Comparative Analysis and Fine-tuning of Transformer-Based Models for Abstractive Summarization of Medical Literature

Abeyankar Giridharan
Department of Statistics
University of Michigan
Ann Arbor, USA
abeygiri@umich.edu

***Abstract*—This study focuses on abstractive summarization of medical literature to address the growing challenge of information overload in the biomedical field. By utilizing transformer-based models such as T5, BART, PEGASUS, BigBird, and their specialized variants, the research evaluates and fine-tunes these models for generating concise, coherent, and accurate summaries of research papers, clinical trials, and case studies. Datasets like PubMed and MS2 form the foundation of this project, offering a comprehensive resource of domain-specific documents. Using evaluation metrics such as ROUGE, BERTScore, and METEOR, the best-performing model—BigBird—was fine-tuned for improved domain adaptation. This project demonstrates the potential of advanced NLP techniques in enhancing accessibility to critical medical insights.**

I. Introduction

The rapid growth of medical literature poses a significant challenge for researchers, clinicians, and students who need to sift through vast amounts of information to derive relevant insights. The field of biomedical research produces thousands of articles, clinical trials, and case studies daily, making it increasingly difficult to keep up with the latest advancements. This challenge highlights the need for effective summarization tools capable of condensing extensive documents into concise, coherent, and accurate summaries. Recent advancements in transformer-based NLP models offer a promising solution for tackling this issue through abstractive summarization.

The goal of this project is to evaluate and fine-tune pre-trained transformer models for summarizing medical literature, focusing on datasets such as PubMed and MS2. These models are assessed based on their ability to retain critical information, handle domain-specific terminology, and generate summaries of approximately 300 words from documents averaging 2,000 words. By leveraging state-of-the-art models like BigBird and PEGASUS, this study seeks to provide actionable insights into their performance and optimize the best model for biomedical applications.

The research involves a structured approach that begins with evaluating models such as T5, BART, and BigBird using metrics like ROUGE, BERTScore, and METEOR. Fine-tuning techniques, including learning rate optimization and domain-specific training, are then applied to improve the chosen model's performance. This methodology ensures the generation of high-quality summaries, which can aid in faster decision-making and improved access to critical medical knowledge.

A. Literature Review

Text summarization in the biomedical field has traditionally been dominated by extractive methods, which focus on identifying key sentences or phrases directly from the text. However, with the rise of transformer-based models, abstractive approaches—capable of generating new sentences and rephrasing content—have shown significant promise in biomedical summarization.

Extractive Methods: Earlier efforts predominantly used extractive summarization techniques due to structured abstracts in datasets like PubMed and clinical reports. Methods like TextRank and Latent Semantic Analysis (LSA) applied statistical and graph-based approaches to identify salient text segments. However, these methods struggled to capture nuanced relationships in medical literature.

Abstractive Methods: Recent advancements in deep learning have led to interest in abstractive summarization. Early encoder-decoder models for summary generation faced challenges due to the complexity and length of biomedical documents. Transformer-based models like BERT and GPT emerged as promising tools, offering contextual understanding and fluency after fine-tuning with domain-specific data.

Transformers in Text Summarization: Transformer architectures by Vaswani et al. (2017) revolutionized NLP.

Models like T5, BART, PEGASUS, and BigBird excel in summarization tasks. T5's versatility and robust performance on biomedical text summarization are evident when fine-tuned with datasets like PubMed. BART's bidirectional encoder and autoregressive decoder make it suitable for summarizing scientific documents. PEGASUS, optimized for abstractive summarization, uses gap-sentence generation for pre-training, making it highly effective for summarizing complex medical literature. BigBird and BigBird-PubMed extend transformers with sparse attention mechanisms, efficiently processing long documents and biomedical text.

Summarization of Long Documents: Summarizing lengthy biomedical documents poses challenges, particularly in maintaining context while condensing information. Techniques like hierarchical encoding (Cohan et al., 2018) and divide-and-conquer approaches (Gidiotis and Tsoumakas, 2020) have proven effective. Models like LED and BigBird address long-document processing by introducing efficient mechanisms. Datasets like PubMed and MS² support training and evaluation, with MS² also addressing the synthesis of multiple sources.

