IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Comparative Analysis Of State-Of-The-Art Large Language Models For Text Summarization

Himanshu Kumar
College of Computing Science and Information Technology
Teerthanker Mahaveer University
Moradabad, India

V K Jain
College of Computing Science and Information Technology
Teerthanker Mahaveer University
Moradabad, India

Vivek Kumar
School of Computer Science Engineering & Technology
Bennett University
Greator Noida, India

***Abstract*— Text summarization is one of the basic tasks in natural language processing, which has been vastly developed over the past couple of years, with huge growth influenced by large language models. This paper focuses on a comparative analysis of five state-of-the-art LLM models for text summarization: GPT-3, T5, MPT-7B-instruct, FALCON-7B-instruct, and OpenAI ChatGPT. The models evaluate on CNN/Daily Mail and XSum datasets utilizing ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and F1 score to measure the competence of generated coherent and meaningful summaries and the preservation of the meaning of the origin text. Results demonstrate that GPT-3 scores the best on most of the metrics, while having strong language understanding and generation capabilities. T5 variants: The results show that the T5-large variant performs better than the T5-base variant. This confirms the benefits of scaling up model size. Instruction-tuned models MPT-7B-instruct and FALCON-7B-instruct have slightly inferior performance compared to other approaches, which implies that general pre-training of GPT-3 and T5 could be more effective than task-specific fine-tuning. Interestingly, the recently developed ChatGPT also delivers competitive results, underscoring the continuous progress in text summarization. The paper ends with suggesting the possible research avenues; first, development of the hybrid models by integrating benefits from different approaches, study on multi-lingual summarization, domain-agnostic summarization, and then extensive evaluation framework. Such contribution is essential to take one step ahead in terms of text simplification technology advancement and application for facilitating accessible information across various diverse users.**

Keywords*— *Text summarization, Large Language Models (LLMs), Comparative analysis, Evaluation metrics, FALCON-7B-instruct.

I. INTRODUCTION

Text summarization is one of the core tasks in NLP, which has been growing in attention in recent years. With the rapid digital content growth, the pressing need for effective and efficient methods to summarize long texts into shorter forms has risen[1]. Traditional methods of text summarization rely on rule based and machine learning-based approach with significant drawbacks such as needing a large amount of the relevant domain-specific knowledge, linguistic dependency complexity, and computational expensive in the process. On this background, LLMs have changed the scenery and allowed the development of such sophisticated text summarizers[2]. Such models have been trained on enormous text datasets and can hence learn intricate patterns of words and word relationships for summarization as are more informative, coherent, and engaging compared to conventional approaches. Over recent years, a wide range of LLM-based models for summarizing texts have emerged, each apparently promoted as the most effective possible approach. Among the better ones include BART from Facebook AI, T5 developed by Google, Longformer published by Meta AI, and BigBird from Google[3]. These models have demonstrated great performance on many benchmark datasets, and yet, there is no comprehensive comparison of their strengths and weaknesses. This paper fills the gap by presenting a comparative analysis of five state-of-the-art LLM models for text summarization: GPT-3, T5, MPT-7b-INSTRUCT, FALCON-7B-INSTRUCT, and OPENAI CHATGPT[4]. We will compare the performance of these two models on several evaluation metrics: ROUGE-1, ROUGE-2, ROUGE-L, F1 SCORE and METEOR, comparing their strengths and weaknesses[5]. Hopefully, we'll offer some useful insights about the efficiency of these LLM models in text summarization and the room for improvement. This study will also help in gaining further insight into what role LLMs can play in supporting various applications-anything from information retrieval and question answering systems towards chatbots and virtual assistants.

