**International Journal of Applied Mathematics**
**Volume 38 No. 1s , 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

# EXPLORING ABSTRACTIVE SUMMARIZATION OF PRE-TRAINED MODELS: A STUDY ON GPT-2, T5, PEGASUS AND BART

**Punam Virendra Khandar[1], Dipak Kumar Mohanty[2], Santos Kumar Baliarsingh[3] and Vishnu Vardhan Budati[4]**

[1]School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubneswar, Odisha

[2]Government B.Ed. Training College Kalinga, Kandhamal, Odisha

[3]School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubneswar, Odisha

[4]Department of Computer Science and Applications, RBU, Nagpur, Maharashtra

[1]2181064@kiit.ac.in, [2]dkmohanty.iitkgp@gmail.com, [3]santos.baliarsinghfcs@kiit.ac.in and [4]budativv@rknec.edu

**Abstract**

Text summarization, a significant challenge within Natural Language Processing (NLP), purposes to distill huge volumes of content into brief, coherent summaries. As the amount of text data continues to expand, the need for efficient and accurate summarization methods has grown, making it a critical task across various sectors. Despite advancements in summarization, especially with the emergence of transfer learning that leverage the capabilities of pre-trained models, questions remain about which models perform best for summarization on specific datasets. This study evaluates how well the pre-trained models—GPT-2, T5, Pegasus, and BART perform for the abstractive text summarization task. We employ the datasets MultiNews, WikiSum, DUC & CNN/Daily Mail, consisting of news related articles accompanied by their human-generated summaries. In the proposed work, we fine-tuned each model on these datasets through transfer learning, by carefully adjusting the parameters. The metrics employed to assess the performance of each model are ROUGE and METEOR. After rigorous experiments, the results indicate that the T5 model beats the others in abstractive summarization on these datasets, achieving superior ROUGE scores of R1, R2 and RL equal to 63.04%, 42.22%, 47.82% respectively and an average precision of 90.10%. Lastly, the further exploration of future research directions is provided.

***Index Terms:*** *Abstractive Text Summarization, Transformer model, Transfer learning, ROUGE.*

## I.INTRODUCTION

The exponential growth in digital textual data has posed a significant challenge in extracting meaningful insights efficiently. Effective text summarization, one of the vital activities in Natural Language Processing (NLP), poses significant challenges due to the need to maintain a balance between accuracy, coherence, and readability. Traditional methods, including extractive and abstractive approaches, often fall short, particularly in generating summaries

Received: August 14 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

that closelymimic human-produced content. Additionally, the limitations of earlier models, such as long short-term memory networks (LSTMs) as well as recurrent neural networks (RNNs), have constrained the scalability and effectiveness of summarization tasks. In the field of NLP, the emergence of the Transformer architecture has marked a significant advancement, improving both the performance and scalability of text summarization models. Starting with the use of encoder-decoder models in 2015, deep learning methods have been effectively appliedto abstractive summarization, resulting in more coherent and contextually relevant summaries. Transformer-based simulation models, viz GPT-2, PEGASUS, T5 and BART, have been fine-tuned on large datasets like MultiNews, XSum, leveraging self-attention mechanisms to generate high-quality summaries. These models, supported by transfer learning and extensive pre-training, have set new benchmarks in the field, offering a broad understanding of language that enhances their summarization capabilities. Despite these advancements, several challenges remain unresolved. Transformer-based models, while effective, sometimes generate summaries lacking the nuance and coherence found in human-generated content. The fine- tuning process, although beneficial, can be resource-intensive and may not generalize well across various text genres or domains. Moreover, available assessment metrics like ROUGE may not completely identify the qualitative nuances of summary content, potentially causing the discrepancies between model performance and human judgment. There is a pressing need for further exploration of hybrid approaches that harness the capabilities of both extractive and abstractive methods to generate summaries that are accurate, natural-sounding, and contextually appropriate. This paper aims to investigate and evaluate the use of Transformer-based models for abstractive text summarization applied to the datasets CNN/Daily Mail(version 3.0.0), MultiNews, WikiSum, DUC. By fine-tuning pre-trained models such as T5, GPT-2, PEGASUS, and BART, this study seeks to measure their effectiveness in generating high-quality summaries. Additionally, the paper examines the current limitations of the models. Further, it proposes probable avenues for future work to advance text summarization in NLP.

## II. LITERATURE REVIEW

The advancements in deep learning have significantly influenced research in abstractive content summarization [1, 2, 3]. In [2], the authors developed a system for abstractive content summarization using Convolutional Neural Networks (CNN) accompanied by Long Short-Term Memory (LSTM) networks.They utilized Multiple Order Semantic Parsing model to identify key sentences from the source material, subsequently utilizing deep learning techniques to generate summaries of text, the CNN/Daily Mail as well as Gigaword datasets were used for comparing the performance of different techniques. In a related study [4], researchers examined the effect of local attention on Long Short-Term Memory models to generate text summaries based on abstractive approach. This study employed the GloVe dataset and Amazon Fine Food Review dataset. Experiments revealed that although general attention-based techniques resulted in higher ROUGE-1 scores due to longer summaries, local attention-based framework exhibited superior scores of ROUGE-2 by producing adequate word pairs that aligned with provided reference summaries.

