Exploring the role of transformer-based language models in medical transcript summarization *

Mariia Zimokha[1,*,†], Kyrylo Yemets [2,†]

[1] Queen Mary University of London, London, United Kingdom
[2] Lviv Polytechnic National University, Lviv, Ukraine

Abstract

The increasing complexity of medical documentation necessitates effective summarization tools that enable healthcare professionals to swiftly access critical patient information. This study investigates the efficacy of four transformer-based models—T5, Flan-T5, GPT-4 Mini, and BART—in summarizing medical transcripts. We conducted a comprehensive comparative analysis utilizing various learning paradigms, including zero-shot learning, one-shot learning, and pretraining methods, to assess the models' capabilities. For evaluation, we employed a diverse set of metrics: ROUGE, METEOR, and BERTScore, providing a holistic view of each model's performance. Our findings indicate that the BART model consistently outperformed the others in summarizing medical texts, demonstrating superior fluency and coherence. In contrast, the GPT-4 Mini model exhibited notable flexibility in domain-specific fine-tuning through zero-shot learning. These results highlight the potential of leveraging transformer-based models to enhance the efficiency of medical documentation processes, ultimately contributing to improved patient care and clinical outcomes. This study underscores the importance of integrating advanced natural language processing techniques into healthcare practices to address the challenges posed by complex medical information.

Keywords

Transformer-based models, medical summarization, BART, GPT-4-mini, T5, Flan-T5, metrics

1. Introduction

Medical documentation consists of many parts and one of them is patient transcripts. This crucial part provides a healthcare professional quick access to important patient details. However, the biggest challenge is medical terminology, the models are required to comprehend domain-specific terminology and complex syntactic structures [1], [2], [3].

Recent developments in natural language processing (NLP) produced a big variety of transformer-based models, such as GPT-4-mini [4], Flan-T5 [5], and BART [6] which have shown effectiveness in summarization tasks due to self-attention mechanisms [7]. Despite their performance on general topics, their effectiveness in summarizing medical transcripts remains less explored.

This study investigates the performance of four transformer-based models—T5 [8], Flan-T5 [5], GPT-4-mini [4], and BART [6]—in summarizing medical transcripts [9]. To compare models' performance on medical data, the technics such as zero-shot learning, one-shot learning, and pretraining methods have been utilized. The research aims to address the following questions:

- How do these models perform in summarization tasks with and without instructions (Zero-shot vs Few-shot learning)?

IDDM'24: 7th International Conference on Informatics & Data-Driven Medicine, November 14 - 16, 2024, Birmingham, UK
* Corresponding author.
m.zimokha@se24.qmul.ac.uk (M. Zimokha), kyrylo.v.yemets@lpnu.ua (K. Yemets)
0009-0004-2795-6739 (M. Zimokha), 0000-0002-5157-9118 (K. Yemets)

- Does pretraining on domain-specific data enhance the performance of transformer models in medical transcript summarization?
- Which model fits the best for medical applications in terms of accuracy, privacy, etc?

By answering these questions, this paper aims to provide insights into how transformer-based models can be leveraged to improve clinical documentation efficiency.

2. Methodology

Over the last few years, the transformer models have become a go-to approach to tackle a variety of NLP problems due to their powerful architecture. It was introduced in the *Attention Is All You Need* paper (2017) [10], [11], and stated that it is possible to improve outcomes of NLP tasks by leveraging self-attention mechanisms to process sequential data. This mechanism allows us to capture long-range statements more efficiently than short-term memory networks (LSTM), recurrent neural networks (RNN), and deep neural networks (DNN) [12], [13].

The transformer architecture offers several advantages over classical architectures (LSTM, RNN, and DNN). By leveraging a self-attention mechanism instead of recurrence, the relationship among all tokens in the sequence can be captured simultaneously. That allows to preservation of the meaning of the long texts, hence, mitigates the vanishing gradient problem that RNN and LSTM can suffer. Also, due to transformer architecture, parallelization across tokens in a sequence can be performed, whereas RNNs are impossible. This parallel processing accelerates training and inference times, making transformers highly efficient, especially when scaled to large datasets and model sizes [15].

In this study, the effectiveness of the four transformer-based models has been evaluated: Flan-T5 [5], T5 [8], BART [6], and GPT-4 Mini [4]. For the given models the advantages and disadvantages will be discussed in more details.