Bridging Legal and Biomedical Summarization: While much of the foundational work in abstractive summarization comes from domains like legal texts, the methodologies apply to biomedical literature. The use of transformer models for legal text summarization (Yoon et al., 2022) demonstrated the potential of BERT2BERT and BART for domain-specific challenges. Similarly, methods for structured legal texts inform strategies for biomedical literature, which follows standardized formats such as IMRaD.

Building on these foundations, this study focuses on fine-tuning transformer-based models for summarizing medical literature. By leveraging models like BigBird-PubMed and PEGASUS, this research bridges the gap between domain-specific adaptation and general summarization techniques, utilizing datasets like PubMed and MS² to train and evaluate models, ensuring concise, coherent, and clinically relevant summaries.

II. METHOD

A. Problem Formulation

The task at hand is the abstractive summarization of scientific articles, specifically in the domain of medical literature. The objective is to evaluate and fine-tune transformer-based models to generate concise and coherent summaries of lengthy research papers, case studies, and clinical trials.

- **Input:** The input to the model is a long scientific document, which includes research papers, case studies, and clinical trials. These documents are sourced from medical databases such as PubMed and MS2.
- **Output:** The output is an abstractive summary of the input document, aiming to condense the content into a concise summary of approximately 300 words, capturing the essence of the original document while maintaining accuracy and coherence.

Dataset Description: The data sets used for this study include PubMed and MS2, which are well-established collections of medical literature. For fine-tuning, 200 relevant files along with their corresponding summaries were selected and organized into two folders: one containing the texts and the other containing the summaries.

PubMed is a comprehensive database of biomedical literature maintained by the U.S. National Library of Medicine. It provides access to millions of articles from journals, research papers, and reviews in the fields of medicine, biology, and health sciences. PubMed is widely used by researchers, clinicians, and students for evidence-based information, offering features like advanced search filters.

The MS² dataset is a large-scale, multi-document scientific summarization dataset designed for generating concise summaries of scientific literature. It contains thousands of clusters of related scientific papers, each with a human-written summary. MS² is a critical resource for advancing natural language processing (NLP) tasks, particularly in summarizing scientific documents. It addresses the challenges of condensing multiple sources into coherent and accurate summaries.

- **Average Document Length:** The documents in these datasets are on average 2,000 words long, with substantial variations depending on the type of document.
- **Average Summary Length:** The summaries generated from these documents typically span about 300 words.
- **Document Types:** The dataset comprises a variety of document types, including research papers, case studies, and clinical trials, all of which have structured abstracts suitable for training.

B. Model Formulation

The medical text summarization task is framed as a supervised learning problem, utilizing biomedical documents or abstracts (input) to generate concise, high-quality summaries (output). Pre-trained transformer models such as T5, BART, PEGASUS, and BigBird are evaluated for their summarization capabilities.

Initially, models are evaluated on both datasets using metrics such as ROUGE, BERTScore, and METEOR to assess their ability to retain critical information and produce semantically coherent summaries. The top-performing model is selected based on this evaluation and is subsequently fine-tuned using the curated dataset. Fine-tuning involves domain adaptation using techniques such as learning rate optimization, dropout regularization, and careful handling of domain-specific terminology to improve the model's summarization quality in biomedical contexts.

C. Base Models and Evaluation

The **base models** were selected on the basis of balancing state-of-the-art performance with domain-specific capabilities, particularly in handling the complexity and length of the biomedical literature. The chosen base models for the project are:

