleviate these limitations, the proposed framework is a two-stage process; an extractive selection stage and an abstractive rewriting stage. DistillBERT model was used for extracting and scoring the most salient sentences from the source document to generate a draft summary set in preparation for the second stage. BART-base model then receives this draft to generate a coherent and readable final summary. The two-stage summarization pipeline significantly reduced the load on the abstractive model by pre-selecting sentences and removing noise and irrelevant data before generating the final outputs. The new model was trained on the publicly available cnotesum dataset. Experimental evaluations were conducted on the generated summaries using both quantitative and qualitative text summarization metrics. ROUGE and BLEU metrics were used to determine the n-gram overlap and lexical similarity and model deviation of the summaries generated, while BERTScore, fluency, and hallucination rate were used to assess the semantic quality of the generated summaries. The final results were further benchmarked against two baseline models to establish the challenges of using a single type of summarization pipeline. The results showed that our two-stage pipeline achieved an outstanding performance with a high BLEU score of 68.4%, which indicated that the model successfully captured all important words. Its high BERTScore of 87.3% showed that the newly generated words used during rewriting retained the same meaning as they appeared in the original context. Furthermore, its high fluency score of 75.2% indicated that the generated outputs were highly readable and cohesive. Finally, its low hallucination rate of 6.0% confirmed that there were few or no fabricated words in the final summaries. Furthermore, human evaluations and factual consistency metrics confirmed that restricting the summarization to purely extractive model generated summaries that were factually correct but are disjointed and lacks coherence and fluent reading. Conversely, using purely abstractive model for such summarization will produce summaries that are logically coherent and fluently readable but it hallucinated heavily, it fabricated words that were not present in the original text. Hence, this paper has established that the abstractive text summarization model performed best when stabilized with extractive techniques.

Keywords: Text Summarization, DistillBERT, BART-base, Summaries

I INTRODUCTION

In today's world, where information is key and vital to any successful task, the amount of information being conceived by human beings is becoming increasingly large and expanding beyond human cognition. Hence, the need for a reduced version of complete information has become a necessity in current human life. The ability to efficiently process and comprehend vast amounts of textual data has become crucial. As digital information is fast growing, spreading, and multiplying fast in number, individuals and organizations have become overwhelmed with the amount of information they need to deal with. The act of removing all irrelevant and extraneous words and sentences, and processing and assimilating

only the germane insights is known as Text Summarization (TS). This has become essential for further action; this has become necessary and essential for timely information and intervention, but it still remains very challenging.

Text Summarization (TS) is defined as the extraction of the most salient sentences or phrases from a document to form a smaller and concise representation. This is often achieved through the identification of sentences exhibiting high lexical readability. TS is defined as an automated process that condenses large texts into shorter, coherent forms while preserving core meaning and minimizing hallucinations [1]. It was further emphasized that TS has emerged as a crucial solution to condense lengthy documents into concise, readable summaries without

compromising essential details [2]. TS is also defined as the process of creating a reduced, concise and accurate summaries that significantly preserves its essential meaning [3]. Similarly, TS is said to be a vital area in natural language processing that focused on condensing lengthy documents into concise, meaningful summaries [4].

However, TS is the process of reducing long voluminous texts into a shorter version while retaining the core idea and meaning of the original (or source) text. TS is a tool that can be used to reduce heavy information and documentation burdens on human with the use of NLP to automatically summarise unstructured long text into shorter versions [5]. Rather than relying on rigid, keyword matching tools of the past, contemporary implementations use domain adapted language models to extract critical words and details in a given document for paraphrasing into shorter notes [6]. When properly optimised, these text summary models can condense multi-page provider entries into precise timelines, effectively helping humans to navigate information overload and accelerating real-time decision making.

Overview of Text Summarization using Clinical Notes

Despite the clear and important advancement in Natural Language Processing (NLP) and Machine Learning (ML), TS remains an open challenge. It remains a highly complex problem that research has not yet solved completely, it does not have a perfect or fully satisfactory solution yet, because of some unique difficulties inherent. Such include coherence loss, hallucinations, evaluation issues, and length limits among others. Hence, it still remains an active area of ongoing research. Clinical narratives often contain fragmented sentence structures, medical jargon, and domain-specific abbreviations, complicating accurate interpretation and summarization [7], [8]. While purely extractive techniques ensure factual accuracy by selecting verbatim sentences from the source text, they tend to produce summaries lacking contextual coherence and narrative flow, rendering them suboptimal for clinical decision-making where nuanced understanding is crucial [9].

Conversely, abstractive methods, although capable of producing coherent and fluent narratives, but has the risk of introducing

hallucinations, that is, generating information not present in the source text [2]. This issue poses a severe threat in the clinical domain, where misinformation could mislead healthcare practitioners and jeopardize patient safety. The high computational demands of state-of-the-art abstractive models, severely limits the real-world application inside local medical facilities or resource-constrained clinical settings that operate on stricter infrastructure bounds [10].

Hence, the research question addressed in this work is how a hybrid text summarization system can be developed by integrating extractive and abstractive models to generate a factually accurate, logically coherent and fluently readable summaries using the cnotesum clinical note dataset. And to intentionally overcome the vital risk of hallucination. Bridging this gap is essential to support clinicians in quick assimilation of patient information and record, this will surely enhance the clinical workflows.

Purely extractive techniques ensure factual accuracy by selecting verbatim sentences from the source text, they tend to produce summaries lacking contextual coherence and narrative flow, rendering them suboptimal for clinical decision-making where nuanced understanding is crucial [9]. While abstractive methods, although capable of producing coherent and fluent narratives, but has the risk of introducing hallucinations, that is, generating information not present in the source text [2]. This issue poses a severe threat in the clinical domain, where misinformation could mislead healthcare practitioners and jeopardise patient safety.

The high computational demands of state-of-the-art abstractive models, severely limits the real-world application inside local medical facilities or resource-constrained clinical settings that operate on stricter infrastructure bounds [10]. Hence, the question addressed in this work is how a hybrid text summarization system can be developed by integrating extractive and abstractive models to generate a factually accurate, logically coherent and fluently readable summaries using the cnotesum clinical note dataset. And to intentionally overcome the vital risk of hallucination. Bridging this gap is essential to support quick assimilation of information and record, this will surely enhance quick decision making in any organization. When such organization rely on such recorded data or

information to make critical decisions, a single inaccurate number, misspelled word, or omitted fact can cause total system failure. Since, there has been a rapid increase in electronic documentation through

Electronic Health Records (EHRs) and the voluminous information required by medical experts is also increasing, hence the need for the summarization, so that the information overload can be reduced and thereafter enhance a real-time decision-making for experts. Clinical notes, including progress reports, discharge summaries, and consultation records, constitute a significant portion of unstructured data and contain essential clinical insights [7], [8].

However, their length, complexity, and specialized language create challenges for healthcare professionals who must quickly extract relevant information under time constraints [11]. Automatic TS offers a solution by transforming extensive clinical narratives into concise and readable summaries without compromising essential details [2]. However, summarizing text with purely extractive approaches protects factuality and at the expense of fluency by producing fragmented and disjointed summaries [1]. Conversely, migrating to an abstractive pipeline introduces deep transformer architectures that fixes the readability flow, yet these models require massive computing resources and remains highly vulnerable to generate factual inconsistency and hallucinated texts [12], [13], [3].

The limitations of the two summarization processes have brought about the need for the development of a hybrid summarization system that successfully summarizes text, based on factual accuracy of the extractive summarization process and simultaneously ensuring a logically coherent and readability flow of the abstractive summarization process. These two complement each other, and best applicable in clinical contexts where accuracy and reading coherence is critical [1]. The collaboration of the two method is essential because it represents an unshakable working relationship. The hybrid extractive - abstractive summarization system is designed to achieve a combined result that is much better than the result which each individual method would have achieved alone. This is also critical to clinical domains, where any minor abstractive fabrications

can severely cause alteration in medical decision-making accuracy [14], [15].