Received: August 14 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Vaswani et al.(2017) [5] introduced a Transformer Architecture shown in figure 1. It magnificently reformed the Natural Language Processing field. It replaced the classic Long Short-Term Memory networks and Recurrent Neural Networks with a more efficient and scalable architecture that relies heavily on mechanisms of self-attention.

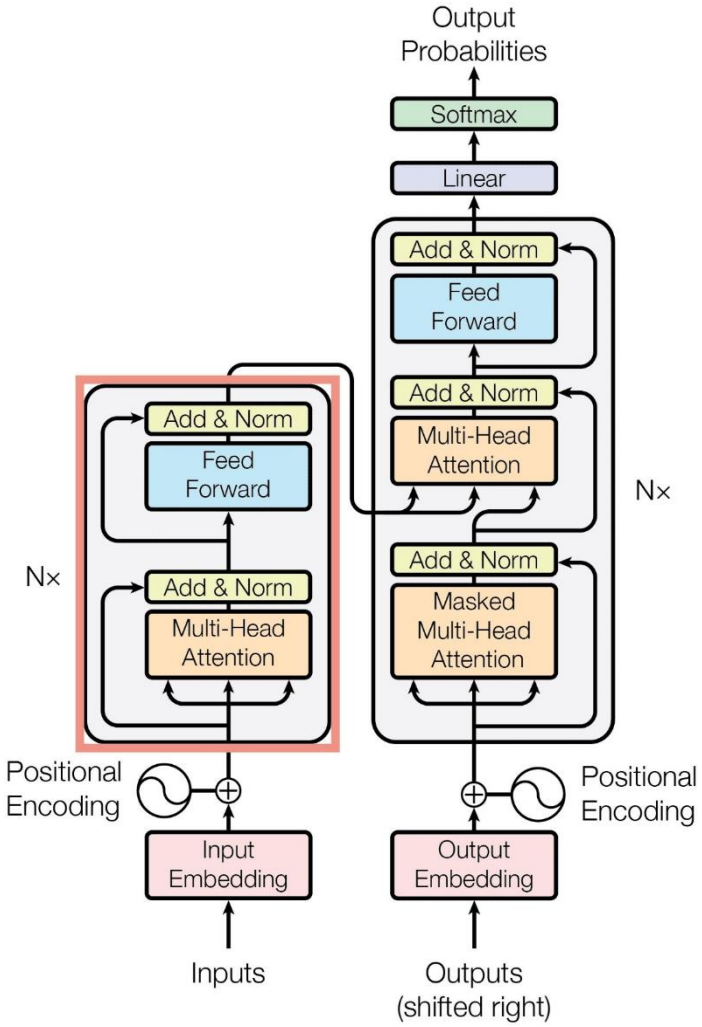


**Fig. 1:** The Architecture of the Transformer model [5]

Madhuri P. Karnik et al. [6] proposed Transformer models utilizing the PGN (Pointer-Generation Network) and coverage mechanism, subsequently developing the Fastformer-architecture based on these concepts. The Fastformer model, characterized by its linear complexity, outperformed another similar techniques. The dataset used for the conduction of experiments was CNN/Daily Mail. The pointer-generator concept, introduced in 2017, has since gained popularity in the NLP domain [7,8].

In [9], the authors presented PEGASUS, a language model built on the Transformer architecture that employs an encoder-decoder approach. Specifically designed for large text datasets with a novel self-supervised objective, PEGASUS demonstrated superior performance in resource-limited summarization tasks, achieving results comparable to human excellence. Similarly, BART, a model employing the sequence- to-sequence concept as a denoising autoencoder, was proposedby Mike Lewis et al. [10]. The BERTSum model has been adapted

Received: August 14 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

for conversational language by the authors in [8], who applied abstractive summarization to spoken tutorial videos across a wide range of topics. They utilized the ROUGE metric and the Content-F1 metric, with manual evaluators grading a randomly selected set of abstracts from the HowTo100M and YouTube datasets.

Recently, researchers have focused on addressing a major issue in abstractive text summarization (ATS): factual error correction [11,12]. Jianfeng et al. [11] proposed a scoring method called SiCF, which evaluates the quality of a summarization model across three main aspects: Faithfulness (factual precision), Coverage (factual recall), and Semantic invariance (model confidence indicator). This score is used to select a subset of unclassified conversations with highly-constructed summaries for model training. To address a conditional- generation cloze problem, the authors in [12] proposed a new factual error correction model called FactCloze. FactCloze assesses the relationships between factual components and determines whether gaps can be accurately filled.