- **T5 (Text-to-Text Transfer Transformer):** Converts all NLP tasks into a text-to-text format, making it versatile for summarization tasks. Pre-trained on the Colossal Clean Crawled Corpus (C4), providing robust language understanding. In this study, the T5-large model was utilized via the Hugging Face pipeline to generate summaries from the input text using a maximum length of 450 tokens.
- **BART (Bidirectional and Auto-Regressive Transformer):** Combines bidirectional encoding and autoregressive decoding, ideal for text generation tasks like summarization. Fine-tuned on various summarization datasets, achieving state-of-the-art performance in many benchmarks. The Google PEGASUS-large variant of BART was employed to summarize the input text using the pipeline provided by the Transformers library.
- **PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization):** Optimized for summarization tasks by masking important sentences during pre-training to simulate summarization objectives. Particularly effective for long-form summarization due to its gap-sentence generation strategy. The Google BigBird-Pegasus-large-arxiv model was applied to biomedical text using a maximum token length of 512 and a minimum length of 100 to ensure adequate summary coverage.
- **BigBird:** Handles long sequences by introducing sparse attention mechanisms, making it efficient for processing lengthy biomedical texts. Supports domain-specific extensions like BigBird-PubMed, fine-tuned for biomedical literature. Both general-purpose BigBird-Pegasus and BigBird-PubMed models were used to summarize text, leveraging domain-specific training to capture key points from medical literature.
- **BigBird-PubMed:** Pre-trained explicitly on PubMed abstracts and full-text articles, offering a significant advantage in summarizing biomedical literature. It combines sparse attention mechanisms with domain-specific data to handle long biomedical documents effectively. In this study, the BigBird-Pegasus-large-PubMed model was applied to summarize biomedical texts using a maximum length of 200 tokens and a minimum length of 50 tokens, leveraging its pre-training on PubMed for domain-specific performance.
- **FLAN-T5 (Fine-tuned Language Net – T5):** Builds on T5 with fine-tuning across various tasks, including summarization, for better task generalization. Demonstrates strong performance in zero-shot and few-shot learning scenarios. Summaries were generated using the Flan-T5-large model with a truncation length set to 150 tokens for concise abstraction.
- **DistilBART:** DistilBART is a smaller, faster, and more efficient variant of BART (Bidirectional and Auto-Regressive Transformer) created through knowledge distillation. It retains BART's strong performance on natural language generation and comprehension tasks while reducing computational requirements, making it suitable for deployment in resource-constrained environments. The 'sshleifer/distilbart-cnn-12-6' variant was used to generate summaries with a maximum length of 200 tokens to balance conciseness and coverage.
- **LFED (Longformer Encoder-Decoder):** Utilizes a sliding window attention mechanism to manage long documents effectively. Suitable for handling extended biomedical abstracts and multi-document summarization tasks. The AllenAI Longformer Encoder-Decoder model was employed for summarizing input text with a maximum length of 512 tokens and a truncation strategy to handle lengthy biomedical documents.

The base models are evaluated on a sample medical text article by comparing their generated summaries with the expert-generated summaries in the domain. The evaluation is conducted using the following metrics:

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Measures the overlap between n-grams (e.g., unigrams, bigrams) or sequences in the generated summary and the reference summary. Variants include ROUGE-1, ROUGE-2, and ROUGE-L, which assess word-level, phrase-level, and longest common subsequence matches, respectively.
- **BERTScore:** Uses contextual embeddings from a pre-trained BERT model to evaluate the semantic similarity between the generated and reference summaries. Focuses on meaning rather than surface-level word overlap.
- **METEOR (Metric for Evaluation of Translation with Explicit ORdering)** Combines precision and recall with alignment strategies, including synonyms and paraphrasing, to evaluate the generated summaries. Provides a balanced assessment of fluency and accuracy.

After analyzing the evaluation metrics for the sample text, BigBird achieved the highest scores in all tests. This was unexpected, as BigBird PubMed was expected to perform the best. However, due to the inclusion of files from the MS2 dataset, BigBird outperformed BigBird PubMed, demonstrating better generalization.

D. Fine-Tuning The Best Performing Model - BigBird:

The fine-tuning of the BigBird model was carried out in two distinct iterations, each with its unique approach to optimization. In the first attempt, the model was trained using a straightforward methodology. A fixed number of three epochs were set, and the learning rate was initialized at 2×10^{-5}. Optimization was performed using the AdamW optimizer, with no additional regularization parameters