To address these limitations, hybrid summarization approaches have emerged, combining extractive and abstractive techniques within a unified framework [16]. These approaches leverage extractive methods to anchor summaries in factual content while utilizing abstractive models to enhance coherence and readability [17]. The integration of the two summarization methods to form the hybridization is very important in domains where accuracy, clarity and fluent reading are critical and extremely important. The healthcare sector is an example of such domain because they deal with human life. In such work environment, simple mistake or recording error in any documented data or information can lead to disastrous, life-threatening or highly destructive consequences.

The scientific purpose that inspired this research arises from the pressing need for automated systems that can summarise long, complex and voluminous texts into shorter, simpler and brief texts that can capture the whole narratives of the source text to facilitate rapid decision-making in any working environment. Looking further into the clinical environment in the healthcare domain, it was discovered that raw voluminous clinical notes and records available in their custody is becoming more huge and larger with each passing day while the number of human healthcare expert keeps dropping due to search for greener pastures outside the country. The large amount of text is becoming a significant challenge for human cognitive processing, as reading through it may likely lead to information overload that surpasses a clinician's functionality, hence, a severe cognitive saturation tends to occur [18], [19].

The Current Challenge

This overload of information does not only exhaust clinical staffs that are directly involved, it also reduces core information awareness, likewise, it can also cause missed placed medical priority like false diagnosis and wrong prescription. In the same vein, it can also cause a costly delay for time-sensitive patient that requires interventions [20]. However, an effective and efficient text summarization system that can deliver accurate, concise, coherent, and readable narratives of any

given texts can alleviate this heavy cognitive load burden, and further allow healthcare providers to focus on mainly critical patients' care activities.

Furthermore, previous research, by [2], emphasizes the pertinent risks of hallucination misinterpretation of facts that has been discovered by using direct abstractive models. Therefore, a hybrid summarization approach that is based on factually extracted content by the extractive layer and refined by rewriting and paraphrasing via the abstractive layer, these addresses both challenges while retaining accurate facts, high coherence, and readability flow of text [16].

The healthcare sector is striving globally on daily basis with their level of digital transformation and use of automated systems, but the strong need and desire for scalable and sustainable solutions for clinical note summarization has never been higher [17]. Automated clinical text summarization is essential for reducing administrative burden and extracting insights and understanding from clinical notes. However, deploying fully generative Large Language Models (LLMs) in clinical workflows presents severe compromise in computational efficiency, factual reliability, and model deviation. Pure abstractive summarization excels at generating fluent summaries but suffer from one major critical flaw when applied to clinical narratives directly. Unstructured clinical notes are very lengthy, complex, and redundant. Processing these massive word loads directly through direct abstractive models creates substantial computational bottlenecks, leading to high risk of real-time, patient care deployment in the application domain.

The model often fails to ensure their outputs are grounded truth, leading to critical errors such as fabricated medical conditions, omitted clinical events, or altered medication dosages. In the healthcare domain, these ungrounded hallucinations pose direct risks to patient safety and clinical decision-making. Consequently, there is an urgent gap in medical AI for a resource-efficient framework that can compress large clinical notes without losing vital details, and ensuring that the final summaries are computationally scalable, factually anchored, logically coherent and fluently readable to understand the patient's actual record.

This study is specifically restricted to text summarization using clinical notes. It is limited to the relative performance of two specific

lightweight transformers, which are; DistillBERT and BART-base. DistillBERT is used for the extractive summarization aspect, which is the first stage of the hybrid summarization process. And BART-base is used for the development of a subsequent hybrid system, which is the second stage of the hybrid system. Cnotesum clinical note dataset was utilized to implement this work. It is available online and hosted on Hugging Face.

II Previous Works

A large-scale dataset called MS² was designed for multi-document summarization in the biomedical domain by [2]. Their work addressed the issue of time-consuming and manual process of synthesizing evidence from numerous studies to evaluate the effectiveness of medical interventions. The authors formulated a texts-to-text summarization task. Their models generated concise summaries from multiple study abstracts and contextual background information gathered for the document used. which was exactly their target statements. A major contribution of this work lies in its scale, domain specificity, it specialized only on multiple biomedical document summarization. It also successfully integrated machine-readable medical information and entities like gens, diseases and drugs and the relationship among them. However, the study revealed that their current model still struggles with the aspect of how to capture contradictory evidence accurately. Though the generated summaries aligned with the reference summary standard with output showing about half of the time. This established the need for further improvements in word logical reasoning and factual consistency of the outputs generated [2].

An exploratory study was conducted on an Automatic Text Summarization (ATS) system for biomedical and healthcare domain. The authors focused on the use of ATS to mitigate information overload. ATS generated concise and relevant summaries that supported clinicians, researchers, and other stakeholders in a faster decision-making domain environment, and also added their own experience and discoveries to the stream of existing knowledge. Their work introduced two major methodological contributions that showed how they built and trained their model. First, they provided the basis for categorizing the summarization techniques into extractive and

abstractive approaches. And also, established the criteria for sorting summaries into indicative and informative summaries. Secondly, their work further emphasized the growing shift as it transferred from the summarization from a single-document to multi-document summarization, and also the change and improvements from extractive summarization methods to abstractive methods. Their work specifically improved how the system responded to the increasing complexity and diversity of biomedical data, such as having a mixture of different elements in the dataset. However, the review was largely descriptive but the performance of their model was not analyzed with any standard evaluation metrics, hence its effectiveness may not be determined. This is another research area for future work [21].

A knowledge-guided multi-objective optimization framework was proposed by [22] to improve the factual accuracy of abstractive clinical text summarization. The authors made use of encoder-decoder transformer models. The study jointly optimized generative loss, entity loss, and knowledge loss to preserve semantic and clinical correctness in generated summaries. The framework combined clinical named entity recognition and knowledge retrieval from medical knowledge bases. Experimental results on Heart Failure, IU X-Ray, and MIMIC-CXR datasets showed an improvement in the entity-level factual accuracy when compared to standard transformer summarization methods. Their work concentrated on enhancing factual accuracy in generated summaries. Despite making use of multi-objective optimization, their work was challenged in the area of computational complexity by depending on extensive fine-tuning, and potential information loss during the summarization stages [22].

Medical reports summarization system was developed by [23] using Text-to-Text-Transfer-Transformer (T5). They addressed the challenge of making medical reports more accessible, readable and understandable to non-expert users, especially patients and their family members who may struggle with complex medical terminology and medical abbreviations. They proposed an abstractive summarization approach which leveraged the T5 encoder-decoder architecture to convert lengthy and technical medical documents into shorter, simplified summaries. Their model was trained on the Indiana dataset. The results they

obtained serves as input and the corresponding outputs were their expected target summaries. This process allowed the task to be treated as a text-to-text generation problem, because it enabled the model to learn how to condense and simplify specific vocabularies and syntax in the medical domain. However, the generated summaries were shorter and simpler, but they lack full clinical correctness and some critical nuances were significantly omitted, this gave room for improvements in future works. Finally, the study focused solely on T5 model, leaving the investigation of larger and more advanced transformer models as an open direction for future research [23].