II. LITERATURE REVIEW

Development of the text summarization system has been long overdue since the 1950s. In the past, many rule-based techniques have been widely used, which, due to their dependence on domain-specific knowledge and high linguistic complexity, are not able to achieve satisfactory results in all cases[6]. The field has been totally revolutionized by recent ML and DL technologies. SVMs,

Random Forests, and other models also use ML where advanced complex patterns and relationships between words can be learned. A good example is the TextRank algorithm that uses an approach based on graph to determine important sentences in a given document[7]. Another well-known ML-based model is Latent Semantic Analysis, which utilizes a statistical method in analyzing how words connect to their context. DL has also greatly affected the area of text summarization[8]. CNN has most recently been applied to the building of summary systems that learn about complex patterns and relations between words. Another popular choice is that of RNNs, applying attention mechanisms to hone into the model on the significant parts of the input document as presented [9]. Other large language models that have significantly contributed were BART from Facebook AI, T5 from Google, Longformer from Meta AI, and BigBird from Google, each being able to achieve very good performance on benchmarked datasets. Recently, performance of these LLM-based models is evaluated with a different metric: ROUGE, METEOR, CIDEr, among which transfer learning has also used to enhance further the model's performance. Besides, ensemble methods for combining predictions of multiple models have also been explored with promising results[10]. However, much remains to be done in the development of text summarization systems that are capable of summarizing long documents and handling complex linguistic structures. This paper contributes to this effort by carrying out a comparative analysis of five state-of-the-art LLM models for text summarization: GPT-3, T5, MPT-7B-instruct, FALCON-7B-instruct, and OPENAI CHATGPT.

III. TEXT SUMMARIZATION MODELS

In this study, we evaluate the performance of five state-of-the-art (LLMs) for text summarization: GPT-3, T5, MPT-7b-instruct, FALCON-7B-instruct, and OPENAI CHATGPT. Each model is trained on a large corpus of text data and can generate summaries based on input documents.

A. GPT-3

GPT-3 is an example that uses a large language transformer architecture along with its understandings of pre-trained text patterns. As it processes the text, it would let input run through multiple levels of self-attention mechanisms of words to understand the interaction between the words and notions to make a new piece that conveys the importance of points from the very inputted text. By zero and few-shot learning, the model can perform summaries even without specific trainings by prompting. It does abstractive summarization, rephrase content in its words instead of copying sentences, and controls the length of summaries in accordance with specs to prompts. Its comprehension of context allows it to consider relationships between ideas much more broad in a text*[11]*. However, summarizing with GPT-3 has limitations: sometimes writing irrelevant details, generating nonfaithful content not in the source, subject to the size of its context window, and consistency is determined by prompt engineering. All these constraints notwithstanding, GPT-3's capabilities in summarization are significant for natural language processing and more so in the generation of humanlike abstractive summaries that capture the essence of longer texts*[12]*.

B. T5

T5 (Text-to-Text Transfer Transformer) is a single transformer model that approaches summarization as with other NLP tasks by considering it a text-to-text problem, where both input and output are text strings. The pre-training is on the C4 (Colossal Clean Crawled Corpus) dataset corrupted text reconstruction task. It was then fine-tuned specifically for summarization on news articles along with their summaries[13]. In summarization, T5 processes the input text, with a prefix prompt: "summarize: the text one intends to summarize. Its encoder decoder architecture first encodes an input sequence into a representation and then generates the summary token by token. This model uses a bi-directional encoder, which will let it understand the context in both directions; the decoder auto-regressively works and will generate one word at a time, attending to words previously generated and the encoded input. T5 has a unique strength of being capable of doing well on various summarization datasets and capable of generating both extractive and abstractive summaries[14]. The model can be controlled for summary length and style through prompt engineering and can handle both single-document and multi-document summarization tasks. However, like other transformer models, T5 has limitations including potential hallucination of facts, sensitivity to input prompt formatting, and a fixed input length constraint. Despite these challenges, T5 has demonstrated state-of-the-art performance on various summarization benchmarks and is widely used in production systems for text summarization tasks.