In [13], the authors emphasized on categorizing resume text. Using a resume dataset, they assessed a number of methods, such as LSTM, pre-trained models, as well as case-based fine-tuned models. Fine-tuning the BART-Large model yielded particularly strong results using the resume dataset. As another dimension in text summarization, aimed at capturing the most essential information while omit- ting less critical details, several authors [14,15,16] explored the concept of soft guidance for salience allocation. Fei Wang et al. [14,15] introduced SEASON, a novel summarization technique offering adaptable and reliable salience guidance. In[16], the authors proposed SDACL, a sentence-level abstraction method for abstractive text summarization, introduced by Ying Huang et al. The research also proposed the SSCL model. The study's results indicate that the proposed approach may lead to superior outcomes in contrastive learning methods and a significant improvement over initial result.

## III. METHODOLOGY

In our study, we assess the effectiveness of pre-trained language frameworks for the task of summarization of abstractive text. The approach involves several stages, including preprocessing the text data, applying tokenization specific to each model, and fine-tuning the models on the chosen dataset. Additionally, we conduct a quantitative evaluation of the models' performance.

### *A. Pre-trained Models*

*1)* ***BART:*** It is short for Bidirectional and Auto-Regressive Transformers. It is recognized as an impactful framework for translation as well as summarization tasks in NLP. This Transformer style architecture is particularly proficient at addressing linguistic challenges with high accuracy. Leveraging its bidirectional and auto-regressive features, BART has proven to be highly effective in tasks such as summarization and translation [17,18,11]. Its bidirectional nature enables it to fully comprehend textual content, efficiently capturing context, while its auto-regressive capability empowers it to produce coherent as well as fluid summaries for the provided input.

*2)* ***PEGASUS:*** It is a framework built on unique self-supervised pre-training objective

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

called Gap-Sentence Generation (GSG). To master language fundamentals and enhance its ability to understand and summarizecomplex textual content, it follows a two-phase learning process: initial training and fine-tuning [12]. This approach excels at condensing input texts into brief,logical, as well as insightful summaries [18,19].

3) ***GPT-2:*** This model is a notable advancement in natural language processing, bringing us closer to creating machines capable of understanding and interacting with human-like language patterns [22]. GPT-2 has gained widespread recognitionas a leading framework in NLP, due to its exceptional performance in language-related tasks and its ability to generate coherent sentences effortlessly [23]. These qualities have madeGPT-2 a preferred choice among academic researchers and industry professionals alike.

4) ***T5:*** It is a Text-To-Text Transfer Transformer framework. A model designed and developed by Google Research, has had a notable impact on natural language processing field [13]. This innovative model supports ample diverse NLP activities, viz summarization, question answering, text classification, as well as content translation, unlike traditional models that are typically designed for specific tasks. By employing a "text-to-text" framework, T5 streamlines multitasking by converting different tasks into text-based formats. With thorough pre-training on huge datasets as well as fine-tuning on task-specific corpora, T5 achieves remarkable performance across numerous applications [4,20,21].

Here, we have given in brief the various versions of these models and the strength of each version. Table I summarizes it.

**Table I:** Comparison of Versions of Pre-Trained Models

| Model | Version | Parameters | Model Size | Strength |
|---|---|---|---|---|
| **BART** | bart-base | 139M | ~500 MB | Strong baseline for text generation and summarization. |
| | bart-large | 406M | ~1.6 GB | Best for abstractive summarization, high-quality generation. |
| | bart-large-cnn | 406M | ~1.6 GB | Fine-tuned for summarization using CNN/ Daily Mail. |
| | bart-large-xsum | 406M | ~1.6 GB | Fine-tuned for abstractive summarization on XSum. |
| | bart-large-mnli | 406M | ~1.6 GB | Fine-tuned for natural language inference (NLI). |
| **Pegasus** | pegasus-base | ~50M | ~200 MB | General-purpose summarization for small tasks. |
| | pegasus-large | 568M | ~2.1 GB | Robust for summarization tasks across domains. |

Received: August 14 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

| | | | | |
|---|---|---|---|---|
| | pegasus-cnn_dailymail | 568M | ~2.1 GB | Pretrained on CNN/Daily Mail dataset for summarization. |
| | pegasus-xsum | 568M | ~2.1 GB | Pretrained for abstractive summarization (XSum dataset). |
| **T5** | t5-small | 60M | ~220 MB | Suitable for small-scale tasks, fast inference. |
| | t5-base | 220M | ~900 MB | Balanced performance and speed. |
| | t5-large | 770M | ~3.1 GB | More expressive, requires higher computational resources. |
| | t5-3b | 3B | ~12 GB | Excellent for complex NLP tasks, needs strong GPUs/TPUs. |
| | t5-11b | 11B | ~45 GB | State-of-the-art performance, very resource-intensive. |
| **GPT (OpenAI)** | GPT-2 Small | 117M | ~500 MB | Fast inference, basic text generation. |
| | GPT-2 Medium | 345M | ~1.5 GB | Balanced size, good for creative tasks. |
| | GPT-2 Large | 774M | ~3.1 GB | Improved coherence for long-text generation. |
| | GPT-2 XL | 1.5B | ~6 GB | High-quality text generation, computationally intensive. |
| | GPT-3 Ada | 350M | Proprietary | Fast and cost-effective. |
| | GPT-3 Curie | 6.7B | Proprietary | Balanced cost vs. quality. |
| | GPT-3 Davinci | 175B | Proprietary | Best performance, state-of-the-art generation. |

The peculiarity of the transformer Models is that they utilize self-attention mechanisms calculated [5] as follows.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})\, V$$

….. (1)

where, Q represents query,

K represents key,

V represents value matrices resulting from the input data,

$d_k$ - dimension vector.

