Indonesian Journal of Electrical Engineering and Computer Science
Vol. 39, No. 1, July 2025, pp. 446~454
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v39.i1.pp446-454

Investigating the recall efficiency in abstractive summarization: an experimental based comparative study

Surabhi Anuradha[1,2], Martha Sheshikala[1]
[1]School of Computer Science and Artificial Intelligence, SR University, Warangal, India
[2]Department of Computer Science and Engineering (AIML), Keshav Memorial Institute of Technology, Hyderabad, India

Article Info

Article history:

Received Aug 5, 2024
Revised Mar 17, 2025
Accepted Mar 26, 2025

Keywords:

Abstractive summary
Claude
Coherence
Falcon-7B-instruct
Large language models
Mistral-7B
MythoMax-13B

ABSTRACT

This study explores text summarization, a critical component of natural language processing (NLP), specifically targeting scientific documents. Traditional extractive summarization, which relies on the original wording, often results in disjointed sequences of sentences and fails to convey key ideas concisely. To address these issues and ensure comprehensive inclusion of relevant details, our research aims to improve the coherence and completeness of summaries. We employed 25 different large language models (LLMs) to evaluate their performance in generating abstractive summaries of scholarly scientific documents. A recall-oriented evaluation of the generated summaries revealed that LLMs such as 'Claude v2.1,' 'PPLX 70B Online,' and 'Mistral 7B Instruct' demonstrated exceptional performance with ROUGE-1 scores of 0.92, 0.88, and 0.85, respectively, supported by high precision and recall values from bidirectional encoder representations from transformers (BERT) scores (0.902, 0.894, and 0.888). These findings offer valuable insights for NLP researchers, laying the foundation for future advancements in LLMs for summarization. The study highlights potential improvements in text summarization techniques, benefiting various NLP applications.

Corresponding Author:

Surabhi Anuradha
School of Computer Science and Artificial Intelligence, SR University
Warangal, India
Email: anuradha@kmit.in

1. INTRODUCTION

Text summarization is essential in today's data-driven world, especially in fields like journalism, law, and academia. In journalism, efficient summarization enables quick digestion of vast amounts of information, facilitating timely news updates. Legal professionals rely on summarization to extract key points from lengthy documents, aiding in case preparation and research. Similarly, in academia, summarization helps researchers sift through extensive literature, identifying relevant studies efficiently.

Text summarization focuses on condensing a given text while retaining its essential information and main ideas. The goal is to produce a concise and coherent summary that accurately reflects the core meaning of the original content. Broadly, text summarization can be classified into two main approaches: abstractive and extractive summarization. Extractive summarization involves selecting key sentences or phrases directly from the original text to form a summary, as noted by Ghadimi and Beigy [1]. This method extracts content verbatim, emphasizing importance or relevance criteria. Algorithms assess sentences based on features like length, word frequency, and key keywords, facilitating the creation of a succinct representation of the original material. Conversely, abstractive summarization is a more advanced approach that does not merely extract sentences from the original text but instead generates entirely new ones. This method demands a deep

understanding of the content to create a summary that is both coherent and natural-sounding. To accomplish this, sophisticated models, particularly large language models (LLMs), are utilized. As stated by Basyal and Sanghvi [2] these LLMs have the capacity to rewrite and rephrase content, capturing the essence of the original text in a more concise and nuanced manner. The use of LLMs contributes to the development of summaries that not only convey the core information but also exhibit a higher level of linguistic fluency and context awareness.

In recent times, text summarization has undergone transformative changes, primarily propelled by the rise of LLMs. Prominent models like OpenAI's ChatGPT, MPT-7b-instruct, flan-t5-xl, falcon-7b-instruct, Mistral-7b-instruct, Mythomist, Llama, Clauda, PaLM, Hermes, Toppy M 7B, PPLX Online LLMs have emerged as pioneers, displaying impressive abilities in understanding and generating text with human-like fluency. This paradigm shift has raised the bar and paved the way for unprecedented progress in text summarization. The article explores the profound impact of these LLMs, highlighting their potent generative capabilities and adaptability across diverse tasks through fine-tuning. The examination of these models unveils new possibilities for enhancing text summarization methodologies, marking a pivotal moment in the evolution of NLP technologies.