A critical challenge of the complexities associated with LLMs in clinical NLP was addressed by [24] using lightweight transformers. The authors developed efficient transformer-based models that produced quality outputs through maximum speed without wasting time, energy and resources. Their models were light weighted, and are capable of processing complex clinical texts within the shortest possible time. They developed a comprehensive framework for building compact clinical language models using Knowledge Distillation (KD) and confusion learning. First, the authors designed three distilled models; DistilClinicalBERT, TinyClinicalBERT, and ClinicalMiniALBERT) as variants extracted from BioClinicalBERT. And Secondly, they introduced an alternative approach using continual learning for the already compressed biomedical models; BioDistillBERT and BioMobileBERT for further pre-training on the clinical text MIMIC-III dataset. This multi-task evaluation demonstrated the generalizability and robustness how the developed lightweight models can be applied in different settings in different real-world applications. However, the models are lightweight, there is limited discussion on the trade-off in terms of some inadequacies that might have occurred in terms of the size and processing speed of the model compression and performance degradation across different tasks. Also, the authors focused on text-based data only. Their work did not integrate multimodal clinical data such imaging, structured medical variables, or real-time monitoring data, like the vital signs. This could further enhance model performance [24].

A hybrid architecture with efficient fine tuning for abstractive patent document summarization was developed by [25]. The authors addressed the challenges of difficulty of generating coherent and concise summaries from long, complex, and technically dense patent texts while maintaining computational efficiency. Their hybrid architecture integrated graph-based extractive component called LexRank and BART as their abstractive techniques. LexRank was used to identify and rank the most important sentences from patent documents based on cosine similarity and eigenvector centrality and BART was used for the final paraphrasing process. This approach ensured that only the most relevant contents were forwarded as input to the abstractive stage. BART fine-tuned the contents received using Low-Rank Adaptation (LRA), this parameter-efficient fine-tuning technique updated only low-rank matrices and restricted the original model weights. This significantly reduced the computational cost without sacrificing the model's performance. The authors made a significant contribution by demonstrating that combining extractive preprocessing with an efficient transformer fine-tuning produced high-quality and readable summaries with a highly reduced computational requirement. Despite achieving these outstanding results, the model did not surpass the performance of the state-of-the-art models, partly due to the limited constrained hyperparameter tuning, their inability to optimize their model's training settings under the strict limitations of computational resources, time and budget [25].

The authors in [26] proposed scientific evidence for clinical text summarization using large language models. The authors addressed the challenge of difficulty in generating accurate, and the ability to understand words, sentences and the overall meanings of the clinical summaries from complex and heterogeneous EHR that contained diverse dissimilar types of data. They also ensured that the LLMs used generated outputs that were readability and useful in real-world medical domain. Their work focused on radiology report summarization, especially the impression section. They incorporated both human evaluation and critical evaluation components for final assessment of their output. They observed that LLMs showed promise in reducing clinician workload and improved information synthesis. However, despite

the encouraging improvements, current evidence is insufficient to support their safe and reliable deployment in clinical settings. Hence, the use of LLMs for clinical text summarization cannot yet be considered fully reliable without substantial improvements in evaluation to verify and validate that it is safe for use in real-world medical domain [26].

Researchers in [27] presented a practical and empirical study on clinical text summarization. The authors designed, implemented and evaluated the LLMs system for patients' discharge summaries in the German hospital. The authors addressed the issue of summarising complex and diverse health records required by the system to handle accurate timelines, understand medical jargon, and maintain strict clinical precision. Their clinical summarization approach was structured around temporal windowing and hierarchical event grouping, to map out patient records. They later evaluated different input encoding strategies, including; Flat Text, JSON, Markdown, and XML to determine their impact on the quality of the summaries generated. This work offers strong real-world analysis, but the performance was limited because it was only tested a single model without any fine-tuning [27].

Evaluation Metrics

Five text summarization metrics was considered for the assessment of the outputs generated. This includes; ROUGE, BLEU, BERTScore, Fluency and Hallucination Rate. ROUGE score was used to calculate recall, the amount of words and phrases the generated summaries was able to capture from the reference summaries. BLEU score is used to calculate precision, that is, how accurate the generated summaries are. BERTscore is used to check semantic and synonyms of words in the summaries. Fluency score is used to check the readability flow of the generated summaries and the hallucination rate is used to check the factual deviations found in the summaries.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) framework measures how a machine-generated summaries overlaps with a human-written reference summary. It evaluates the

model performance by calculating the overlap of matching word sequences against the reference summaries [28]. That is the total number of the exact words or phrases found in the generated summaries, that were originally available in the reference summary. ROUGE focuses on recall. It measures how much information from the reference summaries is captured in the machine-generated summary. ROUGE's core strength lies in its suite of metrics. This collection of different scoring methods is used to measure different lexical overlap, depending on the nature of task to be evaluated. The most common metrics in this suite include; ROUGE-N, it is used to calculate n-gram overlap, the sequence of n words that overlap between the generated summary and the reference summary. ROUGE-SU4 evaluates skip-bigram overlap, this measures the informative content. This are sometimes spread across non-consecutive words.

This ROUGE variant allows for a more comprehensive assessment of the quality of the summaries generated. ROUGE's widespread adoption is as a result of the several advantages it has over other metrics. First, its focus on recall ensures that the generated summaries are highly informative, it captures all the major aspect of the reference summary to ensure that no text is overlooked. Secondly, the multiple metrics it has allows different categories and perspectives of the summaries to be evaluated. This allows researchers to prioritise specific aspects of the summarization quality. Again, ROUGE evaluates summaries the same way a human naturally does, it checks if the summary contains the most important words and phrases from the reference summaries text.

However, ROUGE has the ability to award high scores to summaries that simply copy heavily from the source text without capturing the essence of the content. It grades the reuse of exact words and phrases from the original text, even when the summary misses the exact meaning or main idea of the content. Again, ROUGE's reliance on reference summaries to assess the machine-generated summaries, has the tendency to creates problems with human bias and lack of reproducibility when result changes according to different people's summaries [28].

Bi-Lingual Evaluation Understudy (BLEU)

Bi-Lingual Evaluation Understudy (BLEU) was originally developed to evaluate machine translation tasks, Before the era of text summarization began, string-alignment baselines like the BLEU framework was used to measure and compare generative text for nearly twenty years [28]. However, because exact word-matching strategies cannot identify words that have the same definition but has different connotations or grammar and rewriting of sentences by altering its grammatical architecture and word other but still keeps the core message exactly the same. Relying only on the basic token overlap metrics is becoming insufficient for today's modern advanced language generation tasks [29].

It compares the sequences of n words (n-gram overlap) between the machine-generated summaries and the given reference summaries created by humans. High BLEU scores indicate a greater degree of overlap between the machine summary and the human references, that is they both share a large amount of identical or highly similar elements such as words, phrases or sentences This suggests that the machine captured the key points of the source text. The strength of BLEU lies in its quality of been easy to understand, interpretate, and its ability to handle different lengths of n-grams. Its foundation in n-gram precision aligns with human beings' ability to understand good summaries containing core factual elements from the source instantly without and form of conscious reasoning or step-by-step logic text [16].

Bidirectional Encoder Representations from Transformers (BERT)

Bidirectional Encoder Representations from Transformers (BERT), introduced by [30] has significantly impacted the field of text summarization. BERT leverages deep learning to learn basic grammar, facts, words and relationship among the words. It also uses deep learning to find complex patterns and meanings in human language. BERT's primary advantage is its ability to understand the general rules, context, and nuances of human language before it is trained on a specific task. Instead of starting training from the scratch, it arrives with a massive in-built foundation of language knowledge. BERT understands the context of a whole sentence and

how words relate to each other, since it understands the true meaning of the text, it can easily select the main points and ignore the extra, less important details to create a summary that perfectly captures the core message and meaning of the original article or document. This robust understanding is very useful for any text summarization task [31].

