Analysis of Transformer based Pre-trained Summarization Models

Prof. Manisha Mathur
Assistant Professor, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
manisha.mathur@ves.ac.in

Manraj Singh Virdi
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.manrajsingh.virdi@ves.ac.in

Nikhil Dhanwani
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.nikhil.dhanwani@ves.ac.in

Chirag Santwani
Student, Computer Engineering
Vivekanand Education Society's Institute of Technology
Mumbai, India
d2021.chirag.santwani@ves.ac.in

***Abstract*—In order to efficiently retrieve and comprehend large volumes of information, automated text summarization is essential in compressing the material into brief and understandable summaries. In this study, we present a thorough comparison of six innovative transformer-based summarization models obtained from the Hugging Face model hub. Using ROUGE scores to evaluate sample outputs in both quantitative and qualitative terms, we carefully assess each model's performance and capabilities, clarifying its unique advantages, disadvantages, and complexities. Our results show significant variations in the models' summarization quality and effectiveness, which are probably caused by differences in the models' architectures, pre-training sets, and fine-tuning techniques. Our objective is to provide both students and professionals with the knowledge they need to choose carefully when choosing models for various summarizing jobs by explaining these insights and creating a foundation for future developments in transformer-based text summarization techniques.**

***Index Terms*—Automated text summarization, Transformer-based, ROUGE scores, Model performance, Pre-training sets, Fine-tuning techniques, Text summarization techniques.**

I. Introduction

In the contemporary digital landscape, the proliferation of information accessible at our fingertips has surged to unprecedented levels, necessitating the advancement of robust information processing and comprehension mechanisms. Text summarization emerges as a vital solution to this challenge, adeptly organizing and condensing extensive datasets into coherent summaries, while preserving key intricacies and nuances. The escalating demand for swift and efficient information retrieval further underscores the pivotal role of summarized text in mitigating the effects of information overload.

The field of deep learning and natural language processing has developed so rapidly that models that have been trained are now freely accessible and helpful for many NLP applications, especially text summarization. However, with so many models available—from BERT to T5—selecting the optimal pre-trained model for a specific job might be challenging. In addition, while pre-trained models offer an excellent place to begin, they can be even more effective and adaptable by modifying them with data specific to a certain area.

The purpose of this research is to do a thorough evaluation of 6 innovative text summarization models that are derived using the Hugging Face model hub [1] in consideration of the above challenges. Specifically, the models under scrutiny include Facebook's BART-Large-CNN [2], Falconsai's Text-Summarization [4], MBZUAI's LaMini-Flan-T5-248M [5], sshleifer's DistilBART-CNN-12-6 [7], MBZUAI's LaMini-T5-61M [8], and Tuner007's Pegasus_summarizer [9]. We want to clarify each model's advantages, disadvantages, and possibilities through comprehensive evaluation utilizing recognized metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores [10] and detailed examination of sample outcomes. We also want to investigate the efficacy of fine-tuning techniques in improving these models' performance on text summarizing tasks. By conducting this comparison evaluation, we want to offer practical advice to professionals and scholars on how to efficiently navigate the field of summarized text models and optimize their potential for handling information overload.

II. Background

Text summarization stands as a fundamental task within the realm of Natural Language Processing (NLP), addressing the challenge of condensing extensive volumes of textual data into succinct and coherent summaries. With an exponential surge in the creation of research papers, news articles, social media posts, and various digital materials, the need for effective summarization methodologies becomes increasingly paramount. Beyond mere condensation, text summarization serves as a facilitator for diverse tasks such as document classification, outcome ranking, and tailored content suggestion.

A. Approaches to Text Summarization

There are two primary approaches for text summary in natural language processing: extractive and abstractive summarization. In extractive summarizing [13], the precise wording and organization of the original information is maintained by picking out significant lines or paragraphs from the text as a whole and then presenting them in the form of a summary. In order to pick the most useful words based on factors including relevance, significance, and repetition, this method uses algorithms. Abstractive summary [13], on the other hand, involves developing new content that clearly and logically communicates the key points of the original text. Abstractive approaches clarify and paraphrase text using natural language comprehension and production processes, which frequently provide summaries that are more concise and logical.