### *B. Experimental Details*

This section describes the extensive experimental setup designed to assess the effectiveness of several cutting-edge models in multi-document abstractive summarization. Our goal is to use standardized datasets to guarantee a thorough, repeatable, and equitable comparison across all models.

Received: August 14 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

*1)* ***Dataset:*** A variety of sources, including scholarly publications, news pieces, and online records, provided the data for this investigation. To evaluate the resilience of the summary models, a wide range of subjects and writing styles were covered. The CNN/Daily Mail dataset (version 3.0.0), MultiNews, WikiSum, and DUC were the datasets used in this study. They are described as below.

i. ***CNN/Daily Mail dataset:*** This open-source dataset, which is easily accessible through the Hugging Face Datasets library, is commonly utilized in the areas of machine learning along with natural language processing, especially for text summarization tasks. The dataset consists of a collection of news articles (Article), each paired with a multi-sentence summary (Highlights). Itcontains over 200 million words, including both extensivearticles covering a broad range of topics and shorter snippets that highlight key ideas [25]. For this study, selected articles from the dataset are used for training, validation, and testing.

ii. ***MultiNews:*** The MultiNews dataset, which includes over 56,000 news stories and expertly crafted summaries, is a comprehensive resource for multi-document abstractive summarization. Because it covers a wide range of subjects from many sources, summarizing jobs are guaranteed to be robust. The MultiNews dataset offers superior annotations that improve readability and coherence, in contrast to many datasets with much brief summaries. It is useful for study on temporal summarization since it efficiently captures the temporal and evolutionary components of news, with an average input size of 2,100 words accompanied by a summary length of 260 words.

iii. ***WikiSum:*** WikiSum is a comprehensive dataset for abstractive summary of various documents, where source content is provided by numerous reference documents and Wikipedia lead sections function as summaries. About 1.5 million Wikipedia articles are included, and each summary is based on an average of over ten sources. For effective processing, retrieval-based techniques are frequently necessary due to the unusually large input length.

iv. ***DUC:*** The NIST-created DUC datasets, which include excellent, human-written summaries, are frequently used benchmarks for multi-document summarization. Every year, they publish 500–1,000 news stories, with each summary drawn from 10–50 linked papers. While DUC 2005–2007 added query-focused and update summarization tasks, DUC 2001–2004 concentrated on general summarizing.

*2)* ***Preprocessing:*** Preprocessing is essential to guaranteeing that the dataset is in the proper format and prepared for training, regardless of the selected model. For this purpose, tokenization techniques are employed to divide text into tiny elements, like words, subwords, characters. They are subsequently converted into corresponding IDs.

***Preprocessing Steps:***

i.**Cleaning:** Any non-textual components, such as pictures, special characters, and erratic spacing, are eliminated from text data.

ii. **Normalization:** Normalization involves standardizing date and numeric formats and converting texts to lowercase.

Received: August 14 2025 

**International Journal of Applied Mathematics**
**Volume 38 No. 1s , 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

iii. **Tokenization:** Tokenizing words and sentences makes it easier to process language further.

Preprocessed Data = Tokenize(Clean(D))
…..(2)
where, D represents each document in the dataset,
Clean involves removing headers, footers, and special characters, and
Tokenize converts texts into tokens suitable for model inputs.

iv. **Stop-word Removal:** To highlight the more significant information in the text, common stop words are eliminated.

v. **Stemming and Lemmatization:** Words are broken down to their most basic forms to combine many variations of the same word.

*3) **Evaluation metrics:*** In this section, the pre-trained models mentioned above—GPT-2, T5, Pegasus, and BART—are evaluated for their text summarization capabilities. The summaries generated by above mentioned models are evaluated with the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score. It's a well-known metric for summarization tasks. ROUGE is a statistical measure that assesses token frequencies in both the source and target texts. It is widely used by researchers to gauge the coherence of algorithm-generated summaries by measuring the number of matching words with the original text.