Research has shown that BERT has been used as a starting point and trained further on specific text data to make it effective for summarization [32]. Various encoder-decoder architectures leverage BERT's pre-trained encoder to extract crucial information from the source text and then uses a decoder to generate coherent summary [32]. BERT has several advantages when applied to summarization task. For example, its pre-trained nature has taken over, such that, humans no longer have to manually design specific rules or mathematical formulas to help the model understand words, because it has already learned the structure of language from massive amounts of data. This makes it more efficient and scalable. Next, it has the ability to capture the entire document at once, it doesn't just look at words individually, it looks at how words relate to each other across the entire document. Finally, BERT's is highly flexible and adaptable, it can be easily combined with different text summarization architectures to build a new or better text summarization system.

Fluency

Semantic similarity and factual alignment paradigms are basically used to verify that any how well the generated summaries retain the primary idea of a given source document, they do not check whether the text is readable, fluent or structurally coherent [32]. To address this limitation within computational linguistic of human language from a computer perspective, which serves as the theoretical foundation behind how machines parse, interpret, and generate human communication. Fluency functions as a crucial reference-free dimension. It is used specifically to gauge the surface-level grammatical integrity, prose velocity, and formatting coherence of machine-generated text. It operated separately from the baseline source context. The reference summary is not needed to determine the effectiveness of the fluency metric. Over several decades, the methodology used for fluency assessment has undergone a major

paradigm shift. This shift represents a fundamental transformation in how we measure, test, and validate language data. It moved from era of hiring human linguists, proof-readers or subject matter expert to read and evaluate the content line by line, to the use of LLMs and advanced algorithms to evaluate the summaries. Systematically dividing the natural language evaluation metrics into the dual axes of fluency and semantic adequacy was traced back to the early 1990s.

Hallucination Rate

The hallucination rate serves as a vital evaluation metric developed to quantify the number of factual deviations from original contents and enforce generative safety protocols within text summarization workflows. While surface-level readability metrics verify that a generated summary is grammatically sound, they remain incapable of guaranteeing that the output text is factually accurate or reliably anchored in reality [33]. To systematically measure the propensity of deep neural networks to produce linguistically polished yet unverified or deceptive assertions, modern evaluation frameworks rely explicitly on hallucination rate computations [34]. This diagnostic metric operates as an essential safety indicator, calculating the precise ratio of fabricated, ungrounded, or conflicting propositions relative to the summary's total information payload [33], [34]. Rather than a single tool created by an individual inventor, tracking hallucination velocities represents a major collaborative domain within artificial intelligence, formalizing how engineers detect and audit systemic alignment anomalies during automated summarization tasks.

III Materials and Methods

The development of this two-stage text summarization system for clinical notes started with a comprehensive review on current clinical text summarization literatures. The goal is to bridge the identified literature gaps. The development was carried out through a five-stage operational workflow as shown in Figure 1. The process started with the selection of different clinical notes collected from cnotesum datasets (available online). This work developed a specialized extractive TS framework designed to systematically reduce long, unstructured clinical notes into concise, factually correct versions

tailored for rapid interpretation by healthcare professionals.

Data Preparation and Pre-Processing

This framework utilizes the cnotesum dataset as its choice of selected data source. It was chosen because of its emphasis on real-world clinical notes and its ability to pair full clinical documents with machine generated summaries. It was also considered because of its suitability for transformer-based modelling. It is in a tokenised text format for transfer input, it contains thousands of document-summary pairs (that is, human-written summaries) written in English language. The document contains progress notes, discharge summaries and consultation report. This ensures the research is firmly grounded in authentic clinical language, reflecting the complexities of real clinical documentation.

Figure 2 displayed the dataflow diagram of this work. It showed the direct comparative setup of the new hybrid model and the two baseline models: DistillBERT and BART-base models. The initial dataset is routed to the DistillBERT model, the BART-base model and the DistillBERT-BART models separately to ensure all models are evaluated under the same conditions and the same sets of data. Once each model generates its output, the data stream serializes. It moves strictly from evaluation to analysis and to optimization. This creates a clearer feedback loop for further decision making.

Effective pre-processing of clinical text is crucial for building robust summarization models. Cnotesum dataset, while pre-processed to a degree, still contains jargon, abbreviations, irregular formatting, and sensitive patient information. Hence, still required careful preparation to ensure that the hybridised model operates on clean, consistent, and appropriately tokenised data. To achieve this, the following steps were taken;

Although the cnotesum dataset is publicly released in de-identified form, we went further to normalise all whitespaces ensuring that all extra tabs and line breaks were standardised and all artefacts were also removed. Redundant metadata stripping was further performed because some notes contained header artefacts (such as document generation timestamps) which are not relevant to the clinical contents. With these we were able to reduce the noise and all irrelevant tokens that might want to influence the model's learning were carefully avoided.

Tokenization is common to all transformer models. They operate on sub-word or token units rather than raw text. The following decisions guided tokenisation: Sub-word tokenisation uses Hugging Face's tokenizers corresponding to each chosen model. For truncation and padding we applied maximum sequence lengths specific to each model architecture. For instance, DistillBERT is limited to 512 tokens per document segment. This step is critical to align the dataset with the foundational input configurations of the model.

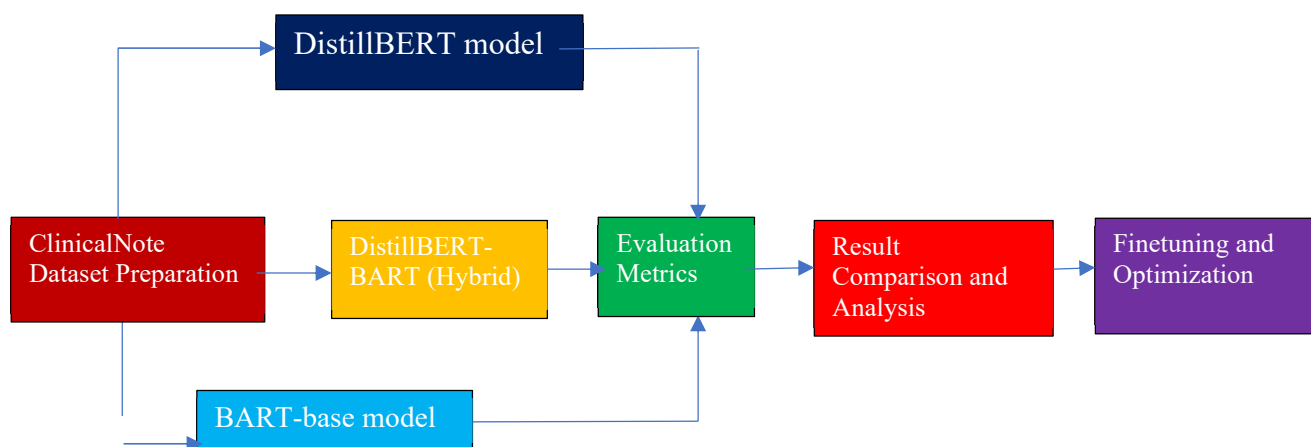


Figure 1: Conceptual Framework

DistillBERT-BART Model Framework Summarization

DistillBERT model was used for the extractive summarization process. It was considered for this framework because of its ability to balance performance and computational efficiency (Sanh et al., 2019). When compared to other variants of BERT, DistillBERT offers: approximately 40% fewer parameters than BERT-base, 60% faster inference time and retention of approximately 97% of BERT's language understanding capacity. Such characteristics are critical to ensure the model's suitability for instantaneous or low latency clinical deployment (Coimbra et al., 2024). The summarization framework for CNoteSum operates by selecting and concatenating the most informative sentences from the reference clinical notes that capture the core information relevant for clinical use.