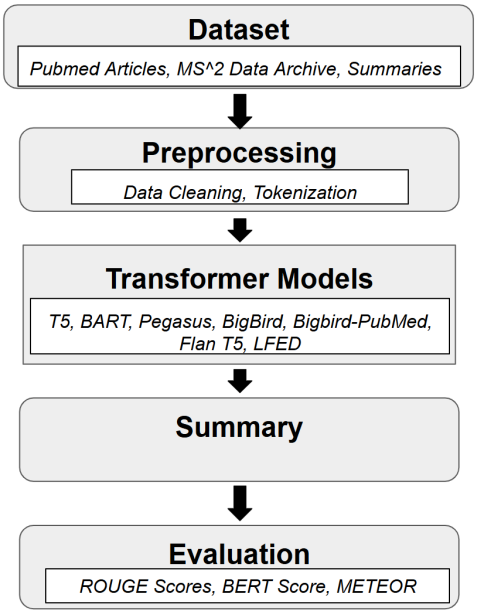

Fig. 1. Schematic Flow Diagram for Base Model Selection and Evaluation

such as weight decay. The training loop was kept simple, with each batch being directly passed through the model for loss computation. The gradients were calculated using backpropagation, and the weights were updated in each iteration. Despite its simplicity, this approach provided an initial insight into how fine-tuning could improve performance over the base model.

In the second attempt, a more sophisticated training regime was employed to further optimize performance. The number of epochs was increased to five, and gradient accumulation was implemented with an accumulation step of four. This allowed for an effective increase in batch size without exceeding GPU memory limitations. Additionally, a weight decay of 0.01 was applied to the optimizer to enhance regularization and prevent overfitting. A learning rate scheduler, OneCycleLR, was used to dynamically adjust the learning rate throughout training, ensuring a smoother convergence process. The use of these advanced techniques aimed to extract more performance improvements from the fine-tuning process.

Both training strategies utilized the curated dataset of PubMed and MS2 Files and their summaries. The input text was tokenized and padded to fit the model's requirements, and the labels were adjusted accordingly. In both cases, the model was fine-tuned to minimize the cross-entropy loss between predicted and actual token sequences. Each training loop involved transferring data to the GPU for computational efficiency, followed by a forward and backward pass through the model. However, the second attempt included optimizations such as gradient accumulation and dynamic

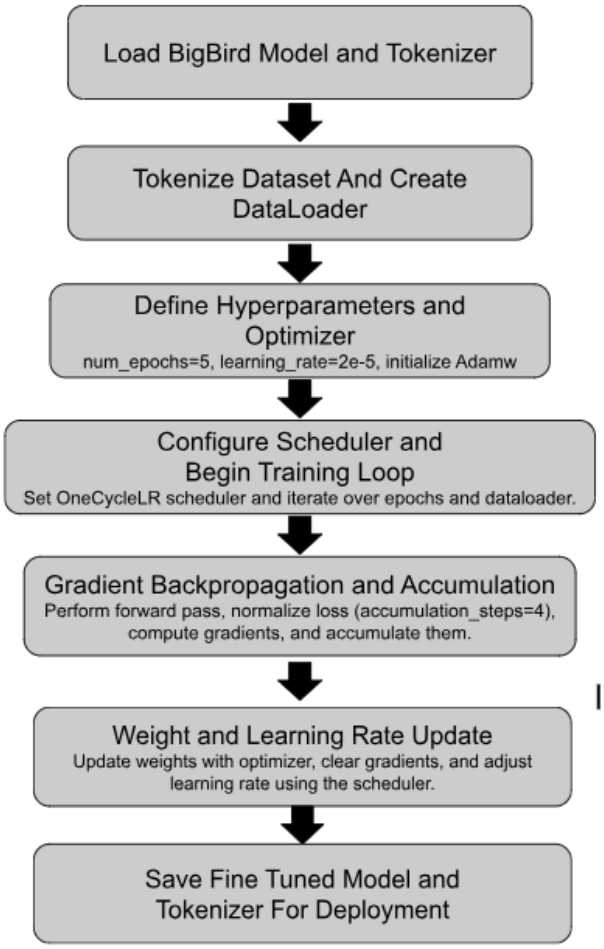

Fig. 2. Schematic Flow Diagram for Fine Tuning the Best Base Model - BigBird

learning rate adjustment, making it more suitable for scenarios involving limited computational resources.

Finally, both fine-tuned models were saved along with their respective tokenizers for evaluation. The outputs generated by these models were compared against the actual summaries in the medical dataset, and evaluation metrics such as ROUGE, BERT Scores and METEOR scores were calculated. These scores served as the primary benchmarks for assessing the quality of fine-tuning and determining the efficacy of each approach in generating accurate summaries.

