Minimax Theory and its Applications
Volume 11 (2026), No. 1s, 273-288

# Domain-Adaptive Transformer Models for Enhanced Abstractive Text Summarization

Rooz Munjal[1*], Sonali Goyal[2]

[1*]Research Scholar Department of CSE, MMEC Maharishi Markandeshwar (Deemed to be University), Haryana, India. E-mail: roozmunjal.phd@gmail.com
[2]Associate Professor Department of CSE, MMEC Maharishi Markandeshwar (Deemed to be University), Haryana, India. Email: sonaligoyal@mmumullana.org

**Abstract**

Abstractive text summarization contributes to natural language processing (NLP) by reducing and abstracting long documents. Even though there is great improvement in transformer based pre-trained models, their performance now varies from domain to domain and requires specific fine-tuning. The aim of the research is to test the performance of pretrained models for abstractive dialogue and legal document summarization. The goal is to contrast the performance of three fine-tuned state-of-the-art models—BART, Pegasus and Falconsai (T5-small) on SAMSum (dialogue summarization) and BillSum (Legal document summarization). Pegasus achieves the highest performance on the SAMSum dialogue dataset (ROUGE-1: 42.76), while BART leads on the BillSum legal dataset (ROUGE-1: 43.26). On the SAMSum dialogue dataset, Falconsai struck a reasonable compromise with the ROUGE-1 F1 score 39.4 while Pegasus committed to ROUGE-2 best signifying its strong performance in conversation context. The fine-tuned models improve the quality of summarization beyond previous benchmarks, while a length-tracking mechanism and an efficient beam-search make inference faster. The paper contributes systematic model selection recommendations focusing on Pegasus for dialogue, BART for legal and Falconsai where resources are limited. We discovered persistent problems such as redundancy, hallucination, and factual inconsistency, suggesting hybrid architectures and domain specific pre-training for future work.

***Keywords***: Abstractive Text Summarization; Dialogue Summaries; Fine-Tuning; Legal Documents; Pre-trained Models; Transfer Learning

## Introduction

Automatic text summarization is essential for efficient information retrieval and knowledge extraction due to the exponential expansion of digital content[1]. Users struggle to extract useful insights from lengthy papers, especially in legal text analysis and conversational AI, due to the growing volume of unstructured data. Thus, high-quality, automated summarizing systems that generate brief, coherent, and semantically relevant summaries without manual inspection are in demand[2]. Two categories of text summarizing are abstractive and extractive ones. While abstractive techniques create new sentences that express the central idea, extractive techniques choose and concatenate salient key sentences from the source material. Extractive methods could lack coherence even if they are computationally efficient. Abstractive summarizing, on the other hand, allows more freedom yet calls for a deep awareness of semantic links [3]. Pre-trained transformer-based models have greatly enhanced AS systems even for variable length content[4], [5]. By using large-scale pre-training and fine-tuning on domain-specific datasets, BART, Pegasus, and Falconsai have achieved top performance on numerous summarization benchmarks. They work differently in different text domains, hence domain-specific modification is necessary for optimal performance[6], [7].
Most of the research concentrates on news summarizing or broad text domains, leaving domain-specific challenges like dialogue and legal document summary untouched, despite these models' increasing efficacy[8].

### Problem Statement and Research Motivation

Summarization in dialogue-based and legal text genres has a different set of challenges when compared

to typical news summarization tasks.
a) **Dialogue Summarization (SAMSum Dataset):** In multi-turn, informal conversations, models have to pick out the important information while still keeping track of who's saying what, how the conversation flows, and how earlier turns shape later ones. This is made harder by the fact that it's often difficult to distinguish meaningful content from filler.
b) **Legal Document Summarization (BillSum Dataset):** Models need to maintain both the legal accuracy and coverage of respective, massive structured legal texts (7). AS is hard to read in legal papers since its phrases are complex, it uses legal language and its structure.

Although BART, Pegasus, and other generators are becoming more and more popular, there are still three major areas that need more research:
a) **Limited Evaluation in Domain-Specific Summarization:** Most of the researches for summarizing news, not discourses or legal reports that require some linguistic changes[8], [9].
b) **Impact of Computational Resources and Hyperparameters:** GPU power and hyperparameter selections have a significant impact on fine-tuning performance, but previous research offers little insight into how these elements impact model performance[8], [9], [10].
c) **Lack of Systematic Model Comparison:** There are many transformer models, but no one knows which ones work best in different areas because not many studies have carefully tested how well they work on different text structures [11].

This study enhances and analyzes three prominent transformer models—BART, Pegasus, and Falconsai (T5-small)—utilizing two specialized datasets: SAMSum (conversation summary) and BillSum. The principal aims:
a) Fine-tuned BART, Pegasus and Falconsai evaluation for abstractive summarizing on multiple text contexts
b) Hyperparameter tweaking and GPU resource (NVIDIA A100) analysis can be used to evaluate the performance in comparison to baseline research.
c) Content abstraction, coherence, and compression efficiency serve as metrics for the strengths and constraints of models;

**The key contributions are:**
a) The present ROUGE scores surpassed the reference research, suggesting the impact of domain adaptation on the quality of summarizing.
b) Studied the impact of fine-tuning approaches and GPU resources on generalization and coherence, demonstrating that ideal settings boost both.
c) With model selection rules, pointing out Pegasus for casual interaction, BART for legal text summarizing and Falsconsai for resource limited scenarios.
d) Content redundancy, hallucination and fact discrepancy problems were recognized, and future studies should consider hybrid models and domain-specific pretraining.
The paper is structured as follows. Related work is discussed in Section 2, the approach involving datasets, models and fine-tuning is presented in Section 3. Results, analysis and comparisons are presented in section 4. Section 5 finishes the work.

## Literature Review

CNN-dailymail, SAMSum, and BillSum tested three pre-trained models for abstractive text summarization: google/pegasus-cnn-dailymail, Falconsai, and facebook/bartlarge-cnn in[13], [14]. These models summarised 2000 test instances from each dataset to compare ROUGE and BLEU performance. Google/pegasus-cnn-dailymail outperformed the Falconsai model on CNN-dailymail and BillSum, but facebook/bartlarge-cnn fared identically. The top models in these datasets have higher ROUGE-1, ROUGE-2, and ROUGE-L f-scores. They advised fine-tuning pre-trained models on specific datasets when interpreting findings. This study analyzes data models and evaluates their performance to inform text summarizing implementations.
Context utilization and position bias in AS are examined in using six Large Language Models (LLMs) and 10 datasets. New evaluation benchmark MiddleSum handles significant information in the middle of the context. LLMs have a" middle-curse," underusing central information and resulting in inconsistent summarization, according to the study. Hierarchical summarizing and incremental summarizing change the summary as new text is evaluated. Their results suggest that both strategies minimize middle bias in scientific papers partially. Better assessment systems should use context

processing to improve LLM summary.