DistillBERT Extractive Design

The extractive summarization system operates as follows:

1. Sentence Segmentation: Clinical notes are divided into sentences using SpaCy's medical model, ensuring medical entities are not incorrectly split.
2. Sentence Embedding: Each sentence is tokenised and passed through DistillBERT to produce contextual embeddings.
3. Document Embedding: A global embedding for the entire note is computed using mean pooling across sentence embeddings.
4. Similarity Scoring: Cosine similarity between each sentence embedding and the document embedding is calculated, producing relevance scores.
5. Sentence Ranking: Sentences are ranked by similarity score.
6. Selection: The top k sentences (hyperparameter) are selected as the extractive summary.

Design of a two-stage DistillBERT-BART text summarization pathway for Clinical Notes

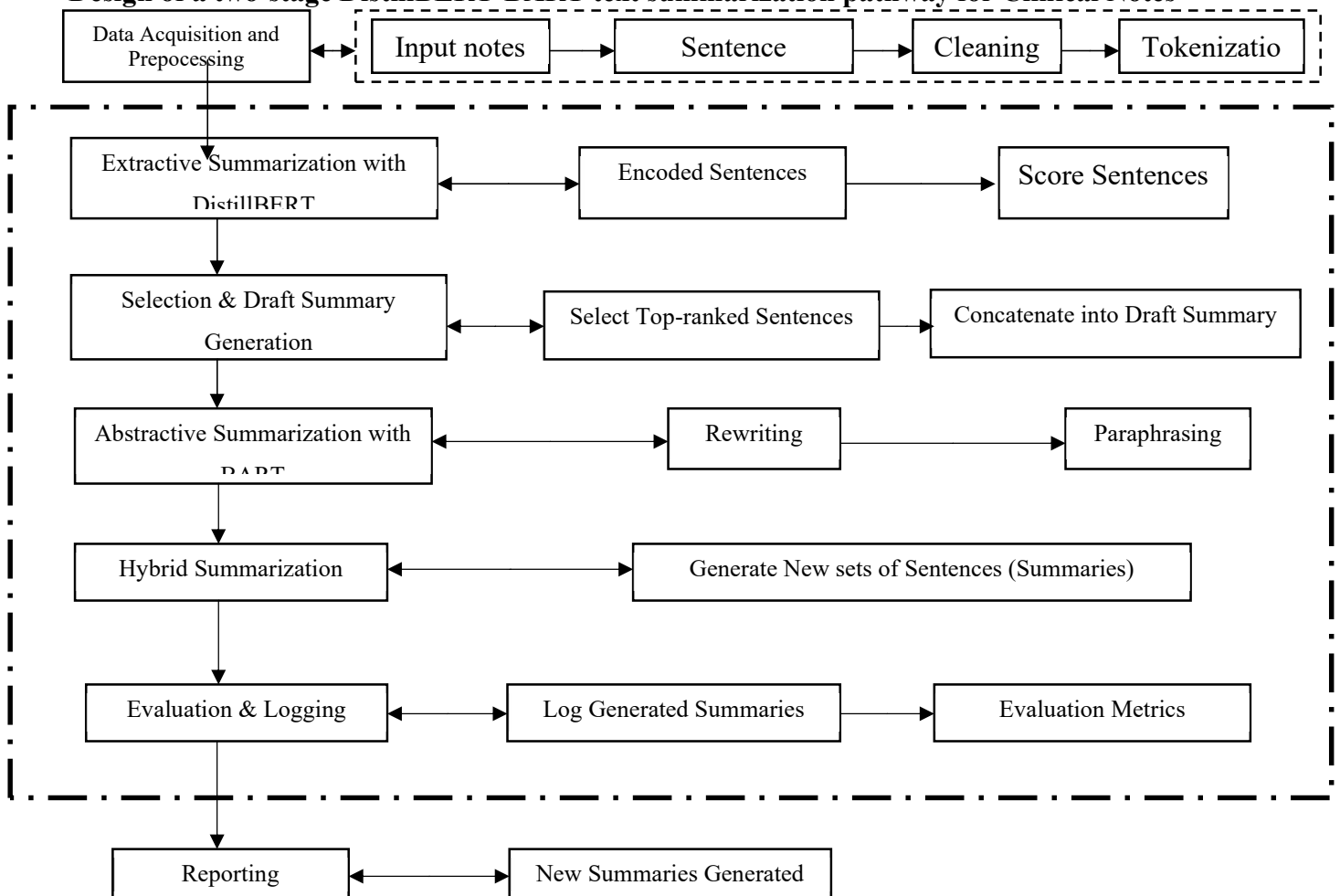


Figure 2: The DistillBERT-BART Summarization Methodology

Stage 1: Sentence Representation DistillBERT Embeddings: Each clinical note (d) is split into n sentences ($S_1, S_2, \dots, S_n$). DistillBERT encodes each sentence into vector representation

$$v_i = \text{fDistillBERT}(s_i), \quad v_i \in \mathbb{R}^m \quad (1)$$

Where $v_i = s_i$ (a sentence in the reference summary)

Stage 2: Sentence Scoring: A scoring function assigns an importance value to each sentence

$$\text{score}(s_i) = w^T v_i \quad (2)$$

Where v_i = the vector representation of each sentence in the reference summary.

w^T = the weight assigned to each vector (the weight is determined by the frequency of each word in the sentence)

Stage 3: Ranking and selection: Sentences are ranked by scores (s_i). The top-k sentences are selected such that

$$S^* = \arg \max_{S \subseteq \{s_1, \dots, s_n\}, |S|=k} \sum_{s_i \in S} \text{score}(s_i) \quad (3)$$

Where $\arg \max$ = argument of the maximum (this selects the sentence with the highest n score. i.e. the top-k sentences. Sentences with the highest score

Stage 4:

Summary Construction: The extractive summary is obtained by concatenating the selected sentences in their original order:

$$\text{Summary} = \bigoplus_{s_i \in S} s_i \quad (4)$$

where

$\bigoplus$ = the summary construction (i.e. concatenating the extracted sentences in their original order)

R = Reference

g = no of n-grams

The Abstractive Algorithm

i. The extractive sentences that were generated in the first stage were concatenated into a single text block.

- ii. The abstractive model takes as input the concatenated sentences extracted from the extractive layer (this significantly reduced the risk of hallucination)
- iii. The input length was limited to 512 tokens for T5-small and 1024 for BART-base which was the maximum tokens permitted by each model.
- iv. Special tokens were appended to signal end-of-input: $\langle \text{pad} \rangle$ and $\langle \text{eos} \rangle$ were added for T5 and $\langle s \rangle$ and $\langle /s \rangle$ were added for BART-base.

Hybrid Pipeline Integration

The full hybrid summarization algorithm runs as follows:

1. Input Clinical Note: segmented into sentences.
2. Extractive Summarization (DistillBERT): top k sentences selected.
3. Concatenation: selected sentences combined into a single string.
4. Abstractive Summarization using BART-base rewritten into a fluent summary.

This ensures the abstractive model only generates text based on factually vetted input, reducing hallucination risk.