T5 (Text-To-Text Transfer Transformer) is a model designed by Google [8] that uses an encoder-decoder architecture. T5 models have different sizes, from 60 million to 11 billion parameters [8]. As the model uses transformer architecture, it is leveraging self-attention mechanisms. However, it differs from other transformer models in terms of its text-to-text framework. Every task is converted to a consistent input text format using a span-corruption objective, where spans of text are masked for the model to predict [8]. Also, the model uses an encoder-decoder structure, compared to GPT models, which use just a decoder structure. The encoder captures the relationships in the text bi-directionally, and the decoder generates output autoregressively. One of the biggest disadvantages of this model is a context window - often limited to 512 tokens in certain variants, and for the longer text, specific techniques must be applied, such as chunking. However, it might lead to information loss due to a lack of coherence. Also, the model is more sensitive to noisy data, such as incomplete or inconsistent inputs, making it less efficient than BART, which was trained on noisy and corrupted data to increase robustness [6].

Flan-T5 is based on the T5 (Text-To-Text Transfer Transformer) architecture designed by Google [8]. Like T5, the Flan-T5 model uses an encoder-decoder architecture and leverages the self-attention mechanism to capture contextual relationships in the text. The Flan version builds on this foundation by fine-tuning the model on a broader range of instruction-based datasets, which leads to task-specific fine-tuning can be avoided due to instruction-based fine-tuning, such as zero-shot and few-shot learning [5]. The Flan-T5 model is available in multiple sizes, and The Flan-T5 Large variant has approximately 770 million parameters. Due to the base specific T5 model, Flan-T5 inherited the same disadvantages. The model has a context window limitation of 512 tokens [5], necessitating additional processing steps for long texts, the same as for the T5 model. Also, the current model is sensitive to noisy inputs, similar to the T5 model. Although the model's outcomes can be improved with instruction-based tunning (few-shot learning), the prompt's quality has a

significant impact on the model's outcome [5]. In domain-specific tasks, achieving desirable results can be challenging; if the prompt lacks coherence, the model's effectiveness can be reduced.

BART (Bidirectional and Auto-Regressive Transformers) is an encoder-decoder model developed by Facebook (Meta) [6]. It combines bidirectional encoding with autoregressive decoding, allowing for effective context processing in both directions and sequential generation. Leveraging this architecture makes BART well suited for tast where the text comprehension and generation is required, such as summarization. This model has been trained on noisy and corrupted data [6], hence, this model handle the noisy data better than some other transformer models, improving robustness and coherence of the output text. As a many other models, BART is available in several sizes, for example the base model has around 400 million parameters and a context window of 1024 tokens, making it suitable for summarization tasks. Despite that the context window is bigger than T5 based model, for larger texts might not fit into the window, the same methods can be utilized such as chunking with the same implications. The same as other transformer models, BART can be resource-intensive, requiring computational power for training and inferencing.

GPT-4 Mini is part of the Generative Pre-trained Transformer family developed by OpenAI [4]. It is known for its autoregressive capabilities and uses a decoder-only architecture that generates output by predicting the next token, taking into account the previous token. This architecture allows the model to excel in zero-shot and few-shot learning scenarios. The model has been trained on a vast general dataset, so the model performs quite well on a variety of NLP tasks without pretraining. Also, the model can be adapted for domain-specific tasks using prompting. However, the quality of the outcome will depend on the quality of the prompt. GPT-4 Mini has a context window limitation of 2048 tokens, and similar chunking techniques can be employed in the case of longer texts. Like other transformer models, the GPT-4 mini is computationally intensive[6].

3. Modeling and results

3.1. Dataset descriptions

This experiment utilized the **Medical Transcriptions** dataset from Kaggle, encompassing various medical transcripts [9]. Each transcript includes patient medical information shared during doctor appointments, such as medical history, examination notes, and medications. The dataset comprises *4,999* rows, focusing on the *transcription* and *keywords* columns. After removing rows with missing transcription or keywords, the dataset contains *3,931* rows for evaluation.