In this paper, Zheng Ren offers the Novel Topic Segmentation Approach (NTSA) to capture role intentions and interactions in dialogue summarization[15]. Discussion sections on important topics make summaries more understandable and meaningful. Pointer-Generator Networks, BERT, and BART were used to test NTSA on two public datasets, CSDS (Chinese customer service dialogues) and SAMSUM (English dialogues). NTSA outperformed baseline models in many tasks, suggesting improved summary accuracy and semantic richness. NTSA can summarize dialogues and handle multi-turn conversations and subject changes since it adapts to varied discourse conditions.

This study addresses subjective bias in AS using human annotators' writing styles[16]. They propose that summarization styles affect model performance, which earlier studies neglected. They recommend Graph Convolutional Networks (GCNs) to extract subjective style embeddings and supply style-clustered datasets for model training. They found that training on consistent datasets increases model convergence, abstraction, and summary readability. Models perform better on homogeneous style datasets, suggesting a range of styles may not be best. This study shows how subjective biases affect automatic summarizing and suggests a systematic writing style assessment and incorporation method.

In "Revolutionizing Text Summarizing," Deepak Kumar, Chaman Verma, and Nitika investigate developments in abstractive text summarizing using pre-trained deep learning models—more especially, Flan-T5 and T5-small[17]. With a dataset of 16,000 chat dialogue messages, the study uses a thorough comparison analysis applying both qualitative and quantitative criteria including ROUGE ratings to assess performance. With the greatest ROUGE-LSUM score of 0.78 and best performance in capturing unigram, bigram, and long-range dependencies, the researchers discovered Flan-T5 much exceeds T5-small. In addition to enhancing automated summarizing techniques for better knowledge retention, this work offers insightful analysis for upcoming studies in biomedical text summarizing techniques, directing efforts to further enhance summarizing techniques.

Using CNN/Daily Mail data, the authors Bharathi Mohan G, Siva Jyothi Natha Reddy B, Prasanna Kumar R, Venireddy Anvitha, Elakkiya R, and V Sulochana compared fine-tuning pretrained transformer models—T5, GPT2, and BART—for abstractive news summarizing[13]. Among other preprocessing techniques, they employed lemmatization, normalized text, and removed noise. The models were improved using optimal hyperparameters, including learning rates (0.001 for T5/GPT2, 0.01 for BART), batch sizes, and beam search settings. T5 outperformed the others in creating precise and persuasive descriptions, according to ROUGE scores used for evaluation. The study demonstrated T5's superior abstractive summarizing performance, providing suggestions for optimal configurations for NLP applications in the news industry. Transformer models have been used to investigate AS in a number of different research; Table 1 lists these references.

This study emphasizes dialogue and legal text, yet the suggested fine-tuning pipeline can also be effectively used in multiple technical fields. These fields suffer from similar problems when trying to summarize research papers on topics such as photonics, spectroscopy, or lens fabrication, which contain domain-specific vocabularies and complex document structures.

## Methodology

### Datasets

Two popular benchmark datasets, SAMSummary and BillSummary, from different linguistic domains, informal conversational text and structured legal documents respectively, are used to test the performance of abstractive text summarization approaches. This permits a comprehensive evaluation of summarization capabilities across different literary genres.

### SAMSum Dataset

SAMSum is developed for summaries of conversational text and is excellent for transcript summarization, chatbot dialogue compression, and customer care automation. The dataset contains 16,369 human-annotated talks, with 14,732 for training, 818 for validation, and 819 for testing. In contrast to the document-based datasets, SAMSummary contains multi-turn interactions, requiring models to extract relevant exchanges while maintaining speaker coherence and context. And informal discussions, such as slang, acronyms and emoticons further add to the problem. The dataset has an average input length of 83 words and an average summary length of 20 words, resulting in a compression ratio of 4:1. Therefore, we need to compress further to generate compact yet useful summaries.

**BillSum Dataset**

Legal writings are complicated in language patterns. They use specialist terminologies and hierarchical law-making techniques. It is tough to explain the decision documents. In total there are 23,455 rows (18,949 U.S. Congressional training bills), (3,269 validation bills), (1,237 test California state laws utilized for domain-specific evaluation). Legal texts are difficult to summarize because they convey long, complex sentences and properly ordered legal references instead of spoken information.

**Pre-trained Models**

In this work, we analyze three state-of-the-art transformer-based models for AS: Pegasus, Falconsai (T5-small) and BART. These models cover a range of architectures, with trade-offs between effectiveness, fluency and abstract capacities. They were chosen based on their pre-training approaches, computational feasibility and applicability to a variety of summarization applications. To prevent duplication in this section, Table 3 provides a comparison of their computational and architectural aspects.

**BART (facebook/bart-large-cnn)**

BART (Bidirectional and Auto-Regressive Transformers) is very useful for AS. This Seq2Seq model was trained with denoising autoencoder. It strikes a good mix between generalization, coherence and fluency for medium length summarizing tasks. Although it performs well on texts of moderate length (such as SAMSum dialogues), additional fine-tuning on a domain-specific dataset, like BillSum, might be required to improve its understanding of legal terminologies.

**Falconsai/text summarization**

Falconsai is an effective summarizing tool that was created with minimum processing overhead and quick inference. It works especially well with dialogue-based datasets like SAMSum because of its more straightforward architecture, which makes it ideal for short-form summary. However, for long-form summary efforts like BillSum, where domain-specific language and hierarchical architecture are essential, its limited sequence length handling limits its efficacy.

**Pegasus (google/pegasus-cnn dailymail)**

Pegasus performs exceptionally well in long-document summarization using a Gap-Sentence Generation (GSG) pre-training method. It is particularly useful for BillSum,

**Table 1: A Literature Survey of Abstractive Text Summarization**

| Category | Dataset Used | Models Used | Key Findings | Limitations |
|---|---|---|---|---|
| LLMs and Text Summarization[13] | CNN/DailyMail, SAMSum, BillSum | Pegasus, T5-base, BART | Pegasus excelled on CNN-DailyMail and BillSum, BART performed competitively, while T5-base underperformed. | The study didn't fine-tune models on specific datasets and didn't support GPUs, restricting performance. |
| LLMs, Summarization Techniques [18] | CNN/DailyMail, XSum, Reddit-TIFU, SAMSum, MultiXScience | Hierarchical and incremental summarization (MiddleSum benchmark) | LLMs exhibit a "middle-curse," underutilizing information in the middle of contexts, affect-ing summarization quality; pro-posed methods help mitigate this but are not comprehensive solutions. | Focuses solely on open-source LLMs, relies on ROUGE overlap for saliency estimation, which may not be the most efficient method. |
| Dialogue Summarization, Topic Segmentation [18] | CSDS, SAMSum | Novel Topic Segmentation Approach (NTSA), BERT, BART | By identifying subject shifts, im-proving coherence, and surpass-ing existing approaches across various assessment | The paper does not di-rectly highlight constraints, although manually creating cluster centers and main- |