Sequence-to-sequence generation:

The conditional probability is:

$$P(y|x) = \prod_{t=1}^T P(y_t | y_{<t}, x; \theta) \quad (5)$$

Training objective (Maximum Likelihood Estimation): The models are trained to minimize negative log-likelihood:

$$\mathcal{L}_{NLL}(\theta) = -\sum_{t=1}^T \log P(y_t | y_{<t}, x; \theta) \quad (6)$$

Refinement-Discourse and Fragment Fixing: To improve coherence, a penalty for fragmented sentences or missing connectors will be introduced, where $\text{FragPenalty}(y)$ increases when output lacks proper discourse structure.

$$\mathcal{L}_{ref} = \lambda_1 \cdot \mathcal{L}_{NLL} + \text{FragPenalty}(y) \quad (7)$$

Where

$P(y|x)$ = probability of the entire target sequence y occurring given the input sequence X

$\mathcal{L}_{ref}$ = Reference Loss (i.e. total training loss to be minimised)

Penalty = (no of chunks)/(no of matches)

Frag(y) = penalises the model when it produces undesirable fragmentation
 λ_1 = prediction accuracy
 λ_2 = penalty for fragmentation

Evaluation (Logging)

ROUGE (n-gram Lexical Match)

ROUGE-N metric was used to evaluate the quality of generated summaries based on the overlap of n-grams between the generated summary and reference summaries. It further assessed how well the model captures key factual information from the source document as shown in the equations below.

N-gram Matching: ROUGE-N calculates the overlap of n-grams (e.g., unigrams, bigrams) between the generated summary and reference summaries.

$$\text{ROUGE} - N = \frac{(\sum \text{Count}(\text{match_n_gram}))}{(\sum \text{Count}(N_gram) \text{ in reference summary})} \quad (8)$$

Precision: The ratio of overlapping n-grams in the generated summary to the total n-grams in the generated summary.

$$\text{Precision} = \frac{(\sum \text{Count}(\text{identified } N_gram \text{ in generated summary}))}{(\sum \text{Count}(N_gram) \text{ in generated summary})} \quad (9)$$

Recall: The ratio of overlapping n-grams in the generated summary to the total n-grams in the reference summaries.

$$\text{Recall} = \frac{(\sum \text{Count}(\text{identified } N_gram \text{ in generated summary}))}{(\sum \text{Count}(N_gram) \text{ in reference summary})} \quad (10)$$

F1 Score is the harmonic mean of precision and recall. It provides a balanced measure of the quality of the generated summary.

$$\text{F1 Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (11)$$

Summarily, the average of all F1 scores across all N-gram sizes gave us the final ROUGE-N score. The average percentage of ROUGE-N scores indicate better summarization quality (more overlap with reference summaries). Low percentage signals poor performance (summaries not well captured) and a very high percentage also signals overfitting (error during process).

ROUGE-L was used to evaluate the Longest Common Subsequence (LCS) between the generated summary and reference summaries. It further

measured the ability of the model to preserve the order and structure of the original text as shown in the equations below

LCS: Length of the Longest Common Subsequence between each generated summary and each reference summary is calculated as shown in equations 9 to 11.

$$\text{ROUGE} - L = \frac{(\sum \text{LCS}(\text{system_summary}, \text{reference_summary}_i))}{(\sum \text{Length}(\text{reference_summary}_i))} \quad (12)$$

$$\text{Precision} = \frac{\text{Length of LCS}}{\text{Length of the generated summary}} \quad (13)$$

$$\text{Recall} = \frac{\text{Length of LCS}}{\text{Length of the reference summary}} \quad (14)$$

$$\text{F1 Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (15)$$

The Average Rouge Score (ARS) outlines the steps to calculate the average ROUGE score (combining the ROUGE_N and ROUGE_L). This provides a single score that combines the strengths of different ROUGE variants.

BLEU (Precision of n-grams with brevity penalty):

$$\text{BLEU} = \text{BP} \cdot \exp(\sum_{n=1}^N w_n \log p_n) \quad (16)$$

Where

BP = Brevity Penalty (i.e. translations that are too short compared to the reference summary)

w_n = weight assigned to n-grams

p_n = precision for n-grams

$$\text{F1 - Score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (17)$$

BERTScore(semantic similarity with embeddings):

$$\text{BERTScore}(H, R) = \frac{1}{|H|} \sum_{h \in H} \max_{r \in R} \cos(e(h), e(r)) \quad (18)$$

where

H = Generated summaries (GS)

R = Reference summaries

|H| = Total number of tokens (words) in GS

$\cos(e(h), e(r))$ = the cosine similarity between the vector embeddings of the two tokens.

The fluency metric as shown in equation 13 and 14 were used to measure how grammatically

correct the generated summaries were. This was done independently not considering whether it is factually correct or matches the reference text.

Perplexity (PPL) =

$$\exp \left(-\frac{1}{N} \sum_{i=0}^n \log P((t_1, \dots, t_{i-1})) \right) \quad (19)$$

This equation represented Perplexity (PPL), which is a primary metric used to evaluate the performance of language models by measuring how well they predict a sample of text.

where:

N = This is the total number of words or tokens in the evaluation test dataset.

Log = Logarithm. It transforms the multiplied probabilities of words into added values.

Fluency Score (Scaled) =

$$\max \left(0, 10 \cdot \left(1 - \frac{\log(PPL)}{\alpha} \right) \right) \quad (20)$$

PPL: Perplexity. The raw evaluation score from your language model. As established, lower perplexity is better, and a baseline model usually has a higher perplexity.

log (PPL): This converts the exponential perplexity score into a linear

Alpha: This is a hyperparameter set by the researcher representing the "maximum acceptable log-perplexity." It dictates the worst allowable performance before the score hits zero.

Hallucination Rate, was used to measure the proportion of factual claims in the generated summaries that were not present in the reference summaries. In clinical summarization, hallucinations are patient safety risks, a fabricated drug name, wrong dosage, or invented diagnosis can cause direct harm.

$$\text{Hallucination Rate}(\%) = \left(\frac{H}{K} \right) \times 100 \quad (21)$$

Where:

H: Number of hallucinated elements.

K = The total number of facts, statements, or tokens checked during the evaluation.

Depending on the specific benchmark used, K can be the total number of sentences/propositions generated or the total number of factual claims tested or the total number of queries or prompts evaluated.

4 RESULTS AND DISCUSSION

This section presents a detailed discussion and analysis of the experimental results obtained from the two-stage DistillBERT-BART text summarization architecture. The performance evaluation of the machine generated summaries against the reference summaries (ground-truth) was conducted on both quantitative and qualitative qualities. The ROUGE and BLEU scores were the quantitative metrics used to assess how well the generated summaries captures the reference summaries (Recall) and the factual accuracy of the generated summaries (Precision) respectively. The metrics used to evaluate the quality of the generated summaries were BERTScore, fluency and hallucination rate. BERTScore was used to assess the semantics and meaning of new words generated, to ensure it has the same meaning with the context of the reference summary. Fluency and hallucination rate were used to check the logical flow (readability) and to check for the presence of redundancies or machine fabricated words (hallucinations) in the machine generated summaries.

4.1 Quantitative and Qualitative Evaluation of the Generated Summaries

Table 1 shows the comprehensive analysis, justification, and architectural reasoning for the performance metrics of each approach. It displayed the raw scores in percentage as evaluated using the generated summaries from the developed two-stage DistillBERT-BART framework system compared with the two baseline models.

TABLE I: Performance Evaluation and Comparison of DistillBERT-BART and Baseline Models

METRIC	Quantitative Metrics		Qualitative Metrics		
	ROUGE	BLEU	BERTScore	Fluency	Hallucination Rate
Purely Extractive (DistillBERT)	58.3%	51.2%	51.2%	32.0%	4.0%

Purely Abstractive (BART-base)	45.5%	-14.4%	24.8%	42.4%	24.0%
Hybrid Two-Stage (DistillBERT-BART)	47.8%	68.4%	87.3%	75.2%	6.0%

DistillBERT extracted sentences directly from the source text without modifying them. Its strength is in ROUGE and hallucination rate. It achieved the highest ROUGE score of 58.3%. This was achievable because it copied the reference texts verbatim, hence this guaranteed a high literal match because ROUGE heavily penalizes variations in n-grams overlaps. Similarly, it has the lowest hallucination score because it cannot fabricate or invent new words, but has the possibility of making selecting errors regarding the context. However, DistillBERT joined together selected sentences from different parts of the reference document, the summaries generated from it lacked heavy narrative flow, readability and coherent structure. Hence the reason for its low fluency score of 32.0%.