ROUGE measures the effectiveness of a text summarization algorithm using N-grams. Specifically, ROUGE-N calculates the N-gram intersection between the output generated by the model and the text available for reference. ROUGE-1 examines the similarity between unigrams (individual words) in the model's output and the source text, while ROUGE-2 evaluates the similarity of bigrams. ROUGE-L is determined based on the longest matching sub-sequences, comparing the output of the model to the text available for reference at the sentence level. ROUGE-Lsum, similar to ROUGE-L, applies a recall penalty to excessively long summaries. It is evaluated as follows.

$$ROUGE - N = \frac{\sum_{s \in Reference\ Summaries} \sum_{\text{gram}_n \in s} Count_{match}(gram_n)}{\sum_{s \in Reference\ Summaries} \sum_{\text{gram}_n \in s} Count(gram_n)}$$

….. (3)

where, N → the length of the n-gram (e.g., ROUGE-1 for unigrams, ROUGE-2 for bigrams).

s → a reference summary (there may be multiple human-written references).

$gram_n$ → an n-gram in the reference summary.

$Count_{match}(gram_n)$ → number of overlapping n-grams between the *candidate summary* and *reference*

*summaries*.

$Count(gram_n)$ → total number of n-grams in the reference summaries.

Similarly, another metric, the METEOR score (Metric for Evaluation of Translation with Explicit Ordering) is employed. This metric is intended to evaluate the quality of the text

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

generated by the models. It provides a harmonic mean of precision and recall, considering the synonymy and paraphrase.

*4)* ***Implementation Details:*** The specific versions of the pre-trained models used in our study are Google/PEGASUS- CNN-DailyMail, GPT2-Medium, Facebook/BART-LARGE-CNN, and T5-Small. We optimized each framework for 20 epochs setting the batch size of 32 to obtain useful results when comparing Google/PEGASUS-CNN-DailyMail, GPT2-Medium, Facebook/BART-LARGE-CNN, and T5-Small on the selected datasets as mentioned above. We employed gradient clipping with a maximum norm of 1.0, and optimized with AdamW. We made a slow start with learning rate of 1e-5 and the loss was monitored. It helped model to converge slowly and the rate was increased till 3e-5 to achieve optimum training. To stabilize training, a linear or cosine scheduler was used, with 500 warmup steps. To guarantee the best possible summarization quality, cross-entropy loss L(θ) was employed [5].

$$L(\theta) = -\sum_{(x,y)\in D} log\ p(y \mid x; \theta) \quad \text{….. (4)}$$

where, L(θ) – Loss function,
p – probability function,
D - training dataset,
x - input,
y - target output,
θ – model parameters

Finally, the model effectiveness is evaluated by computing ROUGE-1, ROUGE-2, and ROUGE-L scores.

## IV. RESULT AND DISCUSSION

Table II shows the results generated by various models—Google/PEGASUS-CNN-DailyMail, GPT2-Medium, Facebook/BART-LARGE-CNN, and T5-Small—evaluated using the ROUGE metric. Based on Table II and Figure 2, it is clear that T5 outperforms the other models used in the present study, achieving the highest scores of 0.6304, 0.4222, and 0.4782 for ROUGE-1, ROUGE-2, and ROUGE-L, respectively, demonstrating superior performance. T5, in fact, has powerful properties like language understanding, generalization, robust architecture, etc. Additionally, it comes in various sizes (small, Base, Large, 3B, 11B), making it scalable based on computational resources as well as task complexity.

BART also demonstrates competitive scores of 0.56, 0.27, particularly- in the ROUGE-1 and ROUGE-2 metrics. In contrast, GPT-2 and PEGASUS show 0.43, 0.27, 0.32 and 0.29, 0.18, 0.17 scores respectively. These are relatively lower scores compared to T5 and BART, suggesting GPT-2 and PEGASUS may not perform much effectively in generating summaries for this specific evaluation. Besides, GPT-2 is comparatively better than BART in terms of ROUGE-L having scores of 0.32 and 0.24 respectively.

Based on the second metric used in the study, that is METEOR, which provides a harmonic mean of precision and recall, we computed average precision across all selected models on different datasets. Looking at the average precision mentioned in the Table-II below, T5 shows outstanding performance over all other models with the highest precision of 90.10%.

Received: August 14 2025

**International Journal of Applied Mathematics**
**Volume 38 No. 1s , 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

**Table II:** Rouge Scores of Different Text Summarization Models On Different Datasets

| Dataset | Model | Rouge1 | Rouge2 | RougeL | Average Precision |
|---|---|---|---|---|---|
| CNN/Daily Mail | GPT2 | 0.432836 | 0.272727 | 0.328358 | 87.90% |
| | T5 | 0.630435 | 0.422222 | 0.478261 | 90.10% |
| | BART | 0.567901 | 0.278481 | 0.246914 | 86.50% |
| | PEGASUS | 0.222222 | 0.176923 | 0.185185 | 85.00% |
| MultiNews | GPT2 | 0.423758 | 0.195768 | 0.298765 | 88.70% |
| | T5 | 0.624357 | 0.398756 | 0.426187 | 87.20% |
| | BART | 0.50901 | 0.248176 | 0.234521 | 89.40% |
| | PEGASUS | 0.231432 | 0.165473 | 0.178654 | 82.30% |
| WikiSum | GPT2 | 0.393687 | 0.199965 | 0.328877 | 83.90% |
| | T5 | 0.599876 | 0.416543 | 0.462621 | 84.50% |
| | BART | 0.51679 | 0.248145 | 0.214345 | 89.70% |
| | PEGASUS | 0.198976 | 0.228765 | 0.171754 | 80.40% |
| DUC | GPT2 | 0.368711 | 0.196534 | 0.312887 | 89.40% |
| | T5 | 0.509876 | 0.416543 | 0.452621 | 82.30% |
| | BART | 0.501979 | 0.347145 | 0.244245 | 83.90% |
| | PEGASUS | 0.296977 | 0.188765 | 0.175423 | 84.50% |