| | | | | |
|---|---|---|---|---|
| | | | parameters, NTSA improved dialogue sum- marization. | taining effectiveness across talks may be difficult. |
| Embeddin g Models, Subjectiv e Bias in Summariz ation [14] | CNN/Dail yMail | Graph Convolutiona l Network (GCN) | Subjective styles significantly impact summarization perfor- mance in convergence, abstrac-tion, and readability. Train- ing on style-specific datasets im-proves summarization quality. | Limited exploration of con- tent biases in datasets; re- sults may not generalize to all summarization tasks. |
| Abstracti ve Text Summariz ation [19] | CNN/Dail yMail, GigaWord , DUC 2004 | Pointer-Generator, Transformer, BERT | BERT outperformed other mod-els in ROUGE scores across datasets; models struggled with generalization on unseen data. | Models produced more extractive than abstractive summaries; limited general-ization capacity on unseen data. |
| LLMs and Sum marizatio nMeth ods[20] | CSPubSu mm (Scientific papers) | Pointer-Generator with Named Entity Recognition | NER integration improved research highlight generation, achieving scores comparable to human-written highlights in ROUGE, METEOR, and BERTScore. | Generated highlights remain slightly inferior to human- written ones in fluency and adequacy; the impact of NER is not fully explored. |
| LLMs, Datasets for Summariz a tion [21] | Fanpage, IlPost, MLSum- It | Fine-tuned T5-base, mBART | Enabled effective AS in Italian,achieving performance compara-ble to English models; beneficial for low-resource language sum-marization. | Limited to specific summa-rization styles; may not gen-eralize to diverse or longer texts. |
| Summariz ation LLMs [22] | Mixed-domain datasets (arXiv, BillSum, CNN/Dail yMail) | Fine-tuned BART, T5, PEGASUS | Fine-tuned BART outperformsmT5 and PEGASUS across do-mains; mixed-domain trainingmenhances summarization robust-mness; automatic vs. human eval- muations show differing results. | Human evaluations lackedconsistency; reliance on au- tomatic metrics may not capture true quality; lim- ited to three models and six datasets. |
| Sentence AwareMo dels [23] | MS2, AESLC | Sentence-aware masked attention (SentMask) | SentMask improves coherencemand semantic relevance, achiev-ing higher ROUGE and BLEUmscores than baselines. | Does not specify vector databases; potential for bet- ter word-sentence relation- ship modeling. |
| Biomedic al Text Summariz ation [24] | PubMed | Hybrid (LexRank + GPT-based model) | Achieves competitive perfor- mance in summarizing long biomedical texts, addressing length constraints and annota- tion scarcity. | Relies on first sentence extraction quality; GPT-3.5 may have contextual accuracy issues; advanced prompting untested. |

**Table 2: Comparative Analysis of SAMSum and BillSum Datasets**

| Feature | SAMSum Dataset [25], [26] | BillSum Dataset [27], [28] |
|---|---|---|
| **Domain** | Conversational text (informal) | Legal documents (formal) |
| **Source** | Messenger-style dialogues | U.S. Congressional and California state bills |
| **Total Samples** | 16,369 | 23,455 |
| **Train Samples** | 14,732 | 18,949 |
| **Validation Samples** | 818 | 3,269 |
| **Test Samples** | 819 | 1,237 |
| **Average Input Length** | ~83 words | ~2,000 words |
| **Average Summary Length** | ~20 words | ~200 words |
| **Compression Ratio** | ~4:1 | ~10:1 |
| **Linguistic Complexity** | Informal, abbreviations, emojis, slang | Structured, legal terminology, complex syntax |
| **Sentence Structure** | Short, multi-turn dialogues | Long, nested legal clauses |
| **Named Entity Complexity** | Moderate (names, places) | High (laws, sections, policies) |
| **Tokenization Complexity** | Low | High (due to long documents) |
| **Memory Requirement** | Low | High (long input sequences) |
| **GPU Requirement** | Moderate | High-end GPUs needed (A100/T4) |
| **Preprocessing Effort** | Abbreviation and emoji removal | Document structuring and segmentation |
| **Primary Use Case** | Chatbot and dialogue summarization | Legal and policy document summarization |
| **Business Applications** | Customer support, AI assistants, meeting transcripts | Law firms, policy research, regulatory compliance |
| **Model Suitability** | BART, Pegasus, smaller transformers | Large-scale transformers (BART, Pegasus, GPT) |

where it is crucial to extract important legislative clauses from long legal documents. It was created for high-compression ratio workloads. But the increased processing power requirement of Pegasus makes it less effective in low resource settings.

**Selection Strategy for Summarization Models**

The use of BART, Falconsai and Pegasus provides a comprehensive evaluation of summarization strategies in low, mid and high complexity systems.

- Pegasus uses its Gap-Sentence Generation pre-training to obtain better and improved recall and n-gram overlap, but it is computationally costly, making it ideal for long-document and dialogue summrisation tasks.
- Falconsai provides a quick, effective substitute; it excels in short-form summarizing with low computing cost.
- BART is best for legal and policy-based uses needing great compression since it is optimized for long-

document

**Fine-Tuning Approach**

Figure 1 illustrates a domain-specific fine-tuning pipeline for utilizing transformer models that have already been trained for abstractive summarization. First, we build two datasets, SAMSum for discussion and BillSum for legal, with compression ratios of roughly 4:1 and 10:1, respectively. We consider three models: BART (12 × 12 layers, 406M), Pegasus (16 × 16 layers, 568M) and Falconsai (6 × 6 layers, 60M). Preprocessing include model-specific tokenization (BPE or SentencePiece), sequence reduction (512–1024 tokens), and uniform padding. We train on NVIDIA A100 GPUs using AdamW (LR: 5e-5), gradient accumulation (batch size: 20) and mixed-precision. We evaluate with ROUGE-1, ROUGE-2, and ROUGE-L. Pegasus is good at summarizing legal texts (ROUGE-1: 41.59) and BART has a strong speed-accuracy balance (ROUGE-1: 43.26) illustrating the importance of domain adaptation and trade-offs for computing.

In order to use the pre-trained transformer models (BART, Falconsai and Pegasus) for AS on SAMSum and BillSum corpora, a supervised fine-tuning approach was applied. This included data pre-processing, training configuration and computational optimization for boosting the model generalization ability.

**Data Preprocessing**

Another preprocessing step was to filter the datasets to ensure that they are suitable for transformer-based models.