B. Large Language Models (LLMs)

Among Large Language Models (LLMs) [14], Transformer-based models such as BART, GPT and T5 have transformed text summarization by their capacity to recognize complex language patterns and meaningful connections. This is because they have been pre-trained on huge amounts of text data. These models show exceptional understanding and generation skills for human-like writing, which makes them useful for a wide range of tasks related to natural language processing, including summarization. By using the encoded knowledge stored inside its boundaries, LLMs are able to efficiently compress complicated documents while maintaining crucial information and context. Analyzing automatically generated summaries is still difficult, though. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores [12] are frequently utilized metrics; however, they have several drawbacks, like their insensitivity to semantic similarity and their incapacity to capture readability and overall consistency. Thus, human assessment continues to be necessary for gaining an in-depth knowledge of summary quality, completing automated measures such as ROUGE and opens the door for future advances in the field.

C. Fine Tuning Models

In addition, optimizing Large Language Models (LLMs) on datasets customized to certain tasks offers an appealing method to improve their summarization performance. By significantly altering the pre-trained model's parameters, fine-tuning allows it to adapt its enormous prior knowledge to the specifics of a given summarization job or topic. It has been shown that this procedure significantly improves performance, helping the model to provide summaries that are more precise, relevant, and consistent. A crucial topic of investigation in our work is the possibility of fine-tuning LLMs for summarization tasks, especially since we are seeking to expand the boundaries of the field by optimizing model performance on particular dataset.

III. Methodology

In this section, we describe the thorough methods used to assess and compare the efficacy of six cutting-edge text summarization algorithms. Our approach aims to evaluate how well each model generates concise and cohesive summaries in various domains. We outline the criteria for model selection, describe the datasets utilized for evaluation and any fine-tuning processes, and specify the evaluation metrics adopted to measure summarization quality. To enhance flexibility and consistency of our results, we also provide the experimental environment, including system data and computational capabilities. This technique attempts to investigate the possible advantages of fine-tuning to enhance model performance in along with highlighting the relative benefits and drawbacks of each model. We want to make a significant contribution to the field of automated text summarizing through the implementation of a systematic and rigorous evaluation approach that will direct further studies and their application innovation.

A. Model Selection

Text summarization tools capabilities and performance are significantly affected by their model architecture. Among the many different designs that are accessible, BART (Bidirectional and Auto-Regressive Transformers) [16] a transformer-based model noted for being flexible to natural language processing applications. It can efficiently collect context and produce coherent text because to its bidirectional architecture and autoregressive decoding. T5 (Text-To-Text Transfer Transformer) [17], designed by Google AI, uses a unified architecture in which the inputs and outputs are expressed as text strings, making it suitable for a variety of NLP applications. FLAN T5 [18] a version of T5 that has been specially optimized for text summary, providing greater efficiency in producing concise and insightful summaries. Another Google AI model, Pegasus [19], is well known for its abstractive summarization skills. It can efficiently extract important information from source texts and reword or rephrase it. Several models utilize these architectures, including 'facebook/bart-large-cnn' based on BART, 'MBZUAI/LaMini-T5-61M' based on T5, 'MBZUAI/LaMini-Flan-T5-248M' based on FLAN T5, and 'tuner007/pegasus_summarizer' based on Pegasus.

Selected Models

- **Model 1: facebook/bart-large-cnn**
  This model, as cited in [3], created by the research team at Facebook AI, is an effective tool for text summarization. It is based on the robust design of BART (Bidirectional and Auto-Regressive Transformers). The architecture allows the model to understand complicated language connections and structures observed in textual data. Designed specifically for news item overviews, 'facebook/BART-Large-CNN' is excellent at collecting important texts, which reduces the amount of information that is consumed. By carefully tuning with the CNN Daily Mail dataset—a comprehensive collection of text-summary pairs—the algorithm achieves outstanding results on summarizing tasks.
- **Model 2: Falconsai/text_summarization**

Referred from [4], designed by Falcon AI, it is aimed at the medical text summarizing industry. Based on a foundation pre-trained on a wide range of medical literature, this innovative approach makes use of an optimized version of the T5 Large transformer model. The model is excellent at producing concise and coherent summaries that are appropriate to the complexities of the medical area because it focuses on medical texts, such as research papers, clinical notes, and documents. By applying a thorough method of fine-tuning on a composite dataset gathered from various fields, the 'falconsai/text_summarization' model ensures its versatility and efficacy in a wide range of tasks.