3.2. Performance indicators

The models—Flan-T5, BART, GPT-4 Mini, and T5 fine-tuned for the task of summarizing medical text —were evaluated using several metrics to assess the alignment of generated summaries with original transcripts [9]:

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [1]:
  - ROUGE-1: Measures unigram overlap between generated summaries and references.
  - ROUGE-2: Measures bigram overlap between generated summaries and references.
  - ROUGE-L: Assesses the Longest Common Subsequence (LCS) to evaluate content order.
- METEOR (Metric for Evaluation of Translation with Explicit ORdering) [2], [14]: Compares generated texts to references through exact, stemmed, and synonym matches, emphasizing precision and recall.
- BERTScore [15]: Based on the BERT model, it computes cosine similarity between generated text and references using BERT embeddings:

- BERT Precision: Measures the precision of the generated text based on semantic similarity.
- BERT Recall: Indicates how many relevant tokens in the reference were captured.
- BERT F1: The harmonic mean of BERT Precision and Recall, offering a balanced performance measure.

3.3. Results

The assessment of the described metrics was performed on the Kaggle Medical Transcriptions Dataset [9]. Table 1 contains the results of all of the evaluations, to evaluate the quality of the generated summaries.

Table
Evaluation Metrics for Models

Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	BERT Precision	BERT Recall	BERT F1
BART	0.5944	0.5372	0.4575	0.3861	0.8314	0.7623	0.7937
Flan-T5 One Shot	0.2238	0.1966	0.2102	0.1117	0.7916	0.5042	0.6120
GPT-4 mini Zero Shot	0.3688	0.2191	0.2439	0.2149	0.6460	0.6135	0.6282
T5 Pretrained	0.3098	0.2375	0.2427	0.1537	0.7566	0.5876	0.6592

4. Evaluation and discussion

4.1. Evaluation

To evaluate the efficiency of the provided models during the summarization of the medical transcript, let's take a look at the results in Table 1. For better understanding and comparison, the data has been represented in bar charts.

Figure 1 illustrates the ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L) for four models. The BART model dominates in all scores: ROUGE-1 reaches almost 60%, ROUGE-2 - around 54%, and ROUGE-L - around 46%. It can be explained by BART architecture (bidirectional encoder and autoregressive decoder), which makes it well suited for tasks such as summarization. Also, the model was trained on the noisy dataset, which helps generate a more coherent summary with sentence structure. It is reflected in the ROUGE sources, especially ROUGE-L. GPT-4 mini in zero-shot learning demonstrates similar performance as pre-trained on medical data T5. ROUGE-1 reaches almost 38%, and 30% is for GPT-4 mini and T5, respectively. However, T5 outperformed GPT-4 mini in ROUGE-2 (almost 24% and 22% respectively). Those numbers indicate that although GPT-4 mini shows great performance, it may lack understanding of medical language without domain-specific fine-tuning (few-shot learning). On the other hand, T5 was pre-trained on medical data and can comprehend medical terms better, which has been reflected in the scores. The Flan-T5 performed around 22%, 20%, and 21% for ROUGE-1, ROUGE-2, and ROUGE-L, respectively. Those numbers are not far from the T5 and GPT models, which might indicate that the model can benefit from an improved prompt or better chunking strategy.