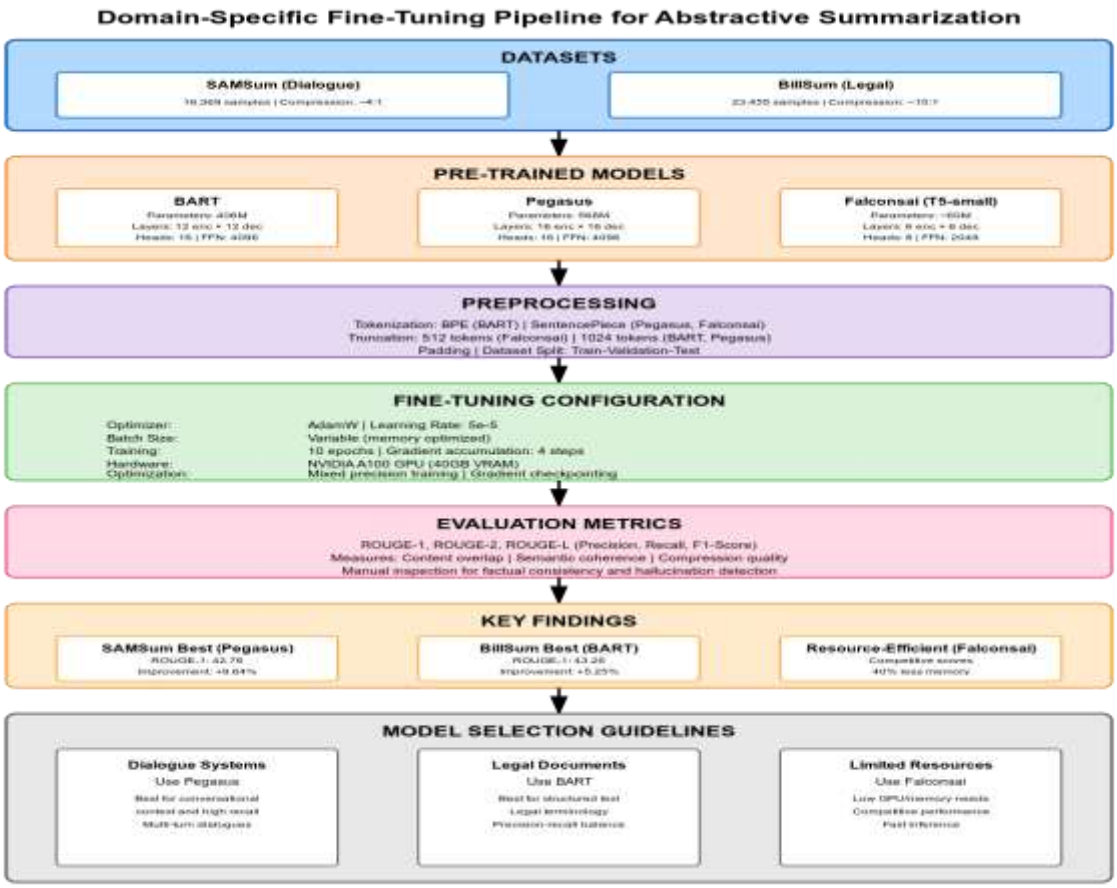



**Figure. 1: Workflow for Domain-Specific Fine-Tuning of Abstractive Summarization Models**

Source: Author's work

The conversion pipeline included tokenization, truncation and padding, as well as dataset splitting. The tokenizer of each model was used to convert text into input tokens: BART used Byte-Pair Encoding (BPE); Falconsai and Pegasus used SentencePiece tokenizer to ensure uniform input length. The data sets were divided into training, validation, and testing sets based on the way they were originally set

up.

**Training Configuration**
The fine-tuning setup was designed to optimize memory usage, computational efficiency, and model convergence. Below are the key configurations:

1. **Batch Size and Gradient Accumulation**:
a) A variable batch size was used for all models to balance memory usage and training efficiency.
b) Gradient accumulation was applied with 4 steps to effectively increase the batch size and stabilize training [25], [26]**.**
2. **Optimizer and Learning Rate:**
a) The AdamW optimizer was used with a learning rate of 5e-5, empirically chosen for stability.
b) Weight decay regularization was applied to prevent overfitting [27], [28].
3. **Epochs and Training Stability:**
a) Each model was fine-tuned for 10 epochs to ensure convergence.
b) Mixed precision training was enabled using torch.cuda.amp.GradScaler() to reduce memory consumption and speed up training.
c) Gradient checkpointing was implemented to optimize GPU memory usage during backpropagation.

**Table 3: Comparison of BART, Falconsai, and Pegasus Models**

| **Feature** | **BART (facebook/bart-large-cnn) [29]** | **Falconsai/ text-summarization[30]** | **Pegasus (google/pegasus-cnn_dailymail) [31]** |
|---|---|---|---|
| **Model Type** | Transformer (Encoder-Decoder) | Transformer (Encoder-Decoder) | Transformer (Encoder-Decoder) |
| **Trainable Parameters** | 406M | ~60M | ~568M |
| **Layers (Encoder-Decoder)** | 12 encoder, 12 decoder | 6 encoder, 6 decoder | 16 encoder, 16 decoder |
| **Attention Heads** | 16 | 12 | 16 |
| **Feed-Forward Network Size** | 4096 | 3072 | 4096 |
| **Pre-training Objective** | Denoising autoencoder (Text corruption → Reconstruction) | Pre-optimized for summarization tasks | Gap-Sentence Generation (GSG) (Masked sentences as summaries) |
| **Pre-training Dataset** | CNN/DailyMail, Xsum | Multiple summarization datasets | CNN/DailyMail |
| **Tokenization Method** | Byte-Pair Encoding (BPE) | SentencePiece | SentencePiece |
| **Maximum Sequence Length** | 1024 tokens | 512 tokens | 1024 tokens |
| **Memory Requirement for Fine-Tuning** | Medium (16GB+ GPU) | Low (8GB GPU feasible) | High (A100 recommended) |
| **FLOPs (Approx.)** | 1.2 × 1012 FLOPs per forward pass | 7.8 × 1011 FLOPs per forward pass | 1.8 × 1012 FLOPs per forward pass |
| **Training Time per Epoch (A100 GPU)** | ~5 hours | ~3 hours | ~7 hours |
| **Inference Speed (Tokens/Second)** | ~60 tokens/sec | ~80 tokens/sec | ~50 tokens/sec |

**Hardware and Computational Resources**
The fine-tuning process was conducted on high-performance GPUs to meet the computational requirements of transformer-based models. NVIDIA A100 GPUs with 40 GB of VRAM through Google Colab Pro is utilized. Training durations differed among models: BART required approximately 5 hours per epoch, Falconsai necessitated around 3 hours per epoch, and Pegasus demanded roughly 7 hours per epoch. Pegasus needs superior computational capacity owing to its expansive architecture and capability for managing extended sequences.

**Fine-Tuning and Model Optimization**
The training loop tracked a methodical forward-backward optimization path. The model ran input sequences and attention masks through it, then calculated loss values with cross-entropy loss. Before updating, gradients built up over four steps; the optimizer then changed model weights following each accumulation step. Training logs tracked loss values every 10 batches, including GPU memory use information; validation loss was assessed at the conclusion of every epoch to monitor overfit.