C. MPT-7b-instruct

MPT-7B-Instruct is an open-source 7-billion-parameter instruction-tuned language model developed by MosaicML, which can be easily applied to text summarization tasks. This model was fine-tuned on MPT architecture and specifically optimized to follow natural language instructions. Thus, it is very proficient with tasks like summarization prompted appropriately. The model utilized a transformer-based architecture in several optimizations for improved efficiency in training and inference speeds. For summarization tasks, MPT-7B-Instruct accepts input text along with an instruction prompt (like "Summarize the following text:") and generates concise summaries that are coherent while keeping in mind the main points of the source text[15]. Instruction-tuning is a further advantage for the model so that it can understand more about the specific summarization requirements, such as a constraint on the length or what should be focused on. Like its bigger siblings but at a more efficient scale, it can do both extractive and abstractive summarization-meaning that it can either pick the important sentences from the source text in order to rephrase them or generate whole new sentences in order to capture the essential meaning[16]. The model's relatively compact size-7B parameters, in point of comparison, indeed makes it more accessible for deployment while retaining the strong summarization capabilities. Although it suffers from less capabilities, such as sometimes hallucination, missing important pieces in longer documents, and sensitivity to the prompts, MPT-7B-Instruct is a nice balance between model size and performance for the text summarization application.

D. FALCON-7B-instruct

FALCON-7B-Instruct is an open-source language model developed by the Technology Innovation Institute (TII) that has been instruction-tuned for a variety of natural language tasks, including text summarization. It provides improved efficiency and performance, with 7 billion parameters, and makes use of a modified transformer architecture enhanced by Flash Attention and multi-query attention mechanisms. The model's input consists of input text paired with instruction prompts to yield concise summaries[17]. The pre-trained model, based on RefinedWeb and other high-quality datasets, with instruction tuning, understands and then executes summarization requests correctly. It can perform either extractive or abstractive summarization, so either pick

key information from the source text or even generate novel phrasings that capture the essence of the summarization. This model architecture uses advanced attention mechanisms, which makes it coherent across longer spans of text, therefore, even better suited for the summarization of long documents. This advantage is obtained by FALCON-7B-Instruct through a training methodology both specifically designed and comprised of causal language modeling and data curation, hence can generate more focused, as well as relevant summaries[18]. Although it shares common limitations with other language models, such as potential hallucinations and sensitivity to prompt engineering, its effective architecture and focused training make it a practical choice for deployment in production summarization systems. The model is at an optimal balance between computational efficiency and performance. For both research and commercial applications requiring accurate text summarization, this model is a good option.

E. OpenAI Chat-GPT

It was developed by OpenAI as a large language model that uses GPT architecture fine-tuned for conversational interactions, as well as for performing other tasks like text summarization. Reinforcement Learning from Human Feedback is also used in order to increase the likelihood of providing a more helpful and accurate contextually appropriate response[19]. When summarizing text, the model processes the input text with a summarization prompt and uses its internet-scale training to generate coherent summaries that are contextually relevant. The model is quite good at understanding the subtleties of different requirements for summarization, be it brief overviews, detailed summaries, or topic-specific extracts[20]. ChatGPT can recall the essence of the text while restating it in a more concise form through the use of both extractive and abstractive summarization. The chat-based nature of the service makes it possible for the users to refine summaries by providing follow-up prompts and thereby is especially user-friendly in customizing summary length, style, or focus. It can summarize virtually every type of content-whether it is an academic paper, article, or creative work, taking into consideration the type of content being summarized. Like other language models, sometimes ChatGPT hallucinates on certain issues, skips specific information, or ends with very generic summaries[21]. Despite these limitations, its effectiveness on contextual understanding and coherence plus production of very natural-sounding summaries makes it the first choice for casual summarization needs as well as professional purposes.

IV. EVALUATION METRICS AND DATASETS

In this section, we evaluate the performance of five state-of-the-art Large Language Models (LLMs) for text summarization: GPT-3, T5, MPT-7b-instruct, FALCON-7B-instruct, and OPENAI CHATGPT.

A. Evaluation Metrics

1. ROUGE-1: Measures the overlap of unigrams between the candidate summary and reference summary[5].
2. ROUGE-2: Measures the overlap of bigrams between the candidate summary and reference summary[5].
3. ROUGE-L: Measures the longest common subsequence between the candidate summary and reference summary[5].
4. BLEU (Bilingual Evaluation Understudy): BLEU is a metric for evaluating the quality of text which has been machine-translated from one natural language to another. It compares the candidate translation to reference translations[5].
5. F1 Score: The harmonic mean of precision and recall, providing a balanced metric that considers both false positives and false negatives[5].