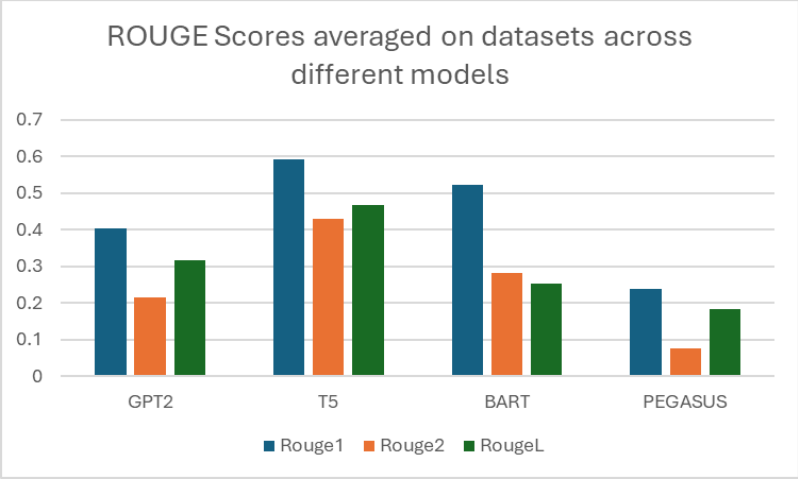


**Fig. 2:** ROUGE Scores averaged on datasets across different models

Received: August 14 2025

International Journal of Applied Mathematics
Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

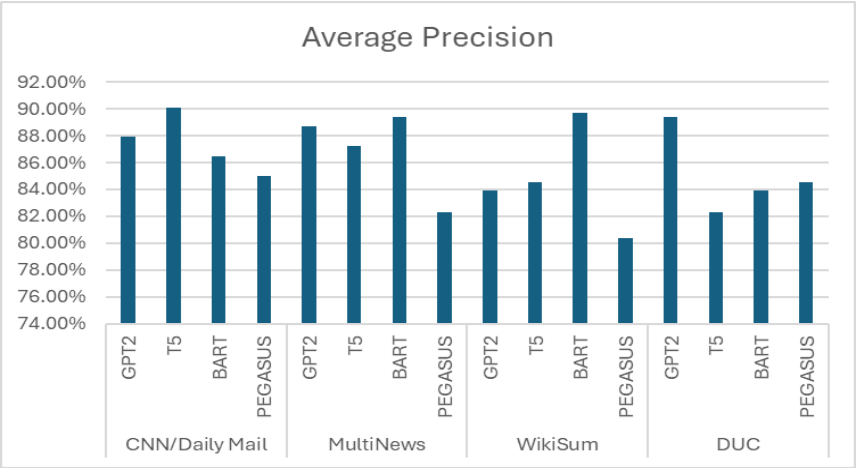


**Fig. 3** Precision Metrics across Models

## V. CONCLUSION AND FUTURE WORK

According to Dima Suleiman et al. [24], earlier studies reported that text summarization using a pre-trained encoder model achieved ROUGE-1, ROUGE-2, and ROUGE-L scores of 43.85%, 20.34%, and 39.90% respectively. Our results, using the datasets CNN/Daily Mail, WikiSum, DUC, MultiNews have exceeded these earlier benchmarks by producing the scores of 63.04%, 42.22%, 47.82% respectively as shown in Table II. Thus, based on these results, we conclude that T5 stands out as the top-performing model for the summarization task, outperforming previous SOTA models. BART also exhibits its strength, closely trailing T5. However, the suitability ofa model may depend on specific use cases and requirements beyond these metrics, including considerations of computational resources, fine-tuning efforts, and task-specific nuances. Future research should focus on enhancing hyper-parameter modifications and extending training durations for the models. It would be valuable to apply these models to larger and more diverse datasets and evaluate their performance on a wider range of text documents. Additionally, expanding the approach to encompass various text generation tasks, such as chatbot development and question answering, could provide further insights into the models' capabilities and applications.

## REFERENCES

[1] Hanunggul, P. M., & Suyanto, S. (2019, December). The impact of local attention in lstm for abstractive text summarization. In 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) (pp. 54-57). IEEE.