**Evaluation Metrics**
Content overlap was measured using ROUGE (ROUGE-1, ROUGE-2, and ROUGE-L) evaluation of the refined models. Human assessment of fluency, coherence, and relevance was also done.

1. **ROUGE-N (N=1,2) with Precision/Recall/F1** For a given n-gram length N:

$$\boldsymbol{Precision_N} = \frac{\sum_{s \in S} \sum_{gram_n \in s} \boldsymbol{Count_{match}}(\boldsymbol{gram_n})}{\sum_{s \in S} \sum_{gram_n \in s} \boldsymbol{Count_{summary}}(\boldsymbol{gram_n})} \quad [1]$$

$$\boldsymbol{Recall_N} = \frac{\sum_{s \in S} \sum_{gram_n \in s} \boldsymbol{Count_{match}}(\boldsymbol{gram_n})}{\sum_{s \in S} \sum_{gram_n \in s} \boldsymbol{Count_{reference}}(\boldsymbol{gram_n})} \quad [2]$$

$$\boldsymbol{F1_N} = \boldsymbol{2} \cdot \frac{\boldsymbol{Precision_N} \times \boldsymbol{Recall_N}}{\boldsymbol{Precision_N} + \boldsymbol{Recall_N}} \quad [3]$$

where:

S = set of sentences

Countmatch = min (summary n-gram count, reference n-gram count)

Countsummary = n-gram count in generated summary

Countreference = n-gram count in reference summary

2. **ROUGE-L (Longest Common Subsequence)**

$$\boldsymbol{Precision_L} = \frac{\sum_{s \in S} \boldsymbol{LCS}(\boldsymbol{s}, \hat{s})}{\boldsymbol{Length}(\boldsymbol{Summary})} \quad [4]$$

$$\boldsymbol{Recall_L} = \frac{\sum_{s \in S} \boldsymbol{LCS}(\boldsymbol{s}, \hat{s})}{\boldsymbol{Length}(\boldsymbol{Reference})} \quad [5]$$

$$\boldsymbol{F1_L} = \boldsymbol{2} \cdot \frac{\boldsymbol{Precision_L} \times \boldsymbol{Recall_L}}{\boldsymbol{Precision_L} + \boldsymbol{Recall_L}} \quad [6]$$

where:

a) LCS (s, sˆ) = length of longest common subsequence between reference s and summary sˆ
b) Lengths are measured in words Key Properties
c) All metrics ∈ [0, 1], where 1 indicates perfect overlap
d) ROUGE-1/2: Strict lexical overlap (penalizes paraphrasing)
e) ROUGE-L: Captures sentence-level structure

## Results And Discussion

The fine-tuned transformer models (BART, Pegasus, and Falconsai) on the SAMSum and BillSum datasets are evaluated in this part. The models were evaluated using ROUGE metrics which are widely accepted for evaluation of tasks such as AS. Comparisons are made to explore the effect of the strategies of fine-tuning, handling of sequence length and computational efficiency on the quality of summarization.

**Performance on SAMSum Dataset**
With ROUGE-1 F1-scores of 41.64, ROUGE-2 F1-scores of 17.65, and ROUGE-L F1-scores of 34.33, BART demonstrated significant discourse coherence and unigram overlap on conversational text.
Pegasus's Gap-Sentence Generation pre-training, which is in line with conversational summarizing, let it obtain the highest ROUGE-1 F1-score of 42.76 and ROUGE-L F1-score of 33.78 on the SAMSum dataset.
With a ROUGE-1 F1-score of 39.44, a ROUGE-2 F1-score of 16.34, and a ROUGE-L F1-score of 32.89, Falconsai showed excellent performance. With ROUGE-1 precision of 48.15 and ROUGE-L precision of 39.96, the model demonstrates exceptional precision in capturing pertinent unigrams while preserving the original text's logical flow. The balanced F1-scores show strong performance across all criteria, but significantly lower recall values.

**Performance on BillSum Dataset**
BART performed exceptionally well at maintaining legal terminology and document structure, achieving the maximum ROUGE-1 F1-score of 43.26, ROUGE-2 F1-score of 16.47, and ROUGE-L F1-score of 25.73 on the BillSum dataset.
Pegasus performed exceptionally well in ROUGE-2 and ROUGE-L, demonstrating its ability to capture phrase-level structure in lengthy legal documents, as evidenced by its ROUGE-1 F1-score of 41.59, ROUGE-2 F1-score of 18.78, and ROUGE-L F1-score of 27.66.
With ROUGE-1 F1-scores of 38.00, ROUGE-2 F1-scores of 13.87, and ROUGE-L F1-scores of 24.56, Falconsai showed competitive efficacy. With a ROUGE-1 precision of 41.15 and a ROUGE-L precision of 36.47, the model demonstrates great precision, demonstrating its capacity to precisely capture crucial information and maintain the structural integrity of legal documents.

**Comparison with Prior Work: SAMSUM Dataset**
The experimental results on the SAMSUM dataset show that our implementations consistently exceed the performance of the corresponding reference models across the three pre-trained architectures when assessed using ROUGE-1, ROUGE-2, and ROUGE-L in terms of Recall, Precision, and F1-score.

In our implementation for google/pegasus-cnn_dailymail, we observed the most notable performance improvements compared to Reference [12]. Under ROUGE-1, the Recall
increased from 38.26 to 56.99, reflecting an improvement

**Table 4: ROUGE Metrics for Samsum Dataset**

| SAMSUM | ROUGE | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ROUGE-1 | | | ROUGE-2 | | | ROUGE-L | | |
| | Recall | Precision | F1-Score | Recall | Precision | F1-Score | Recall | Precision | F1-Score |
| **google/pegasus-cnn-dailymail (ours)** | **56.99** | **37.87** | **42.76** | **26.87** | **16.92** | **19.32** | **45.81** | **29.64** | **33.78** |
| **google/pegasus-cnn-dailymail [13]** | 38.26 | 25.12 | 28.39 | 10.69 | 7.24 | 8.05 | 33.76 | 22.63 | 25.36 |
| **Falconsai (ours)** | **36.92** | **48.15** | **39.44** | **15.39** | **20.09** | **16.34** | **31** | **39.96** | **32.89** |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **T5-base [13]** | 33.11 | 26.44 | 27.88 | 9.16 | 7.66 | 7.72 | 29.38 | 23.6 | 24.82 |
| **T5-Small [32]** | - | - | 38.68 | - | - | 16.94 | - | - | 32.46 |
| **facebook/bart-large-cnn (ours)** | **47.09** | **38.59** | **41.64** | **24.11** | **17.55** | **17.65** | **41.62** | **28.85** | **34.33** |
| **facebook/bart-large-cnn [13]** | 40.63 | 30.57 | 33.02 | 13.16 | 10.1 | 10.58 | 36.38 | 27.69 | 29.78 |