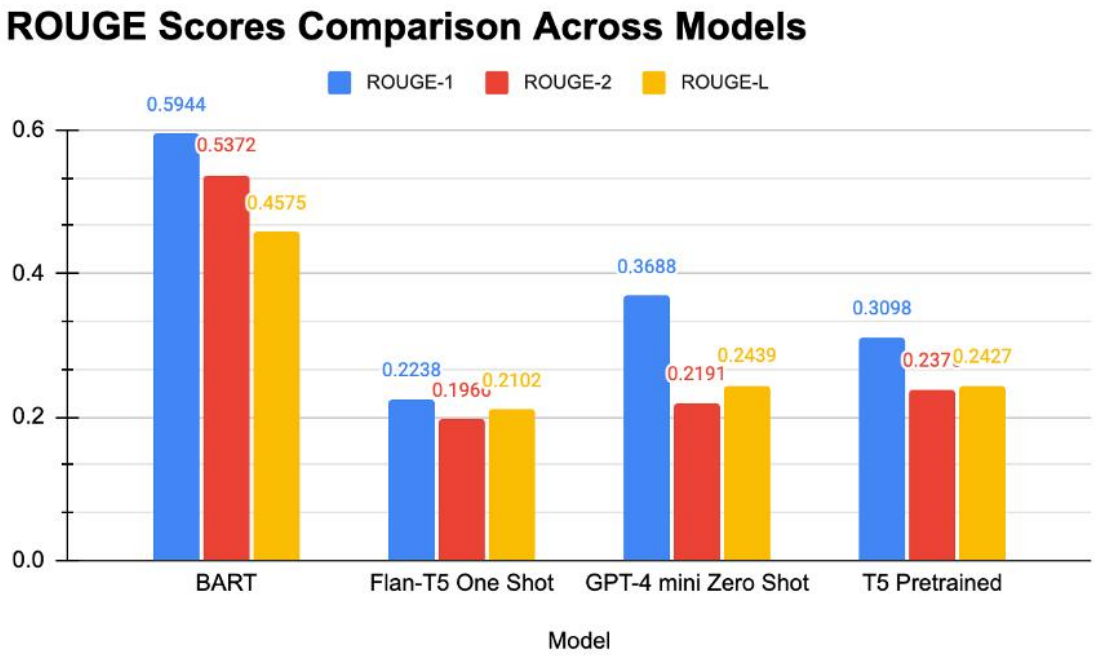

Figure 1: ROUGE Scores Comparison Across Models

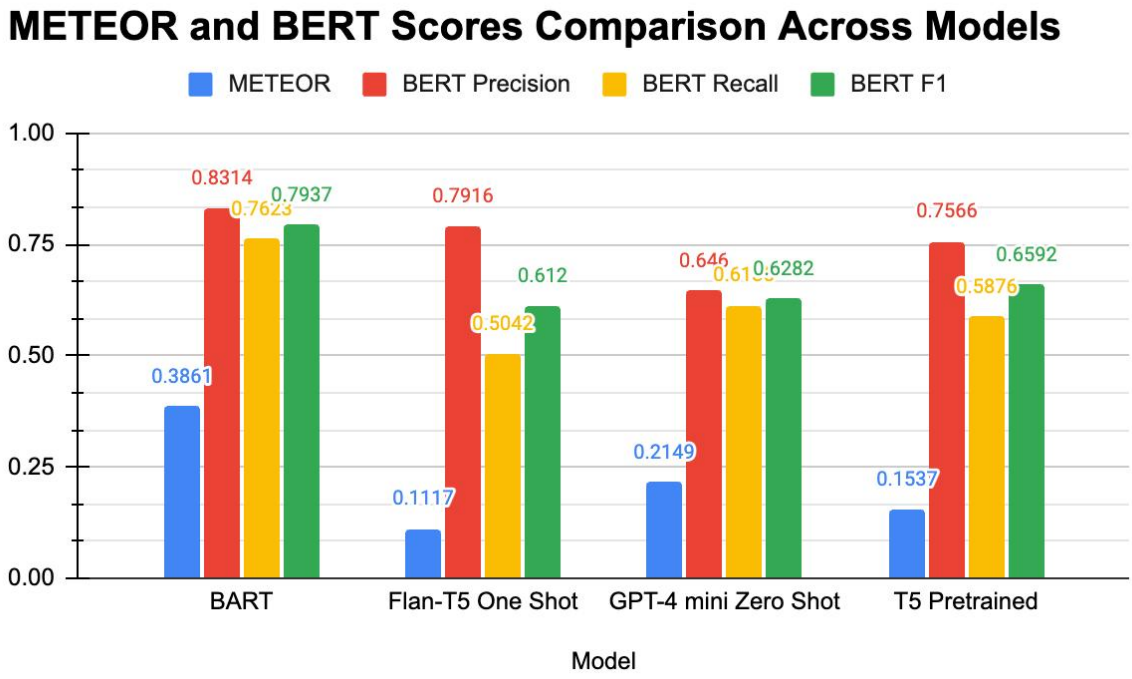

Figure 2: METEOR and BERT Scores Comparison Across Models

Figure 2 illustrates the model performance in METEOR and BERTScore. BART's model showed the highest scores across all metrics, around 38%, 83%, 76%, and 79% in METEOR, Bert precision, Bert recall, and Bert f1, respectively. These results align with the model architecture and its pretraining on noisy data. GPT-4 mini has achieved around 21% in METEOR and has demonstrated moderate but consistent BERT scores across precision, recall, and f1 (around 61%, 62%, and 62%, respectively). These scores indicate that the model performs reasonably well. However, it can