- **Model 3: MBZUAI/LaMini-Flan-T5-248M**
A refined variant of Google's Flan-T5-base model, called 'MBZUAI/LaMini-Flan-T5-248M' [6], is optimized to work with the LaMini-instruction dataset, which contains about 2.58 million instruction samples. It provides compact, effective language models with a range of sizes, checkpoints, and structures as part of the LaMini-LM series. In contrast to other models in the series, the 'MBZUAI/LaMini-Flan-T5-248M' utilizes an encoder-decoder architecture with 248 million parameters, that is based on the T5 model. Its adaptability and usability in a variety of environments are guaranteed by its rigorous evaluations across several NLP tasks and human assessments, demonstrating its flexibility to a wide range of natural language processing activities, including text generation.
- **Model 4: sshleifer/distilbart-cnn-12-6**
The 'sshleifer DistilBART CNN 12-6' [7] model is a pre-trained language model designed for text summarization tasks based on the CNN/DailyMail dataset. It is a compressed and simplified variation of the BART model. Using a distillation process, DistilBART transfers the knowledge and performance of the bigger BART model by teaching a smaller student version to behave like it. Specifically designed for text summarization, the 'distilbart-cnn-12-6' variation uses a Convolutional Neural Network (CNN) encoder architecture with 12 layers of encoder and 6 decoder levels. This architecture improves the model's summarizing abilities by assisting in the acquisition of local contextual information in the input text.
- **Model 5: MBZUAI/LaMini-T5-61M**
A key part of the LaMini-LM series, the 'MBZUAI/LaMini-T5-61M' [8] model was carefully developed by the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI). Optimized for instruction fine-tuning, this version of the T5-small model has been developed on the LaMini-instruction dataset, an extensive collection of 2.58 million samples. This model, which has 61 million parameters overall, balances both efficiency and performance, which makes it a great choice for a range of natural language processing applications. For this study, we have chosen 'MBZUAI/LaMini-T5-61M' because of its stable structure, tailored instruction using domain-specific data, and alignment with the goals stated in the research report "LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions." [6]. This model is a good fit for our summarization experiments because it provides a strong combination of computational effectiveness and task-specific expertise.
- **Model 6: tuner007/pegasus_summarizer**
The pegasus_summarizer [9] is a text summarizing model that is available to the public and has been developed to maximize the potential of Google AI's Pegasus architecture. Since Pegasus serves as its foundation, it undoubtedly specializes at abstractive summarization and is skilled at utilizing various words and phrasing to express the main idea of the source material. A refined version of Pegasus, the 'tuner007/pegasus_summarizer' model is designed to improve performance on text summarizing work. It seeks to further improve summarization abilities by optimizing the original Pegasus model, giving consistency and informativeness in produced summaries.

TABLE I
OVERVIEW OF TEXT SUMMARIZATION MODELS

Model Name	Description	Architecture	Size
facebook/bart-large-cnn	BART with CNN encoder, fine-tuned on CNN Daily Mail dataset	Encoder-Decoder (BART) with CNN encoder	406M params
Falconsai/ text_summarization	Fine-tuned T5 Large for medical text summarization	Transformer (T5)	60.5M params
MBZUAI/LaMini-Flan-T5-248M	Fine-tuned Flan-T5-based on Lamini-instruction dataset	Transformer (Flan-T5)	248M params
sshleifer/distilbart-cnn-12-6	Distilled BART with CNN encoder for efficient summarization tasks	Encoder-Decoder (Distilled BART) with CNN encoder	306M params
MBZUAI/LaMini-T5-61M	Compact T5 variant for summarization	Transformer (T5)	61M params
tuner007/ pegasus_summarizer	An optimized model based on the powerful Pegasus architecture	Pegasus-based model	-

Table I provides a summary of the models that were examined in this study. After providing an overview of each model, we go into the process of preparing the data, describing the dataset that was used for assessment and any processing measures that were taken to ensure consistency and precision in the evaluation of the model's performance.

B. Dataset

SAMSum Dataset: The SAMSum dataset [20] includes around 16,000 messenger-like conversations with summaries for each. These dialogues were created by English-speaking linguists who were given the task to create interactions that matched their regular talks. The goal of the dataset is to replicate the subject matter distribution seen in actual messenger chats. Discussions display a broad range of records and styles, including semi-formal, formal, and casual conversations. The conversations were then annotated with brief summaries that were meant to capture the key points addressed in the dialogues. The Samsung R&D Institute Poland created the SAMSum dataset, which is licensed under a non-commercial agreement (CC BY-NC-ND 4.0) for use in research.

Data Splits:

- Train: 14,732 conversations
- Validation: 818 conversations
- Test: 819 conversations

Sample Instance:

- ”id”: ”13818513”,
  ”dialogue”:
  - ”Amanda: I baked cookies. Do you want some? Jerry: Sure! Amanda: I'll bring you some tomorrow.”