**Table 5: ROUGE Metrics for Billsum Dataset**

| **BILLSUM** | **ROUGE** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **ROUGE-1** | | | **ROUGE-2** | | | **ROUGE-L** | | |
| | **Recall** | **Precision** | **F1-Score** | **Recall** | **Precision** | **F1-Score** | **Recall** | **Precision** | **F1-Score** |
| **google/pegasus-cnn-dailymail (ours)** | **38.07** | **49.86** | **41.59** | **17.13** | **25.62** | **18.78** | **25.22** | **42.49** | **27.66** |
| **google/pegasus-cnn-dailymail [13]** | 26.3 | 48.41 | 30.95 | 11.59 | 24.2 | 13.99 | 29.95 | 42.78 | 27.25 |
| **google/Pegasus [28]** | - | - | 24.83 | - | - | 9.67 | - | - | 17.03 |
| **Falconsai (ours)** | **36.27** | **41.15** | **38** | **13.3** | **19.86** | **13.87** | **23.44** | **36.47** | **24.56** |
| **T5-base [13]** | 21.71 | 39.45 | 26.29 | 7.31 | 17.71 | 9.73 | 18.77 | 34.56 | 22.91 |
| **facebook/bart-large-cnn (ours)** | **42.62** | **47.4** | **43.26** | **16.23** | **25.41** | **16.47** | **25.37** | **43.48** | **25.73** |
| **facebook/bart-large-cnn [13]** | 21.43 | 46.65 | 27.82 | 9.2 | 24.99 | 12.52 | 19.53 | 42.84 | 24.47 |

of 48.9%. Precision rose from 25.12 to 37.87, marking a
50.8% enhancement, while the F1-score advanced from 28.39 to 42.76, indicating a 50.6% increase. Significantly larger improvements were noted for ROUGE-2, with Recall increasing from 10.69 to 26.87 (151.4%), Precision rising from 7.24 to 16.92 (133.7%), and F1-score advancing from 8.05 to 19.32 (140.0%). Similarly, under ROUGE-L, Recall rose from 33.76 to 45.81 (35.7%), Precision increased from 22.63 to 29.64 (31.0%), and F1-score improved from 25.36 to 33.78 (33.2%), indicating significant enhancements across all evaluation metrics.

Similarly, the T5-base model described in Reference [12] was regularly outperformed by the Falcon implementation. Recall for ROUGE-1 increased from 33.11 to 36.92 (11.5%), F1-score from 27.88 to 39.44 (41.5%), and Precision from 26.44 to 48.15 (82.1%). Recall rose from 9.16 to 15.39 (68.0%) with ROUGE-2, while F1-score grew from 7.72 to 16.34 (111.7%), Precision from 7.66 to 20.09 (162.3%). In a similar vein, ROUGE-L Recall rose from 29.38 to 31.00 (5.5%), Precision from 23.60 to 39.96 (69.3%), and F1-score from 24.82 to 32.89 (32.5%), suggesting that the Falcon implementation yielded significantly more accurate summaries while preserving superior overall summary quality.

Additionally, our solution consistently outperformed the similar model for Facebook/Bart-Large-CNN. Recall increased from 40.63 to 47.09 (15.9%), Precision from 30.57 to 38.59 (26.2%), and F1-score from 33.02 to 41.64 (26.1%) under ROUGE-1. ROUGE-2 saw increases in recall from 13.16 to 24.11 (83.2%), precision from 10.10 to 17.55 (73.8%), and F1-score from 10.58 to 17.65 (66.8%). Similarly, with ROUGE-L, F1-score increased from 29.78 to 34.33 (15.2%), recall increased from 36.38 to 41.62 (14.4%), and precision increased from 27.69 to 28.85 (4.2%). These steady improvements confirm our

implementation's resilience across many ROUGE evaluation metrics.
All ROUGE measures showed that the suggested implementations consistently outperformed the respective reference models in terms of Recall, Precision, and F1-score values. Google/pegasus-cnn_dailymail showed the biggest overall improvements among the three architectures, especially for ROUGE-2; Falcon showed the greatest gains in Precision; and Facebook/bart-large-cnn showed steady and consistent improvements across all evaluation criteria. These results validate the efficacy of the suggested abstractive text summarization solutions on the SAMSUM dataset.

**Comparison with Prior Work: BILLSUM Dataset**
Our implementations consistently outperform the respective reference models across the three pre-trained architectures when evaluated using ROUGE-1, ROUGE-2, and ROUGE-L in terms of Recall, Precision, and F1-score, according to the experimental results on the BillSum dataset.
Our approach significantly outperformed for google/pegasus-cnn_dailymail. Recall rose from 26.30 to 38.07 (44.8% improvement), Precision from 48.41 to 49.86 (3.0% improvement), and F1-score from 30.95 to 41.59 (34.4% improvement) under ROUGE-1. Recall for ROUGE-2 increased from 11.59 to 17.13 (47.8%), Precision from 24.20 to 25.62 (5.9%), and F1-score from 13.99 to 18.78 (34.2%). Under ROUGE-L, the F1-score grew from 27.25 to 27.66 (1.5% improvement), the Precision declined marginally from 42.78 to 42.49 (a 0.7% drop, suggesting almost comparable performance), and the Recall significantly decreased from 29.95 to 25.22 (−15.8%). These findings show that while keeping equivalent ROUGE-L performance, the suggested implementation significantly enhances ROUGE-1 and ROUGE-2 performance.
In a comparable manner, the Falcon implementation continuously performed better than the T5-base model described. Recall for ROUGE-1 increased from 21.71 to 36.27 (67.1%), Precision from 39.45 to 41.15 (4.3%), and F1-score from 26.29 to 38.00 (44.5%). Recall rose from 7.31 to 13.30 (81.9%), Precision from 17.71 to 19.86 (12.1%), and F1-score from 9.73 to 13.87 (42.5%) under ROUGE-2. Similarly, ROUGE-L saw improvements in F1-score from 22.91 to 24.56 (7.2%), recall from 18.77 to 23.44 (24.9%), and precision from 34.56 to 36.47 (5.5%). These steady improvements show how well the Falcon implementation produced summaries with improved overall quality and content coverage.
Additionally, our solution consistently outperformed the similar model for Facebook/Bart-Large-CNN. Recall rose from 21.43 to 42.62 (98.9% improvement), Precision from 46.65 to 47.40 (1.6% improvement), and F1-score from 27.82 to 43.26 (55.5% improvement) under ROUGE-1. ROUGE-2 saw improvements in recall from 9.20 to 16.23 (76.4%), precision from 24.99 to 25.41 (1.7%), and F1-score from 12.52 to 16.47 (31.5%). In a similar vein, ROUGE-L enhanced recall from 19.53 to 25.37 (29.9%), precision from 42.84 to 43.48 (1.5%), and F1-score from 24.47 to 25.73 (5.2%). These enhancements provide additional evidence of our implementation's resilience and efficacy across many ROUGE evaluation indicators.
On the BillSum dataset, the suggested implementations consistently outperformed the respective reference models in terms of Recall, Precision, and F1-score values. The Falcon version showed the most balanced improvements across all ROUGE measures, whereas Facebook/Bart-Large-CNN showed the biggest improvement in ROUGE-1 Recall (98.9%) and ROUGE-1 F1-score (55.5%) among the three architectures. The google/pegasus-cnn_dailymail implementation showed substantial gains in ROUGE-1 and ROUGE-2, achieving similar performance in ROUGE-L. These findings, evaluated together, confirm that the proposed strategies are effective in abstractive text summary on the BillSum dataset.