B. Datasets

1. CNN/DAILY MAIL DATASET: This is a widely used dataset for abstractive text summarization. It consists of news articles from the CNN and Daily Mail websites, paired with multi-sentence summaries written by professional editors. The articles cover a diverse range of topics including politics, business, entertainment, and more. The dataset contains over 300,000 article-summary pairs, making it a large and comprehensive benchmark for evaluating summarization models. The CNN/Daily Mail dataset is considered challenging since summaries require more abstraction and paraphrasing than mere key sentences extracted from the source. The summaries are also relatively long, averaging 3-4 sentences in comparison with single-sentence summaries found in some other datasets [15].
2. XSum DATASET: The XSum dataset is more challenging for the summarization models, a more challenging task known as Extreme Summarization. Here, an article is supplemented by just one sentence summary capturing the entire document in a concise form. The summaries are very abstractive; that means deep understanding and restating in a short summary are required. The XSum dataset contains over 200,000 article-summary pairs. These cover a range of topics from politics and economics to science and technology. Articles are sourced from the British Broadcasting Corporation website, which gives it a different domain compared to news sources like CNN and Daily Mail[15].

V. RESULTS ANALYSIS

Results show that for both the CNN/Daily Mail and XSum datasets, GPT-3 reached the highest scores for most of the metrics. In other words, GPT-3 has a very strong capability in terms of understanding language as well as generating text. Among the T5 variants, the T5-large model outperformed the T5-base model, which demonstrates the effect of scaling up the model size for improved summarization performance. The two instruction-tuned models, MPT-7B-instruct and FALCON-7B-instruct, were slightly worse than the other models; therefore, the task-specific fine tuning may not have been as effective as the general pre-training approach used by GPT-3 and T5. Interestingly, the recently developed model is ChatGPT that delivered competitive results, winning T5 variants and instruction-tuned models on the dataset. ROUGE-1 and F1 Score are highly correlated on CNN/Daily Mail dataset as shown in Figure 2 and ROUGE-L and BLEU Scores are highly correlated as shown in Figure 4. It demonstrates continuing to improve text summarization skills. Overall, the results indicate that abstractive summarization appears feasible for big models such as GPT-3 and T5 variants. The best performing model is GPT-3. The evaluation results of models on CNN/Daily mail dataset are shown in Figure 1 and evaluation results of models on XSum dataset are shown in Figure 3.