  ”summary”:
  - ”Amanda baked cookies and will bring Jerry some tomorrow.”,

C. Evaluation Strategy

In our research, the main criteria utilized to evaluate the level of quality of the produced summaries was ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores [11]. The ROUGE set of measures assesses the difference between the generated summary and a collection of prior summaries. We employed the ROUGE algorithm's versions, ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and ROUGE-L (largest common sub sequence), to test accuracy, recall, and content overlap specifically.

1) ROUGE-1 (Unigram Overlap): This metric calculates the amount of overlap between the reference summaries and the generated summary for unigrams, or single words. It evaluates the recall as well as accuracy of unigrams, giving information on the level of content overlap among the references and the generated summary. Greater agreement between the generated and reference summaries in terms of individual words is indicated by higher ROUGE-1 scores.
2) ROUGE-2 (Bigram Overlap): This version of the evaluation expands to take into account bigram overlap—sequences of two consecutive words—between the reference summaries and the produced summary. ROUGE-2 offers a more comprehensive assessment of subject matter overlap than ROUGE-1 since it records the occurrence of word sequences. Better agreement between the produced and reference summaries' word sequences is indicated by higher ROUGE-2 scores.
3) ROUGE-L (Longest Common Subsequence): ROUGE-L calculates the longest common subsequence. In contrast to ROUGE-1 and ROUGE-2, which emphasize on word matches exactly, ROUGE-L takes into consideration the longest continuous word sequence that shows up in both the reference summaries and the produced summary. This metric is particularly useful in determining the overall overlap of information and the degree of conceptual similarity between summaries, irrespective of word order.

Sample Instance: We calculated ROUGE scores for a sample generated summary using the following metrics:

- ROUGE-1 (Unigram Overlap):
  1) Recall (R): 0.9
  2) Precision (P): 0.5625
  3) F1-score (F): 0.6923076875739645
- ROUGE-2 (Bigram Overlap):
  1) Recall (R): 0.7777777777777778
  2) Precision (P): 0.30434782608695654
  3) F1-score (F): 0.4374999959570314
- ROUGE-L (Longest Common Subsequence):
  1) Recall (R): 0.9
  2) Precision (P): 0.5625
  3) F1-score (F): 0.6923076875739645

Although ROUGE scores offer important insights into how effectively summarization models work, it's essential to understand their limitations. ROUGE scores can fail to accurately represent the semantic equality between the generated summary and the reference summaries, which is one of their major drawbacks. They are also sensitive to semantic similarity. Furthermore, factors like coherence, readability, and fluency—all important variables in evaluating the general caliber of summaries—are not taken into consideration by ROUGE scores. Thus, in order to get an in-depth understanding of summary quality, ROUGE scores—while giving useful quantitative measurements—should be combined with qualitative evaluation and human judgment.

D. Fine-Tuning

In natural language processing (NLP), fine-tuning [15] is the process of altering a language model that was previously trained to fit a specific assignment or dataset by training it with data particular to the task. With this method, the model can gain task-specific patterns and complications while utilizing its previous expertise, which improves performance on the target job.

Motivation for Fine-Tuning: The need for adapting pre-trained models to particular datasets or tasks pushes fine-tuning. It's possible that generic pre-trained models fail to perform at the highest level across every domain. Customization is made feasible by fine-tuning, which also guarantees that the model learns to produce outputs for the desired task or dataset that are more accurate and contextually appropriate.

Use of Fine-Tuning: Text classification, recognition of named entities, and text summarization are only a few of the NLP tasks that often make use of fine-tuning. It works particularly well in situations when the knowledge of the pre-trained model may be used but has to be modified to certain domains or situations.

How it helps in improving a good or service: By giving the model the ability to learn from specific to the task data, fine-tuning enhances the model's output. Through this process, the model's parameters are modified in order to better represent the nuanced characteristics and details of the target dataset, producing outputs that are both more precise and contextually relevant. Therefore, fine-tuning aids in improving the model's efficiency on the intended job, which eventually results in better summaries and overall outcomes.