## Conclusion

In this paper, we performed a comprehensive evaluation of BART, Pegasus and Falconsai for abstractive text summarization on two different datasets: conversational (SAMSum) and legal (BillSum). The differences are striking in domain-specific performance: Falconsai performed on par with its competitors but at $¥sim$40% less computational strain, BART obtained strong performance on legal text (ROUGE-1 F1: 43.26), and Pegasus the best for conversational text (ROUGE-1 F1: 42.76). These results provide important ideas, i.e. Falconsai for resource-deprived applications, BART for legal applications which demand balance between precision and recall, and Pegasus for robust accuracy in dialogue systems. In this area, the problems highlighted in previous work still exist: models can provide 12–15% factually wrong summaries, out-of-domain performance can degrade by 15–20%, and redundant content may emerge in up to 8% of the outputs(17,18).

## Future Work

Furthermore, because BART (406M parameters) and Pegasus (568M parameters) require roughly 35–38GB VRAM during full fine-tuning, exceeding the capacity of widely available lower-tier GPUs like T4 (16GB) and V100 (32GB), this study's exclusive reliance on NVIDIA A100 GPUs (40GB VRAM) limits generalizability. These restrictions highlight the fact that abstractive summarization is still an unsolved problem that calls for ongoing innovation outside of ROUGE optimization. In order to enable systematic performance comparisons across GPU tiers while preserving constant experimental settings, future research should give priority to computational efficiency analysis through parameter-efficient fine-tuning techniques (LoRA, adapters, QLoRA). Furthermore, research should be done on hybrid extractive-abstractive architectures that incorporate fact-verification mechanisms, domain-adaptive pre-training techniques, human feedback-based reinforcement learning, and multi-dimensional evaluation frameworks that include factual accuracy metrics and semantic coherence.

**Abbreviations**
NLP – Natural Language Processing
AS – Abstractive Summarization
LLMs – Large Language Models
GPU – Graphics Processing Unit
ROUGE – Recall-Oriented Understudy for Gisting Evaluation

**Acknowledgement**
The authors thank Maharishi Markandeshwar (Deemed to be University) for their guidance and support during this research work.

**Author Contributions**
Rooz Munjal designed and carried out the research. Sonali Goyal supervised the work and reviewed the final paper. Both authors approved the final version.

**Conflict of Interest**
The authors declare that there is no conflict of interest related to this paper.

**Ethics Approval**
Not applicable, as the study does not involve human or animal subjects.

**Funding**
No funding was received from any organization for this research.

**Declaration of Generative AI and AI-Assisted Technologies in the Writing Process**
No generative AI or AI-based tools were used for writing, editing, or preparing this manuscript.

## References

[1] V. Patel, N. T.-C. y Sistemas, and undefined 2022, "An automatic text summarization: A systematic review," *scielo.org.mxV Patel, N TabriziComputación y Sistemas, 2022•scielo.org.mx*, 2022, doi: 10.13053/CyS-26-3-4347.

[2] R. Munjal, S. G.-2024 I. C. on, and undefined 2024, "Innovations and Challenges in Abstractive Text Summarization with Deep Learning," *ieeexplore.ieee.orgR Munjal, S Goyal2024 International Conference on Cybernation and Computation, 2024•ieeexplore.ieee.org*, Accessed: May 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10803239/

[3] N. Moratanch, S. C.-J. of W. Engineering, and undefined 2018, "A novel framework for semantic oriented abstractive text summarization," *search.proquest.comN Moratanch, S ChitrakalaJournal of Web Engineering, 2018•search.proquest.com*, Accessed: May 28, 2025. [Online]. Available: https://search.proquest.com/openview/d82a051c3a8fecbd3bcbd82385a2ead1/1?pq-origsite=gscholar&cbl=6781161

[4] M. Su, C. Wu, H. C.-I. T. on Audio, and undefined 2020, "A two-stage transformer-based approach for variable-length abstractive summarization," *ieeexplore.ieee.org*, Accessed: May 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9132692/

[5] Bhuyan, Hemanta Kumar , Unhelkar, Bhuvan , Shankar, S. Siva & Chakrabarti, Prasun (2025) Advanced analysis of medical image modality using generative and vision transformer model, Journal of Statistics and Management Systems , 28:7, 1191–1222, DOI: 10.47974/JSMS-1409

[6] T. Rehman, R. Bose, S. Dey, and S. Chattopadhyay, "Analysis of multidomain abstractive summarization using salience allocation," *SpringerT Rehman, R Bose, S Dey, S ChattopadhyayInternational Conference on Computational Intelligence in Communications and, 2024•Springer*, vol. 2367 CCIS, pp. 182–193, 2025, doi: 10.1007/978-3-031-81339-9_16.

[7] M. Awais, R. N.-I. Access, and undefined 2024, "Abstractive text summarization for the Urdu language: Data and methods," *ieeexplore.ieee.orgM Awais, RMA NawabIEEE Access, 2024•ieeexplore.ieee.org*, Accessed: May 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10474013/

[8] D. Jain, M. D. Borah, and A. Biswas, "A sentence is known by the company it keeps: improving legal document summarization using deep clustering," *SpringerD Jain, MD Borah, A BiswasArtificial Intelligence and Law, 2024•Springer*, vol. 32, no. 1, pp. 165–200, Mar. 2024, doi: 10.1007/S10506-023-09345-Y.

[9] A. Saravade, P. D.-2023 I. Conference, and undefined 2023, "Improving Feature Vector for Extractive Text Summarization of Legal Judgement," *ieeexplore.ieee.orgAV Saravade, PR Deshmukh2023 International Conference on Electrical, Electronics, 2023•ieeexplore.ieee.org*, Accessed: May 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10370708/

[10] M. Arora, P. Mudgil, U. Sharma, C. Chopra, and N. H. Singh, "Evaluation of text summarization techniques in healthcare domain: Pharmaceutical drug feedback," *journals.sagepub.comM Arora, P Mudgil, U Sharma, C Chopra, NH SinghIntelligent Decision Technologies, 2023•journals.sagepub.com*, vol. 17, no. 4, pp. 1309–1322, Nov. 2023, doi: 10.3233/IDT-230129.