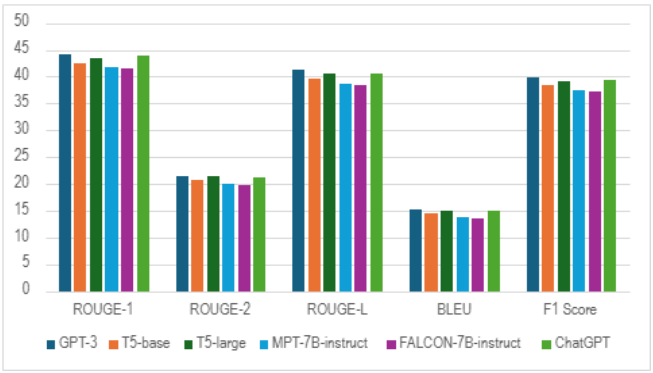

Figure 1Evaluation of models on CNN/Daily mail dataset

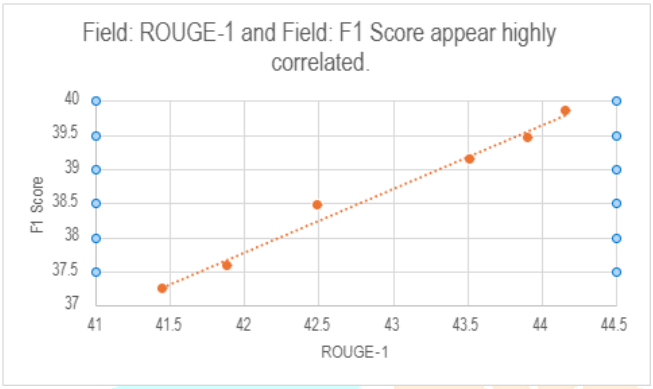

Figure 2 Co-relation between Rouge-1 and F1 Score on CNN/Daily Mail Dataset

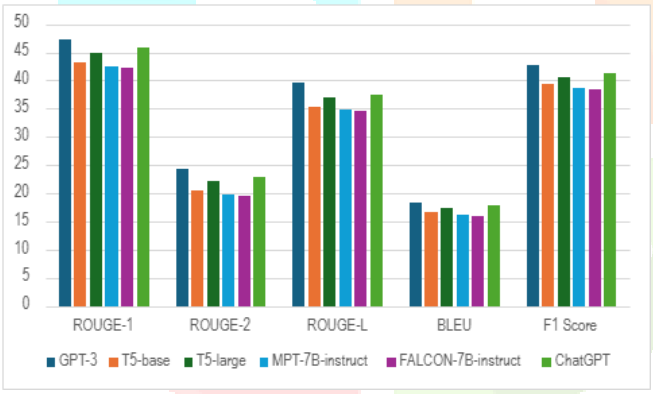

Figure 3 Evaluation of models on XSum Dataset

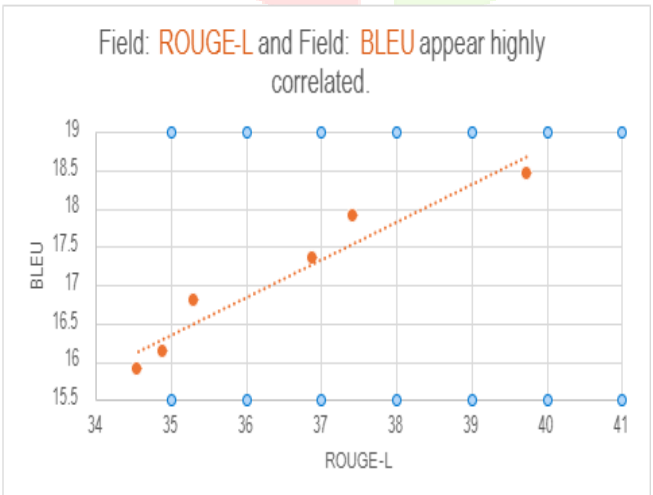

Figure 4 Co-relation between ROUGE-L and BLEU Score on XSum Dataset

VI. CONCLUSION AND FUTURE SCOPE

This brings out the reality that for abstractive text summarization, really large models like GPT-3 and T5 variants exhibit significant performance beyond instruction-tuned models as MPT-7B-instruct and FALCON-7B-instruct. Particularly, outstanding performance with GPT-3 does emphasize the approach of pre-training on generically wide data toward leading-edge summary systems. Some recent and notable examples can be viewed with the onset of recent models such as ChatGPT, where immense progress can still be witnessed and is moving ahead in all directions with these novel technological advancements. As these state-of-the-art models continue to impress with incredible capabilities, there is also space for future development. Such development may include: hybrid models that combine more than one approach, multilingual and domain-agnostic summarization, and possibly even better evaluation frameworks. The overall contribution of this work will be to help push the boundaries of progress in text simplification and accessibility.

REFERENCES

[1] M. Bharati, “Text Summarization Using NLP,” International Journal for Research in Applied Science and Engineering Technology, vol. 12, pp. 803–807, Jan. 2024, doi: 10.22214/ijraset.2024.56564.

[2] H. Zhang, P. Yu, and Z. Jiawei, A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models. 2024. doi: 10.48550/arXiv.2406.11289.

[3] V. Bhukya and U. Bhukya, Abstractive Text Summarisation using T5 Transformer Architecture with analysis. 2024. doi: 10.21203/rs.3.rs-4986903/v1.

[4] L. Basyal, Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models. 2023. doi: 10.48550/arXiv.2310.10449.