IV. Results

A. Rouge Results

The performance of the summarization models differs across criteria, which can be shown by analyzing the ROUGE scores shown below in the Table: II. The ROUGE scores [10] were calculated through a comparison of the model-generated summaries to the dataset's reference summaries, using the ROUGE-1, ROUGE-2, and ROUGE-L standards. To ensure an accurate evaluation of the models' summarization ability, the average ROUGE scores for each model were calculated across the first 10 rows of the unknown dataset (SAMSum) [20]. This method offers a thorough evaluation of the model's effectiveness on several summarizing quality factors. MBZUAI/LaMini-Flan-T5-248M [5] is a standout model among those analyzed, having the best ROUGE ratings in nearly every parameter. This suggests that it is better than the reference summary at capturing unigram, bigram, and longest common subsequence overlaps. LaMini performs very well because of its strong architecture and efficient training process, which allow it to provide summaries that closely match the content of the original texts.

The MBZUAI/LaMini-Flan-T5-248M model [5] performed better than the others because of its greater size, fine-tuning process, and robust design. Because of its T5-based design, that allows for efficient encoding and decoding of textual data, it is well-suited for summarization work. Furthermore, the model may be strongly adjusted to the complexities of the task by performing fine-tuning using the LaMini-instruction dataset. With 248 million parameters, the model is bigger and has a higher representational ability. On the other hand, smaller model sizes, insufficient fine-tuning, or architectural limitations could have contributed to the lower scores of other models. When everything is considered, the MBZUAI/LaMini-Flan-T5-248M model [5] performs better than other models due to its strong architecture, efficient fine-tuning, and large model size, which allows it to produce summaries of excellent quality.

B. Fine tuning Results

We decided on to fine-tune the 'MBZUAI/LaMini-Flan-T5-248M' model [5] based on its impressive performance in text summarizing tasks. A pre-trained model is easily fine-tuned to perform better on a particular task or dataset. We used Google Colab [21], which is a cloud-based platform that provides open access to GPU resources for model training. The Samsum dataset, which includes messenger-like interactions with summaries to go along with them, was used for improving the model. SAMSum dataset [20] was selected because it closely matches the conversational style and material of the model's initial training data, enabling the model to do tasks that seem more similar to one another. The overall goal of optimizing the 'MBZUAI/LaMini-Flan-T5-248M' model's [5] performance to generate excellent summaries in a conversational setting was accomplished via fine-tuning it on the Samsum dataset.

After choosing and fine-tuning the MBZUAI/LaMini-Flan-T5-248M model on the Samsum dataset, we observed significant gains in its ROUGE scores, as shown in Table III below. The model obtained ROUGE-1, ROUGE-2, and ROUGE-L scores of 37.83, 13.96 and 34.14, respectively, before fine-tuning. These scores improved to 41.29, 13.25, and 38.11, respectively, following adjustments. Hence, gains in recall, precision, and F1 scores across all three ROUGE versions demonstrate enhanced overlaps between the produced summaries and reference summaries. Overall, the MBZUAI/LaMini-Flan-T5-248M model's summarization abilities have improved significantly as a result of the fine-tuning procedure, increasing its efficacy in extracting important information from conversational text data. Furthermore, fine-tuning the MBZUAI/LaMini-Flan-T5-248M model [5] on the SAMSum dataset [20] highlights the potential of leveraging domain-specific data to optimize summarization performance.

V. Conclusion

In conclusion, our research examined how well different text summarizing algorithms performed side by side, providing insight into how well they captured the main ideas of the original texts. It was pointed out that models such as MBZUAI/LaMini-Flan-T5-248M displayed better summarization performance, with higher ROUGE scores on several criteria. Moreover, the significant enhancements in ROUGE scores following to the model's adaptation to the Samsum dataset demonstrated the clear impact of fine-tuning on model performance. This highlights how crucial customized training strategies are for maximizing model performance for certain tasks.

While automatic assessment measures like as ROUGE provide useful insights about summarization quality, they must be accompanied with human evaluation to identify details such as consistency and readability. With consideration of these factors, this dual evaluation technique guarantees a thorough assessment of summarization models.

Our findings demonstrate how important text summarizing algorithms are for compressing large volumes of information into understandable summaries. These models provide efficiency and simplicity in processing textual data, with various number of uses ranging from content development to information retrieval. In the long run, text summarization research has a great deal of promise to enhance natural language processing and make information sharing easier in the digital era.