[2] Song, S., Huang, H., & Ruan, T. (2019). Abstractive text summarization using LSTM-CNN based deep learning. Multimedia Tools and Applications, 78(1), 857-875.

[3] Zolotareva, E., Tashu, T. M., & Horváth, T. (2020, September). Abstractive Text Summarization using Transfer Learning. In ITAT (pp. 75-80).

[4] Ranganathan, J., & Abuka, G. (2022, November). Text summarization using transformer model. In 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS) (pp. 1-5). IEEE.

Received: August 14 2025

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[5] Vaswani, A. (2017). "Attention is all you need", Advances in Neural Information Processing Systems, 31st Conference on Neural Information Processing Systems (NIPS 2017), CA, USA, 2017.

[6] Karnik, M. P., & Kodavade, D. V. (2023). Abstractive Summarization with Efficient Transformer Based Approach. International Journal on Recent and Innovation Trends in Computing and Communication, 11(4).

[7] Chen, Z., Xu, J., Liao, M., Xue, T., & He, K. (2022). Two-phase Multi-document Event Summarization on Core Event Graphs. Journal of Artificial Intelligence Research, 74, 1037-1057.

[8] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020, November). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International conference on machine learning (pp. 11328-11339). PMLR.

[9] Lewis, M. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

[10] Jianfeng He, Hang Su, Jason Cai, Igor Shalyminov, Hwanjun Song, Saab Mansour. "Semi-Supervised Dialogue Abstractive Summarization via High-Quality Pseudolabel Selection". arXiv preprint arXiv:2403.04073, 2024.

[11] He, J., Su, H., Cai, J., Shalyminov, I., Song, H., & Mansour, S. (2024). "Semi-supervised dialogue abstractive summarization via high-quality pseudolabel selection" arXiv preprint arXiv:2403.04073.

[12] OMercan, Ö. B., Cavsak, S. N., Deliahmetoglu, A., & Tanberk, S. (2023, October). Abstractive text summarization for resumes with cutting edge NLP transformers and LSTM. In 2023 Innovations in Intelligent Systems and Applications Conference, pp. 1-6, IEEE.

[13] Wang, F., Song, K., Zhang, H., Jin, L., Cho, S., Yao, W. & Yu, D. (2022). Salience allocation as guidance for abstractive summarization. arXiv preprint arXiv:2210.12330.

[14] Rehman, T., Bose, R., Dey, S., & Chattopadhyay, S. (2024). Analysis of Multidomain Abstractive Summarization Using Salience Allocation. arXiv preprint arXiv:2402.11955.

[15] Ying Huang, Zhixin Li, Zhenbin Chen, Canlong Zhang, Huifang M. "Sentence salience contrastive learning for abstractive text summariza- tion". 0925-2312/© 2024 Elsevier; Neurocomputing 593 (2024) 127808.

[16] Yadav, Hemant, Nehal Patel, and Dishank Jani. "Fine-Tuning BART for Abstractive Reviews Summarization." Computational Intelligence: Se- lect Proceedings of InCITe 2022. Singapore: Springer Nature Singapore, 2023. 375-385.

[17] Rehman, T., Das, S., Sanyal, D. K., & Chattopadhyay, S. (2022, August). An analysis of abstractive text summarization using pre-trained models. In Proceedings of International Conference on Computational Intelligence, Data Science and Cloud

Volume 38 No. 1s , 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Computing: IEM-ICDC 2021 (pp. 253-264). Singapore: Springer Nature Singapore.

[18] Lalitha, E., Ramani, K., Shahida, D., Deepak, E. V. S., Bindu, M. H., & Shaikshavali, D. (2023, May). Text summarization of medical documents using abstractive techniques. In 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 939-943). IEEE.

[19] Borah, M., Dadure, P., & Pakray, P. (2022). Comparative analysis of T5 model for abstractive text summarization on different datasets.

[20] Gupta, A., Chugh, D., Anjum, & Katarya, R. (2022). Automated news summarization using transformers. In Sustainable Advanced Computing: Select Proceedings of ICSAC 2021 (pp. 249-259). Singapore: Springer Singapore.

[21] Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., & Gadekallu, T. R. (2024). Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. IEEE Access.

[22] Liu, S., & Healey, C. G. (2023). Abstractive summarization of large document collections using gpt. arXiv preprint arXiv:2310.05690.

[23] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1-67.

[24] Suleiman, D., & Awajan, A. (2020). Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges. Mathematical problems in engineering, 2020(1), 9365340.

[25] See, A., & Manning, C. D. (2017). CNN/DailyMail: A large-scale dataset for abstractive text summarization [GitHub repository]. Retrieved from https://github.com/abisee/cnn-dailymail.

## APPENDICES

*Appendix A: Additional Figures and Charts*

The data and analyses offered in the major sections of the research study on multi-document abstractive summarization are supplemented by additional visual representations in this appendix. Deeper understanding of model performances, comparisons, and patterns noted throughout the study is provided by these figures and charts.