BART-base model generated a set of entirely new sentences to summarize the reference text to generate human-readable text. Its strength in fluency with a relatively high score of 42.4%, this was achievable because of its sequence-to-sequence architecture ability to generate grammatically smooth transitions when paraphrasing. But a major weakness is the rate at which it fabricated new words and meanings to existing words, this was the major reason for its poor performance in both semantics metrics. This includes a disastrous negative BLEU score of -14.4% and a low BERTScore of 24.8%. This was as a result of its high rate of hallucination with a score of 24.0%. If there is no further improvement on this, BART model is liable to generate inaccurate facts or fabricated details.

The hybrid two-stage model used DistillBERT to filter out irrelevant information and extract the core sentences first. Then, BART-base was used to process only the extracted sentences to generate a polished, paraphrased summary. By feeding the BART model with only a curated factual subset of text, rather than the whole reference text, the search space is narrowed. This was responsible for the high BLEU score of 68.4% and a very high

BERTScore of 87.3%. This proved that the generated summaries were both exact in phrasing

and strongly aligned in meaning. Again, the filtered text eliminated the background noise, and allowed BART to focus only on articulation. This was responsible for the high fluency score of 75.2%. It also significantly anchored the new two-stage model to factually correct only, thereby dropping the hallucination rate to a safe 6.0%. This is very close to the DistillBERT (purely extractive model) score.

The hybrid two-stage (DistillBERT-BART) model performed the best for text summarization. It achieved a stable balance between factual accuracy and semantic quality. While the DistillBERT extractive model excel at eliminating hallucination and the BART abstractive model offer superiority in fluency, the hybrid two-stage framework successfully combined the strengths of both paradigms while mitigating their individual flaws.

Visual Performance

Figure 3 displayed the visual representation of the performance metrics of the three distinct text summarization paradigms: Purely Extractive (DistillBERT), Purely Abstractive (BART) and a Hybrid Two-Stage (DistillBERT-BART) models. Evaluation metrics used include both token-matching heuristics which are ROUGE and BLEU, Semantic embedding alignments which is the BERTScore and the linguistic evaluation profiles which are fluency and hallucination rate. The empirical data demonstrates that while purely extractive models mitigate factual hallucinations and abstractive architectures offer smooth vocabulary synthesis, the hybrid two-stage approach provided a more superior summarization, maximizing semantic precision and structural fluency while suppressing text generation errors.

The purely extractive DistillBERT is represented by the green line, the purely abstractive BART-base is represented by the blue line while the two-stage hybrid DistillBERT-BART is represented by the yellow line. The performance of each summarization pathway is

measured across five distinct evaluation metrics, broadly grouped into two categories, which are the quantitative metrics: ROUGE and BLEU metrics, they measured the quantity of summaries captured from the source document and the number of summaries that are factually correct. The other

category which is the qualitative metrics, measured the quality of summaries generated. They checked the semantic quality, the readability flow and the presence of fabrications in the generated summaries.

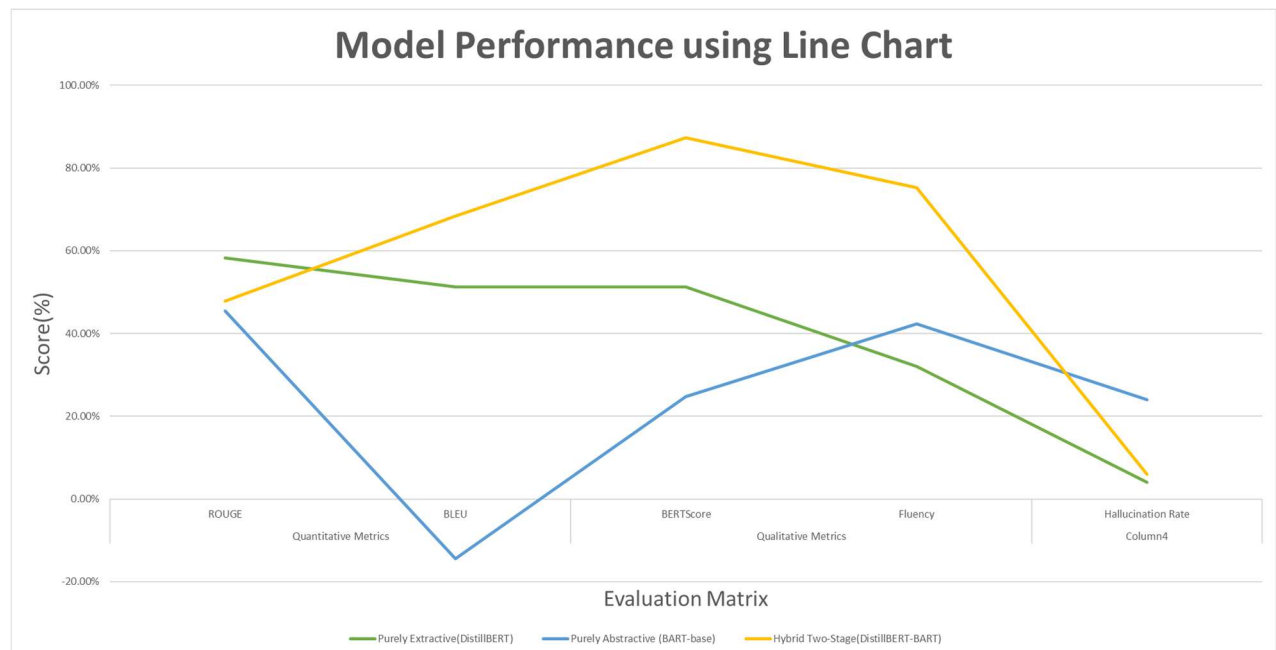


Figure 3: Visualization of the Model Performance using bar chart

The two-stage hybrid model performed the best among the three models. This was possible because it was able to successfully combine DistillBERT and BART to achieve outstanding results across all metrics. It has the highest score in BLEU, BERTScore and fluency, while maintain a very low hallucination rate. DistillBERT remains highly consistent across ROUGE, BLEU and BERTScore. Though, it has the lowest hallucination rate. its performance dropped significantly when evaluated for fluency, which is the human-readable check. The BART-base mode is highly unstable, it struggles heavily with standard the word overlaps matching check BLEU, yet it produces highly human-readable summaries with its high fluency score. Another major drawback highlighted here is a high risk of generating incorrect text, as shown by its very high score in hallucination rate.

The performance evaluation of text summarization approaches showed that the

DistillBERT-BART model consistently delivered the most robust performance, particularly within qualitative metrics. It achieved a peak performance of approximately 88% in BERTScore and 75% in fluency score, while successfully maintaining a low hallucination rate of roughly 6%. In contrast, the pure DistillBERT model exhibited high stability across quantitative benchmarks, securing the highest ROUGE score at 58% and the lowest overall error rate with a 4% hallucination rate. However, its performance noticeably dropped to 32% when evaluated for readability check fluency. Conversely, the pure BART-base model showed that there was a substantial instability by dropping to a very low score of -15% in BLEU score and equally had the highest error present with a 24% hallucination rate despite having a higher score of 42% when compared with the direct extractive model. Ultimately, these findings indicated that integrating extractive and abstractive methodologies in a hybrid pipeline successfully

leverages the factual precision of extractive models alongside the logical coherence and linguistic readability of abstractive architectures.