TABLE II
RESULTS OF ROUGE TEST

LLM Model	ROUGE-1			ROUGE-2			ROUGE-L		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
facebook/bart-large-cnn	35.43	20.19	25.04	9.53	5.40	6.70	34.26	19.73	24.39
Falconsai/text _summarization	23.54	16.66	18.90	7.17	5.31	5.81	20.95	14.77	16.79
MBZUAI/LaMini- Flan-T5-248M	51.61	30.71	37.83	19.94	11.19	13.96	45.97	27.95	34.14
sshleifer/distilbart -cnn-12-6	29.73	15.84	18.04	6.42	3.55	4.34	26.59	14.01	18.04
MBZUAI/LaMini-T5-61M	43.50	32.62	35.21	9.58	8.57	8.37	37.41	28.85	30.80
tuner007/pegasus _summarizer	31.99	18.82	23.39	6.94	3.95	5.02	30.96	18.13	22.56

TABLE III
FINE TUNING RESULTS

LLM Model	ROUGE-1			ROUGE-2			ROUGE-L		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
MBZUAI/LaMini- Flan-T5-248M	51.61	30.71	37.83	19.94	11.19	13.96	45.97	27.95	34.14
MBZUAI/LaMini-Flan-T5-248M (**Fine tuned**)	49.18	38.02	41.29	17.19	12.14	13.25	45.91	34.72	38.11

REFERENCES

[1] "Hugging Face – The AI community building the future." Available: https://huggingface.co/

[2] "facebook/bart-large-cnn · Hugging Face," Jan. 18, 2001. Available: https://huggingface.co/facebook/bart-large-cnn

[3] M. Lewis et al., 'BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension', arXiv [cs.CL]. 2019. Available: https://arxiv.org/pdf/1910.13461.pdf

[4] "Falconsai/text_summarization · Hugging Face." Available: https://huggingface.co/Falconsai/text_summarization

[5] "MBZUAI/LAMINI-FLAN-T5-248M · Hugging face." Available: https://huggingface.co/MBZUAI/LaMini-Flan-T5-248M

[6] M. Wu, A. Waheed, C. Zhang, M. Abdul-Mageed, and A. F. Aji, 'LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions', arXiv [cs.CL]. 2024. Available: https://arxiv.org/abs/2304.14402

[7] "sshleifer/distilbart-cnn-12-6 · Hugging Face," Jan. 18, 2001. Available: https://huggingface.co/sshleifer/distilbart-cnn-12-6

[8] "MBZUAI/LAMINI-T5-61M · Hugging face." Available: https://huggingface.co/MBZUAI/LaMini-T5-61M

[9] "tuner007/pegasus_summarizer · Hugging Face." Available: https://huggingface.co/tuner007/pegasus_summarizer

[10] "Metric Card for ROUGE, Hugging Face Metrics." [Online]. Available: https://huggingface.co/spaces/evaluate-metric/rouge

[11] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics. Available: https://aclanthology.org/W04-1013.pdf

[12] F. Chiusano, "Two minutes NLP — Learn the ROUGE metric by examples," Medium, Aug. 04, 2023. [Online]. Available: https://medium.com/nlplanet/two-minutes-nlp-learn-the-rouge-metric-by-examples-f179cc285499

[13] M. Mayo, "Approaches to text Summarization: An Overview - KDNuggets," KDnuggets. Avaiable: https://www.kdnuggets.com/2019/01/approaches-text-summarization-overview.html

[14] Cloudflare. "What Is a Large Language Model?" Cloudflare, Available: https://www.cloudflare.com/learning/ai/what-is-large-language-model/.

[15] "The Ultimate Guide to LLM Fine Tuning: Best Practices & Tools — Lakera – Protecting AI teams that disrupt the world." Available: https://www.lakera.ai/blog/llm-fine-tuning-guide

[16] R. Rastogi, "Papers explained 09: BART - DAIR.AI - medium," Medium, Sep. 20, 2023. [Online]. Available: https://medium.com/dair-ai/papers-explained-09-bart-7f56138175bd

[17] Q. Chen, "T5: a detailed explanation - Analytics Vidhya - Medium," Medium, Dec. 14, 2021. [Online]. Available: https://medium.com/analytics-vidhya/t5-a-detailed-explanation-a0ac9bc53e51

[18] Exemplary AI, "Accurate transcription and AI-Assisted editing & analysis," Exemplary AI, Feb. 02, 2023. Available: https://exemplary.ai/blog/flan-t5

[19] Shan, Ming-Chien. "Pegasus architecture and design principles." Proceedings of the 1993 ACM SIGMOD international conference on Management of data (1993): n. pag. Available: https://dl.acm.org/doi/pdf/10.1145/170036.170111

[20] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, 'SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization', in Proceedings of the 2nd Workshop on New Frontiers in Summarization, 2019. Available: https://arxiv.org/pdf/1911.12237.pdf

[21] "Google Colaboratory." Available: https://colab.research.google.com/