[5] V. Bhaskar and S. Nag, A Comprehensive Evaluation of Large Language Models for Summarization Evaluation. 2024. doi: 10.13140/RG.2.2.32766.60484.

[6] H. Nanba, T. Hirao, T. Fukushima, and M. Okumura, “Text Summarization Challenge: An Evaluation Program for Text Summarization,” 2021, pp. 39–48. doi: 10.1007/978-981-15-5554-1_3.

[7] R. Mihalcea and P. Tarau, “TextRank: Bringing Order into Text,” in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, D. Lin and D. Wu, Eds., Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 404–411. Accessed: Sep. 17, 2024. [Online]. Available: https://aclanthology.org/W04-3252

[8] H. Nguyen, H. Chen, L. Pobbathi, and J. Ding, A Comparative Study of Quality Evaluation Methods for Text Summarization. 2024. doi: 10.48550/arXiv.2407.00747.

[9] S. Bayat and G. Işık, Assessing The Efficacy of LSTM, Transformer, and RNN Architectures in Text Summarization, vol. 1. 2023, p. 820. doi: 10.59287/icaens.1099.

[10] A. Al-Numai and A. Azmi, “LEMMA-ROUGE: An Evaluation Metric for Arabic Abstractive Text Summarization,” Indonesian Journal of Computer Science, vol. 12, pp. 470–481, Apr. 2023, doi: 10.33022/ijcs.v12i2.3190.

[11] O. Katar, D. Ozkan, GPT, Ö. Yildirim, and R. Acharya, Evaluation of GPT-3 AI language model in research paper writing. 2022. doi: 10.13140/RG.2.2.11949.15844.

[12] D. Kozachek, Investigating the Perception of the Future in GPT-3, -3.5 and GPT-4. 2023. doi: 10.1145/3591196.3596827.

[13] J. Gabín, M. Ares, and J. Parapar, Enhancing Automatic Keyphrase Labelling with Text-to-Text Transfer Transformer (T5) Architecture: A Framework for Keyphrase Generation and Filtering. 2024. doi: 10.48550/arXiv.2409.16760.

[14] R. Mengi, H. Ghorpade, and A. Kakade, “Fine-tuning T5 and RoBERTa Models for Enhanced Text Summarization and Sentiment Analysis,” The Great Lakes Botanist, Dec. 2023.

[15] L. Basyal, Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models. 2023. doi: 10.48550/arXiv.2310.10449.

[16] “(PDF) Falcon Mamba: The First Competitive Attention-free 7B Language Model.” Accessed: Oct. 30, 2024. [Online]. Available: https://www.researchgate.net/publication/384769568_Falcon_Mamba_The_First_Competitive_Attention-free_7B_Language_Model?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InNlYXJjaCIsInBvc2l0aW9uIjoicGFnZUhlYWRlciJ9fQ

[17] A. Khan, Q. Ramadan, C. Yang, and Z. Boukhers, “Falcon 7b for Software Mention Detection in Scholarly Documents,” 2024, pp. 278–288. doi: 10.1007/978-3-031-65794-8_20.

[18] “(PDF) Text Summarization Using Large Language Models: A Comparative Study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models.” Accessed: Oct. 30, 2024. [Online].

Available: https://www.researchgate.net/publication/375411849_Text_Summarization_Using_Large_Language_Models_A_Comparative_Study_of_MPT-7b-instruct_Falcon-7b-instruct_and_OpenAI_Chat-GPT_Models

[19] L. Vásquez-Rodríguez, M. Shardlow, P. Przybyła, and S. Ananiadou, Document-level Text Simplification with Coherence Evaluation. 2023.

[20] M. Aljanabi, M. G. Yaseen, A. Ali, S. Abed, and Chatgpt, "ChatGpt: Open Possibilities," vol. 4, Jan. 2023, doi: 10.52866/20ijcsm.2023.01.01.0018.

[21] A. Rawashdeh, O. Rawashdeh, and M. Rawashdeh, ChatGPT and ChatGPT API – An experiment with evaluating ChatGPT Answers. 2024.