Conclusion

In conclusion, this paper demonstrated that balancing factually accurate summaries and logically coherent and readability flow in clinical text summarization does not require scaling up to using LLMs or expensive models that requires financial burden of cloud computing and infrastructure. Our proposed two-stage DistillBERT-BART pipeline proved that a single architecture can extract, rewrite and paraphrase critical clinical notes with a significantly smaller amount of computer resources compared to standard or older resources. By using an extractive stage to remove noise and irrelevances, identify and select salient sentences in the original texts, and using an abstractive stage to paraphrase and rewrite new sentences, this two-stage summarization framework effectively overcomes the dual challenges of disjointed and unreadable summaries, and the presence of high risk of factual hallucinations in the generated output. Restricting the abstractive rewriting to only factually checked summaries from the extractive selection offered a stable and usable automated text summarization system for clinicians and medical experts. Ultimately, this research has provided a scalable solution for healthcare systems, proving that smaller, lightweight models can deliver dependable summarization under strict operational constraints.

References

- [1] El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165, 113679.
- [2] DeYoung, J., Beltagy, I., van Zuylen, M., Kuehl, B. and Wang, L. L. (2021) 'MS2: Multi-Document Summarization of Medical Studies', *arXiv preprint arXiv:2104.06486*.
- [3] Zheng, W., Chen, X., & Liu, Y. (2025). Benchmarking retrieval-augmented generation: Factuality, speed, and infrastructure trade-offs. *IEEE Transactions on Knowledge and Data Engineering*, 37(2), 845–859.
- [4] Huang, Z., Chen, X., Wang, Y., Huang, J., & Zhao, X. (2025). A survey on biomedical automatic text summarization with large language models. *Information Processing & Management*, 62(5), 104216.
- [5] Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., and Wu, H. (2019). ERNIE: Enhanced Representation through Knowledge Integration. *arXiv preprint arXiv:1904.09223*.
- [6] Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J. B., Aali, A., Bluethgen, C., ... & Chaudhari, A. S. (2024). Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4), 1134-1142.
- [7] Lin, C., Miller, T., Dligach, D., Bethard, S. and Savova, G. (2019) 'Scaling clinical natural language processing with structured data', *npj Digital Medicine*, 2(1), pp. 1–5.
- [8] Zhang, J., Zhao, Y., Saleh, M. and Liu, P.J. (2020) 'PEGASUS: pre-training with extracted gap-sentences for abstractive summarization', *Proceedings of the 37th International Conference on Machine Learning (ICML '20)*, pp. 11328–113
- [9] Islam, M. M., Muhammad, U., & Oussalah, M. (2024). Evaluating text summarization techniques and factual consistency with language models. In *2024 IEEE International Conference on Big Data (BigData)* (pp. 116-122). IEEE.
- [10] Denecke, K., May, R., and Rivera-Romero, O. (2024). Transformer models in healthcare: a survey and thematic analysis of potentials, shortcomings and risks. *Journal of medical systems*, 48(1), 23.
- [11] Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L. and Liu, H. (2018) 'A clinical text classification paradigm using weak supervision and deep representation', *BMC Medical Informatics and Decision Making*, 18(1), p. 36.
- [12] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. and Zettlemoyer, L. (2020) 'BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension', *Proceedings of ACL 2020*, pp. 7871–7880
- [13] Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1906-1919).
- [14] Rupa, M. C., and Ramani, K. (2025). Hybrid Approaches for Advanced Medical Text Summarization: Combining TF-IDF, BERT, and Seq2Seq Models. *Advance Sustainable Science Engineering and Technology*, 7(3), 0250301-0250301.
- [15] Sztabiński, M., & Weichbroth, P. (2026). Deep Learning in Medical Speech to Text: Methods and Challenges. *Symmetry-Basel*, 18, 885.
- [16] Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., and Kaiser, L. (2019). *Text Summarization with Pretrained Encoders*. *arXiv preprint arXiv:1908.08345*.

- [17] Yadav, S., Gupta, D. and Demner-Fushman, D. (2022) 'CHQ-Summ: A dataset for consumer healthcare question summarization', *arXiv preprint arXiv:2206.06581*.
- [18] Nijor, S., Rallis, G., Lad, N., & Gokcen, E. (2022). Patient safety issues from information overload in electronic medical records. *Journal of patient safety*, 18(6), e999.
- [19] Asgari, E., Kaur, J., Nuredini, G., Balloch, J., Taylor, A. M., Sebire, N., ... & Pimenta, D. (2024). Impact of electronic health record use on cognitive load and burnout among clinicians: narrative review. *JMIR Medical Informatics*, 12, e55499.
- [20] Abdulmgid Mrwy Alenezi, Ganiah Alnasmi T Alenezi, Intisar Taleb Nasser Alanazi, Fidaa Saleh Alanazi, Nawader Qasem Fahaad Alruwaili, Reem Shannan Khazin Alanazi, ... Najwa Owaidh Aljohani. (2024). Nursing Management of Cognitive Overload in High-Acuity Clinical Environments. *International Journal of Computational and Experimental Science and Engineering*, 10(4).
- [21] Rohil, M. K., and Magotra, V. (2022). An exploratory study of automatic text summarization in biomedical and healthcare domain. *Healthcare Analytics*, 2, 100058.
- [22] Alambo, A., Banerjee, T., Thirunarayan, K., & Cajita, M. (2022). Improving the factual accuracy of abstractive clinical text summarization using multi-objective optimisation. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1615-1618). IEEE.
- [23] Helwan, A., Azar, D., & Ozsahin, D. U. (2023). Medical reports summarization using text-to-text transformer. In *2023 Advances in Science and Engineering Technology International Conferences (ASET)* (pp. 01-04). IEEE.
- [24] Rohanian, O., Nouriborji, M., Jauncey, H., Kouchaki, S., Nooralahzadeh, F., Clifton, L., ... and ISARIC Clinical Characterisation Group. (2024). Lightweight transformers for clinical natural language processing. *Natural language engineering*, 30(5), 887-914.
- [25] Jayatilleke, N., and Weerasinghe, R. (2025). A Hybrid Architecture with Efficient Fine Tuning for Abstractive Patent Document Summarization. In *2025 International Research Conference on Smart Computing and Systems Engineering (SCSE)* (pp. 1-6). IEEE.
- [26] Bednarczyk, L., Reichenpfader, D., Gaudet-Blavignac, C., Ette, A.K., Zaghir, J., Zheng, Y., Bensahla, A., Bjelogrljic, M. and Lovis, C. (2025) 'Scientific evidence for clinical text summarization using large language models: scoping review', *Journal of Medical Internet Research*, 27, e68998.
- [27] Rauf, M., and Papay, S. (2026). Medical Summarization in Practice: Design, Deployment, and Analysis of a Clinical Summarization System for a German Hospital. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)* (pp. 455-466).
- [28] Suleiman, D., & Awajan, A. (2020). Deep learning based abstractive text summarization: Approaches, datasets, evaluation metrics, and challenges. *IEEE Access*, 8, 109983–110001. doi.org
- [29] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- [30] Devlin, J., Chang, Ma., W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', *NAACL 2019*, pp. 4171–4186.
- [31] Liu, Y. and Lapata, M. (2019) 'Text summarization with pretrained encoders', *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3730–3740.
- [32] See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1073–1083). Association for Computational Linguistics. aclanthology.org
- [33] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, J., Xu, B., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 1–38. doi.org
- [34] Wang, Y., Chen, X., & Zhang, L. (2024). Controlling generative risks: Grounded text summarization via automated hallucination auditing. *Journal of Data and Information Quality*, 16(2), 112–131.