benefit from medical fine-tuning. Flan-T5 achieved 11% in METEOR, the lowest score among the models, but moderate Bert scores, with about 79% in precision, 50% in recall, and 61% in F1. As in the GPT model, these results suggest that the model might benefit from domain-specific fine-tuning. T5, pre-trained on medical data, showed METEOR at 15% and BERT scores of approximately 76% in precision, 58% in recall, and 66% in F1. The Bert F1 and recall scores suggest that the model understands the medical language due to its pretraining.

4.2. Discussion

The evaluation results illustrate significant differences among the transformer models in their summarization capabilities.

- **BART** outperformed Flan-T5, GPT-4 Mini and T5 fine-tuned model across all metrics, highlighting its superior ability to capture the essence of medical transcripts.
- **Flan-T5** demonstrated a notable decrease in performance in zero-shot settings compared to one-shot scenarios, indicating a greater reliance on task-specific data.
- **GPT-4 Mini** showed moderate performance, especially in zero-shot learning, suggesting that while it can generate plausible summaries, it lacks the contextual precision of BART.
- **T5 pretrained** demonstrated solid performance, comparable to GPT4 Mini, and even outperformed it in terms of the BERT F1 score.

The comparative analysis emphasizes the importance of model selection based on the intended application. Models such as BART and T5 are preferable for the tasks requiring a domain- specific knowledge. Due to their size, they can be fine-tuned efficiently using **LoRA** (Low-Rank Adaptation of Large Language Models), which accelerates iteration while preserves prior knowledge from previous training.

Although a variety of metrics can help determine the best model for summarization, ithe most appropriate model may vary depending on the domain-specific task. For medical documentation summarization, the main goal is not only to preserve semantic meaning but inclusion of exact words, such as diagnosis, medications, etc. To have a better understanding of models performance, the keyword-based recall metric was evaluated. It was calculated based on the keywords column, which contains the important terms in transcription and generated summaries. From the provided data from Table 2, BART and GPT-40-Mini performed quite well.

Table 2
Recall Scores of Keywords Across Models

Model	Recall Score
BART	0.6504
GPT Mini	0.6409
T5 Pretrained	0.377
Flan-T5	0.2986

4.3. Challenges

When summarization is needed, several factors must be considered, such as budget, computational resources, privacy, etc. Although this study was conducted using publicly available data and privacy was not a concern, this is not the case for real applications. The GPT model is perfect for the instruction-based use case and is easy to use; however, constructing effective instructions (prompt) can be challenging, especially for large documents, which require specific strategies for processing. Additionally, since this model is hosted by OpenAI, there is a risk of data leakage.

On the other hand, the BART model is open for use and can be pretrained/fine-tuned for specific tasks. However, it is the developer's responsibility to host it, and in some cases, it is preferable as PHI (Protected Health Information) data can be restricted to specific locations or hardware, etc.

5. Conclusion

This study reveals the potential of transformer-based language models in medical transcript summarization. Accurate summarization allows healthcare providers to have quick access to essential patient information, and automated summarization tools can improve patient care. The ability to generate an accurate summary allows the professionals to focus on a patient rather than on the documentation. In domain-specific areas, it is essential to choose/train the model most suitable for the task.

In this study, four transformer-based models (BART, T5, Flan-T5, and GPT-4 Mini) were evaluated by the ability to create an efficient medical summary of the transcripts. They were chosen due to their architecture and proven success track with NLP tasks. To evaluate them, the Medical Transcriptions Dataset from Kaggle and multiple metrics were used to assess the performance. The ROUGE and METEOR use recall in their calculations, and they are useful for tasks where the capture of the core content is prioritized over fluent language production (e.g., diagnoses and medications). Also, the BERTScore metric was incorporated to capture semantic similarity. These metric combinations provide a great view of the model's capabilities to produce accurate and relevant summaries.