III. RESULTS

A. ROUGE Score Evaluation:

After analyzing the evaluation metrics for the sample text, BigBird achieved the highest scores in all tests. This was unexpected, as BigBird PubMed was anticipated to perform the best due to its specialized pre-training on PubMed abstracts and full-text articles. However, the inclusion of files from the MS2 dataset, which represents a broader domain,

may have given BigBird's generalization capabilities an edge over BigBird PubMed. The sparse attention mechanism employed by BigBird proved particularly effective for processing lengthy biomedical texts, leading to the highest ROUGE-1 (0.523985) and ROUGE-Lsum (0.435424) scores among all tested models.

BigBird PubMed, while performing slightly below BigBird, still showcased impressive results with the second-highest scores in ROUGE-1 (0.480565) and ROUGE-Lsum (0.395760). This highlights its domain-specific strength in processing biomedical documents. The performance gap between BigBird and BigBird PubMed underscores the importance of dataset composition in determining model effectiveness. BigBird PubMed's pre-training on PubMed might have limited its flexibility when applied to the MS2 dataset, which could explain its relatively lower ROUGE-2 (0.120996) and ROUGE-L (0.226148) scores compared to BigBird.

Other models, such as LFED and BART, also performed reasonably well, with ROUGE-1 scores of 0.309198 and 0.272727, respectively. LFED's sliding window attention mechanism likely contributed to its capability to handle long documents effectively, while BART's bidirectional encoding and autoregressive decoding facilitated strong performance across summarization tasks. Meanwhile, T5, PEGASUS, FLAN-T5, and DistilBART exhibited lower scores overall, reflecting their relatively weaker performance on domain-specific biomedical data. These findings emphasize the trade-offs between generalization and specialization, with BigBird demonstrating superior adaptability across diverse datasets.

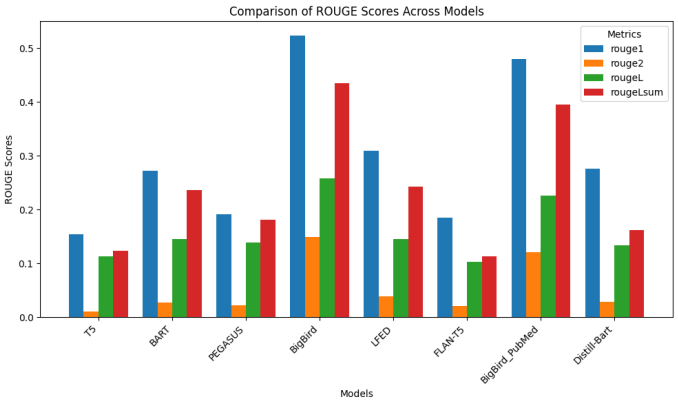

Fig. 3. Comparision of ROUGE Scores across Base Models

B. BERTScore Evaluation:

BigBird achieved the highest BERTScore F1 of 0.8763, indicating its superior ability to generate summaries closely aligned with the reference texts, likely due to its capacity to handle long sequences efficiently. BigBird_PubMed followed with a commendable F1 of 0.8654, showcasing its optimization for medical texts. In contrast, models like LFED and FLAN-T5 demonstrated relatively lower BERTScore F1 values of 0.8226 and 0.8383, respectively, reflecting limitations in their contextual understanding and sequence-to-sequence alignment for this domain.

Traditional transformer-based models such as T5, BART, and PEGASUS displayed consistent but moderate BERTScore F1 scores ranging from 0.8411 to 0.8417. Distill-Bart, with an F1 of 0.8424, demonstrated an effective balance between efficiency and performance, outperforming more computationally intensive models like LFED. These results emphasize the importance of model architecture and domain-specific pre-training, with BigBird and BigBird_PubMed standing out for their ability to capture the unique characteristics of scientific literature.