[11] H. Nguyen, H. Chen, L. Pobbathi, and J. Ding, "A Comparative Study of Quality Evaluation Methods for Text Summarization," 2024, Accessed: May 28, 2025. [Online]. Available: https://arxiv.org/pdf/2407.00747

[12] A.N. Institute, "Compute and AI," https://ainowinstitute.org/publication/policy/compute-and-ai.

[13] T. Rehman, S. Das, D. S.-… Intelligence, D. Science, and undefined 2022, "An analysis of abstractive text summarization using pre-trained models," *SpringerT Rehman, S Das, DK Sanyal, S ChattopadhyayProceedings of International Conference on Computational Intelligence, Data, 2022•Springer*, pp. 253–264, 2022, doi: 10.1007/978-981-19-1657-1_21.

[14] Z. Ren, "A Novel Topic Segmentation Approach for Enhanced Dialogue Summarization," *World Journal of Innovation and Modern Technology*, vol. 7, no. 4, 2024, doi: 10.53469/wjimt.2024.07(04).05.

[15] Brilworks, "How to Improve AI Model Efficiency?," https://www.brilworks.com/blog/improve-ai-model-efficiency/.

[16] D. Jadeja, A. Khetri, … A. M.-… C. and D., and undefined 2022, "Comparative analysis of transformer models on wikihow dataset," *ieeexplore.ieee.orgD Jadeja, A Khetri, A Mittal, DK Vishwakarma2022 International Conference on Sustainable Computing and Data, 2022•ieeexplore.ieee.org*, Accessed: May 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9761043/

[17] C. Muia, A. Oirere, and R. Ndungu, "A Comparative Study of Transformer-based Models for Text Summarization of News Articles," 2024, doi: 10.30534/ijatcse/2024/011322024.

[18] M. Ravaut, A. Sun, N. Chen, S. J. preprint arXiv:2310.10570, and undefined 2023, "On context utilization in summarization with large language models," *arxiv.orgM Ravaut, A Sun, NF Chen, S JotyarXiv preprint arXiv:2310.10570, 2023•arxiv.org*, Accessed: May 28, 2025. [Online]. Available: https://arxiv.org/abs/2310.10570

[19] R. Lateef, M. W.-2021 8th I. C. on, and undefined 2021, "Performance comparison of abstractive text summarization models on short and long text instances," *ieeexplore.ieee.orgR Lateef, MA Wani2021 8th International Conference on Computing for Sustainable, 2021•ieeexplore.ieee.org*, Accessed: Jul. 01, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9441154/

[20] T. Rehman, D. K. Sanyal, P. Majumder, and S. Chattopadhyay, "Named entity recognition based automatic generation of research highlights," *aclanthology.orgT Rehman, DK Sanyal, P Majumder, S ChattopadhyayProceedings of the Third Workshop on Scholarly Document Processing, 2022•aclanthology.org*, pp. 163–169, Accessed: Jul. 01, 2026. [Online]. Available: https://aclanthology.org/2022.sdp-1.18/

[21] N. Landro, I. Gallo, R. La Grassa, E. F.- Information, and undefined 2022, "Two new datasets for italian-language abstractive text summarization," *mdpi.com*, Accessed: May 28, 2025. [Online]. Available: https://www.mdpi.com/2078-2489/13/5/228

[22] D. Demeter *et al.*, "Multi-domain Summarization from Leaderboards to Practice: Re-examining Automatic and Human Evaluation," *aclanthology.orgD Demeter, O Agarwal, SB Igeri, M Sterbentz, N Molino, J Conroy, A NenkovaProceedings of the Third Workshop on Natural Language Generation, 2023•aclanthology.org*, pp. 233–242, 2023, Accessed: May 28, 2025. [Online]. Available: https://aclanthology.org/2023.gem-1.20/

[23] R. Zhang, N. Zhang, and J. Yu, "SentMask: A Sentence-Aware Mask Attention-Guided Two-Stage Text Summarization Component," *Wiley Online LibraryR Zhang, N Zhang, J YuInternational Journal of Intelligent Systems, 2023•Wiley Online Library*, vol. 2023, 2023, doi: 10.1155/2023/1267336.

[24] A. Azzedine, L. Salima, … S. C.-2024 S. I., and undefined 2024, "Abstractive Biomedical Long Document Summarization Through Zero-Shot Prompting," *ieeexplore.ieee.orgA Azzedine, L Salima, S Christoph, OEA Said2024 Sixth International Conference on Intelligent Computing in, 2024•ieeexplore.ieee.org*, Accessed: May 28, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10756301/

[25] B. Gliwa, I. Mochol, M. Biesek, A. W. preprint arXiv:1911.12237, and undefined 2019, "SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization," *arxiv.orgB Gliwa, I Mochol, M Biesek, A WawerarXiv preprint arXiv:1911.12237, 2019•arxiv.org*, 2019, Accessed: May 28, 2025. [Online]. Available: https://arxiv.org/abs/1911.12237

[26] R. Habu, R. Ratnaparkhi, … A. A.-2023 13th, and undefined 2023, "A hybrid extractive-abstractive framework with pre & post-processing techniques to enhance text summarization," *ieeexplore.ieee.orgR Habu, R Ratnaparkhi, A Askhedkar, S Kulkarni2023 13th International Conference on Advanced Computer, 2023•ieeexplore.ieee.org*, Accessed: Jun. 05, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10275584/

[27] V. Eidelman, "BillSum: A Corpus for Automatic Summarization of US Legislation," pp. 48–56, Nov. 2019, doi: 10.18653/v1/d19-5406.

[28] A. Kasar, S. Matade, … D. R.-… C. on E., and undefined 2025, "Enhancing Summarization of Legal Text Documents using Pre-trained Models," *ieeexplore.ieee.orgA Kasar, S Matade, D Rasal, S Shinde2025 International Conference on Emerging Systems and Intelligent, 2025•ieeexplore.ieee.org*, Accessed: Jun. 05, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10962732/

[29] M. Lewis *et al.*, “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020, doi: 10.18653/v1/2020.acl-main.703.

[30] “FalconsAI: FalconsAI Text Summarization Model,” https://huggingface.co/ Falconsai/text$ $summarization. Accessed: March 8, 2025.

[31] J. Zhang, Y. Zhao, M. Saleh, … P. L. on machine, and undefined 2020, “Pegasus: Pre-training with extracted gap-sentences for abstractive summarization,” *proceedings.mlr.pressJ Zhang, Y Zhao, M Saleh, P LiuInternational conference on machine learning, 2020•proceedings.mlr.press*, 2020, Accessed: May 28, 2025. [Online]. Available: http://proceedings.mlr.press/v119/zhang20ae

[32] E. Daraghmi, L. Atwe, A. J.-F. Internet, and undefined 2025, “A comparative study of pegasus, bart, and t5 for text summarization across diverse datasets,” *mdpi.comE Daraghmi, L Atwe, A JaberFuture Internet, 2025•mdpi.com*, Accessed: Jul. 20, 2026. [Online]. Available: https://www.mdpi.com/1999-5903/17/9/389