Collected results showed that BART outperformed others across all metrics. GPT-4 mini demonstrated moderate results utilizing a zero-shot learning technique with weaker medical comprehension than T5, which was pre-trained on medical data. While Flan-T5 is a versatile model, it displayed low performance because it has been trained on instruction-based datasets. These results demonstrate how important it is to pre-train/ fine-tune the model for domain-specific tasks to obtain the best outcome.

The medical NLP is a field of ongoing research. The industry has models such as MedPalm or Amazon Comprehend Medical. However, they are neither publicly accessible nor free. Hence, alternative possibilities should be explored. Based on this study, further exploration can be performed, such as investigating prompt engineering strategies for models like Flan-T5 and GPT-4 mini or pretraining using the LoRA (Low-Rank Adaptation) technique. Also, additional metrics can be defined for evaluation depending on the nature of the specific requirements.

Declaration on Generative AI

While writing this work, the authors used Grammarly to ensure that the language, including the spelling, is coherent and grammatically correct and ChatGPT to check facts with the provided bibliography. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the publication’s content.

References

[1] C.-Y. Lin, ‘ROUGE: A Package for Automatic Evaluation of Summaries’, in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. Accessed: Nov. 02, 2024. [Online]. Available: https://aclanthology.org/W04-1013

[2] M. Denkowski and A. Lavie, ‘Meteor Universal: Language Specific Translation Evaluation for Any Target Language’, in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, Maryland, USA: Association for Computational Linguistics, 2014, pp. 376–380. doi: 10.3115/v1/W14-3348.

[3] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, 'BLEU: a method for automatic evaluation of machine translation', in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, in ACL '02. USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. doi: 10.3115/1073083.1073135.

[4] OpenAI *et al.*, 'GPT-4 Technical Report', Mar. 04, 2024, *arXiv*: arXiv:2303.08774. Accessed: Nov. 02, 2024. [Online]. Available: http://arxiv.org/abs/2303.08774

[5] H. W. Chung *et al.*, 'Scaling Instruction-Finetuned Language Models', Dec. 06, 2022, *arXiv*: arXiv:2210.11416. Accessed: Nov. 02, 2024. [Online]. Available: http://arxiv.org/abs/2210.11416

[6] M. Lewis *et al.*, 'BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension', Oct. 29, 2019, *arXiv*: arXiv:1910.13461. Accessed: Nov. 02, 2024. [Online]. Available: http://arxiv.org/abs/1910.13461

[7] E. J. Hu *et al.*, 'LoRA: Low-Rank Adaptation of Large Language Models', Oct. 16, 2021, *arXiv*: arXiv:2106.09685. Accessed: Nov. 02, 2024. [Online]. Available: http://arxiv.org/abs/2106.09685

[8] C. Raffel *et al.*, 'Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer', Sep. 19, 2023, *arXiv*: arXiv:1910.10683. Accessed: Nov. 02, 2024. [Online]. Available: http://arxiv.org/abs/1910.10683

[9] 'Medical Transcriptions'. Accessed: Nov. 02, 2024. [Online]. Available: https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions

[10] A. Vaswani *et al.*, 'Attention Is All You Need', 2017, *arXiv*. doi: 10.48550/ARXIV.1706.03762.

[11] A. Vaswani *et al.*, 'Attention Is All You Need', Aug. 02, 2023, *arXiv*: arXiv:1706.03762. Accessed: Nov. 02, 2024. [Online]. Available: http://arxiv.org/abs/1706.03762

[12] K. Yemets and M. Gregus, 'A Transformer-based time series forecasting model with an efficient data preprocessing scheme for enhancing wind farm energy production', *Bulletin of Electrical Engineering and Informatics*, p. (in press), 2024.

[13] K. Yemets and M. Gregus, 'Schedule-Free Optimization of the Transformers-based Time Series Forecasting Model', *IAES International Journal of Artificial Intelligence (IJ-AI)*, p. (in press), 2024.

[14] 'NLTK :: nltk.translate.meteor_score module'. Accessed: Nov. 02, 2024. [Online]. Available: https://www.nltk.org/api/nltk.translate.meteor_score.html

[15] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, 'BERTScore: Evaluating Text Generation with BERT', Feb. 24, 2020, *arXiv*: arXiv:1904.09675. Accessed: Nov. 02, 2024. [Online]. Available: http://arxiv.org/abs/1904.09675