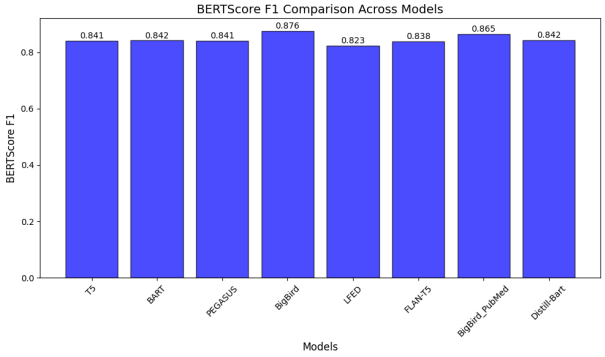

Fig. 4. Comparision of BERTScores across Base Models

C. METEOR Evaluation:

The METEOR scores for various abstractive text summarization models on PubMed data show clear distinctions in performance. BigBird achieved the highest METEOR score of 0.3637, demonstrating its ability to generate high-quality summaries that closely match the reference texts, likely due to its effective handling of long sequences. BigBird_PubMed also performed well with a score of 0.3191, highlighting the benefits of domain-specific optimization for medical content. In contrast, models like T5 and FLAN-T5 showed much lower METEOR scores (0.0727 and 0.0965), indicating their limitations in capturing the complexities of scientific text.

Other transformer-based models such as BART, PEGASUS, and Distill-Bart exhibited moderate METEOR scores between 0.1300 and 0.1394, suggesting they are suitable for general summarization tasks but struggle with specialized medical content. The results emphasize the importance of domain adaptation, with BigBird and BigBird_PubMed outperforming other models in the context of PubMed summarization. These findings reinforce the need for tailored models to achieve optimal performance in domain-specific tasks.

D. Comparision of Fine-Tuned Models with Base Model:

The results obtained from fine-tuning clearly demonstrated an improvement in ROUGE and METEOR scores over the base BigBird model. The base model without fine-tuning

TABLE I
ROUGE SCORES FOR DIFFERENT BASE MODELS USED FOR THE PROJECT.

Model	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
T5	0.153846	0.010363	0.112821	0.123077
BART	0.272727	0.027523	0.145455	0.236364
PEGASUS	0.191489	0.021505	0.138298	0.180851
BigBird	0.523985	0.148699	0.258303	0.435424
LFED	0.309198	0.039293	0.144814	0.242661
FLAN-T5	0.185567	0.020833	0.103093	0.113402
BigBird PubMed	0.480565	0.120996	0.226148	0.395760
Distill-Bart	0.276190	0.028846	0.133333	0.161905

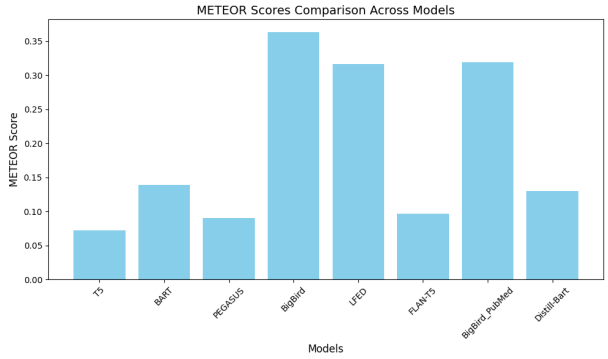

Fig. 5. Comparision of METEOR across Base Models

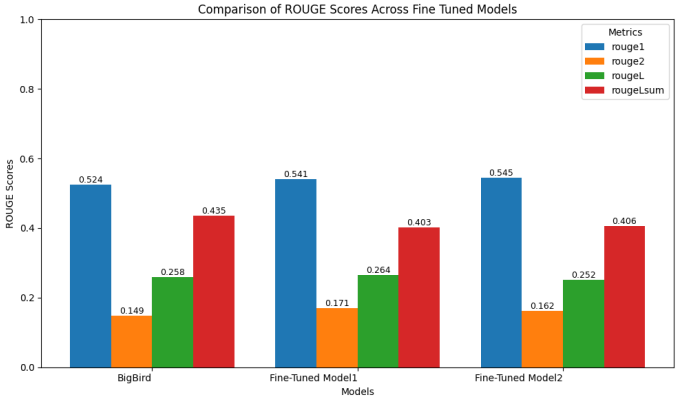

Fig. 6. Comparision of ROUGE Scores across Fine Tuned Models

achieved ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum scores of 0.523, 0.148, 0.258, and 0.435, respectively, alongside a METEOR score of 0.3637. Fine-tuned Model 1 exhibited significant enhancements, with ROUGE-1 improving to 0.540, ROUGE-2 to 0.170, and METEOR to 0.4125, indicating a marked increase in summarization quality. Fine-tuned Model 2 further increased ROUGE-1 to 0.545, reflecting enhanced coherence, though ROUGE-2 declined slightly to 0.161 and METEOR to 0.3537, suggesting a trade-off in precision for improved generalization.