Received: August 14 2025 

**International Journal of Applied Mathematics**
**Volume 38 No. 1s , 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

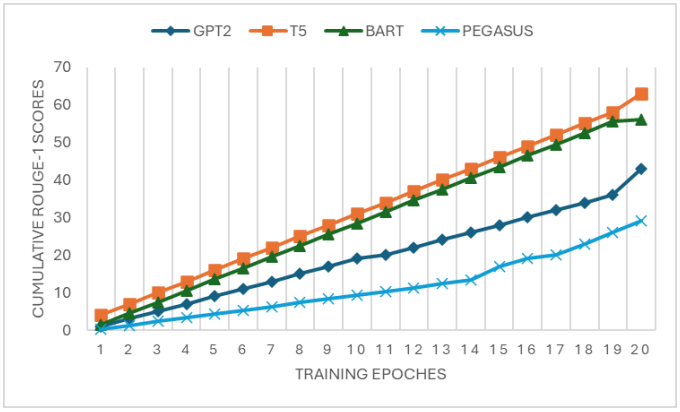


Fig. A1. Model Performance over Time

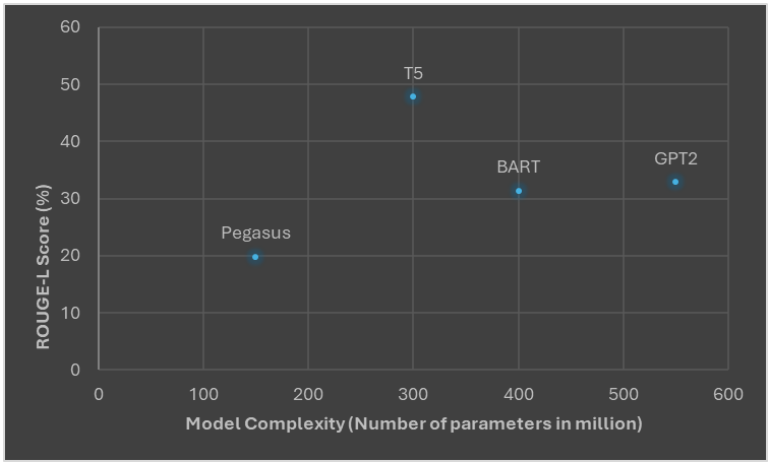


Fig. A2. Model Complexity Vs Performance

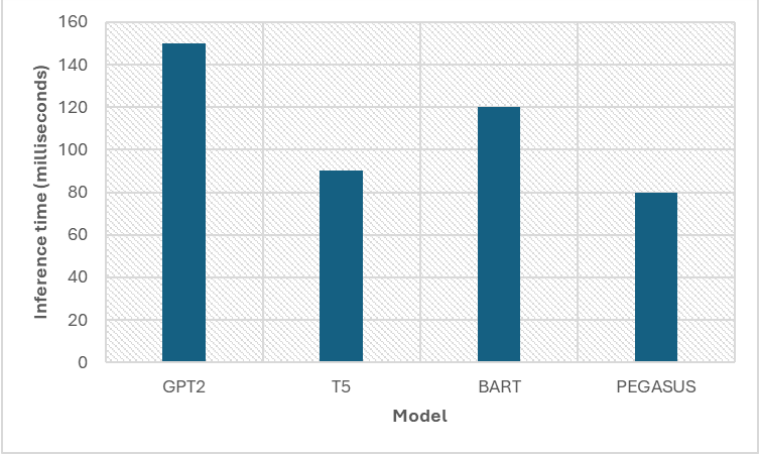


**Fig. A3.** Comparison of Inference Time across Models

Received: August 14 2025

**International Journal of Applied Mathematics**
**Volume 38 No. 1s , 2025**
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

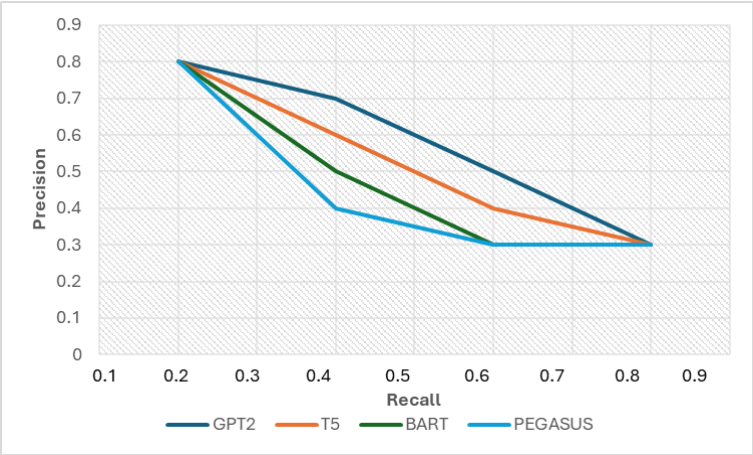


**Fig. A4.** Precision-Recall Curves for Each Model

Received: August 14 2025