The observed improvements in these evaluation metrics underline the effectiveness of fine-tuning in tailoring pre-trained models to domain-specific tasks. By aligning the model's weights with the medical dataset, fine-tuning significantly improved its ability to produce accurate and context-aware summaries. The advanced techniques utilized in Fine-tuned Model 2, including gradient accumulation and dynamic learning rate scheduling, contributed to stable training and enhanced generalization, further optimizing the model's performance.

Overall, these findings demonstrate the successful optimization of the base BigBird model through fine-tuning for the specific use case of generating high-quality summaries of medical documents. This highlights the transformative potential of domain-specific fine-tuning in leveraging pre-trained language models for specialized applications.

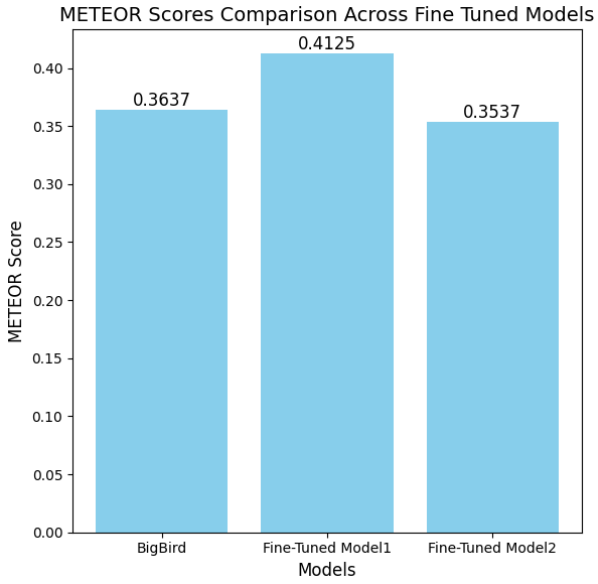

Fig. 7. Comparision of METEOR across Fine Tuned Models

IV. CONCLUSION

This research aimed to address the challenge of abstractive summarization in the domain of medical literature by evaluating and fine-tuning state-of-the-art transformer models. Among the various models explored, BigBird emerged as the best performer, demonstrating its capability to process lengthy biomedical texts effectively and generate concise, accurate summaries. While models like BigBird-PubMed showed strong domain-specific performance, the inclusion of a broader dataset like MS2 highlighted the generalization advantages of BigBird. These findings underscore the importance of selecting the right model architecture and dataset composition when tackling complex natural language processing tasks in specialized domains.

Fine-tuning played a pivotal role in enhancing model performance, with advanced techniques such as gradient accumulation, dynamic learning rate scheduling, and weight decay yielding significant improvements. The iterative approach to fine-tuning BigBird demonstrated the value of combining computational efficiency with optimization strategies to achieve superior results. Evaluation metrics, including ROUGE, BERTScore, and METEOR, validated the effectiveness of the fine-tuned model, establishing a benchmark for future efforts in medical literature summarization.

In summary, this study contributes to the growing body of research in biomedical natural language processing by presenting a robust methodology for abstractive summarization. The insights gained from the evaluation of multiple transformer models and the iterative fine-tuning process can guide future developments in the field. By enabling the generation of accurate and coherent summaries of complex medical documents, this research has the potential to significantly aid researchers, clinicians, and policymakers in accessing critical information more efficiently.

REFERENCES

[1] Vaswani, Ashish, et al. "Attention Is All You Need." Advances in Neural Information Processing Systems, 2017.

[2] Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint arXiv:1810.04805, 2018.

[3] Liu, Yinhan, et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv preprint arXiv:1907.11692, 2019.

[4] Raffel, Colin, et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." Journal of Machine Learning Research, 21(140), 1-67, 2020.

[5] Yang, Zhilin, et al. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." Advances in Neural Information Processing Systems, 32, 2019.

[6] Brown, Tom B., et al. "Language Models are Few-Shot Learners." Advances in Neural Information Processing Systems, 33, 1877-1901, 2020.

[7] National Center for Biotechnology Information. "PubMed." Hugging Face Datasets, 2024.

[8] DeYoung, Jay, et al. "MŜ2: Multi-Document Summarization of Medical Studies." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021.