Having its roots in English language pre-training, the bidirectional and auto-regressive transformers (BART) model has undergone a transformative fine-tuning process on the CNN Daily Mail dataset. Operating as a transformer-based encoder-decoder model with a bidirectional encoder akin to bidirectional encoder representations from transformers (BERT) and an autoregressive decoder resembling GPT, BART emerges as a versatile powerhouse. Its excellence extends to text generation tasks like summarization and translation, while also proving adept in comprehension tasks such as text classification and question answering. This specific iteration, ‘facebook/bart-large-cnn’, fine-tuned on the extensive CNN Daily Mail dataset, further amplifies BART's proficiency, leveraging a vast collection of text-summary pairs to refine its language understanding and generation capabilities [3].

The 'text-davinci-003 (Legacy)' model represents a significant advancement in natural language processing (NLP), demonstrating exceptional accuracy and proficiency across a wide range of language tasks. It surpasses its predecessors, such as the Curie, Babbage, and Ada models, by consistently producing higher-quality and more extended text outputs while adhering closely to given instructions. With a token capacity of 4,097, this legacy model efficiently handles extensive text generation [4]. Notably, ‘GPT-4’ has been recognized for its improvements over ‘GPT-3.5’ OpenAI emphasizes that ‘GPT-4’ is more reliable, creative, and capable of handling more nuanced instructions. Notably, ‘GPT-4’ introduces two versions with significantly expanded context windows, allowing for the processing of 8,192 and 32,768 tokens. This marks a substantial improvement compared to the limitations of ‘GPT-3.5’ and ‘GPT-3’, which were confined to 4,096 and 2,049 tokens, respectively [5].

As stated in reference [6], the ‘MPT-7B-Instruct’ model is designed specifically for short-form instruction-following tasks, making it an excellent choice for various instruction-based applications. This 7-billion-parameter LLM was trained on 1 trillion tokens over 9.5 days using 440 A100-40G graphics processing units (GPUs). It is developed by fine-tuning the base model, ‘MPT-7B,’ with the anthropic helpful and harmless (HH-RLHF) dataset and Databricks Dolly-15k dataset. This tailored approach results in a model that excels at accurately comprehending and following instructions with precision. Featuring a decoder-only architecture, the model is optimized for scenarios where users ask questions and expect concise, direct responses rather than an extended continuation of their input.

‘Falcon-7B-Instruct’ is a highly capable 7-billion-parameter causal decoder-only model, meticulously developed by the Technology Innovation Institute (TII) as an extension of ‘Falcon-7B’. Fine-tuned on a diverse dataset from chat and instruction-based domains, it is released under the Apache 2.0 license. As a significant advancement in language models, ‘Falcon-7B-Instruct’ serves as a powerful and openly licensed contribution to the field [7].

The ‘Mistral-7B-v0.1’ LLM is a 7-billion-parameter generative text model that outperforms ‘Llama-2-13B’ across all assessed benchmarks. Designed as a transformer-based architecture, it integrates advanced features such as Grouped-Query Attention, Sliding-Window Attention, and a Byte-fallback BPE tokenizer. Notably, ‘Mistral 7B’ functions as a base model and does not include moderation mechanisms, as it is purely a pre-trained model [8].

The 'Llama-v2-13B' model is a carefully fine-tuned language model tailored for dialogue-based applications and commercial use. Built on an optimized transformer architecture, ‘Llama 2’ functions as an auto-regressive model. It employs a dual optimization strategy combining supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences, ensuring both usefulness and safety [9]. Meanwhile, ‘CodeLlama 34B v2,’ refined from ‘Phind-CodeLlama-34B-v1,’ achieves an impressive 73.8% pass@1 on HumanEval, establishing itself as the leading open-source model in its domain [10].

‘MythoMax 13B’, as cited in [11], represents a pinnacle in fine-tuned language models, originating from the robust ‘Llama 2 13B’. An evolution of Gryphe's ‘MythoMax L2 13B’ model card, this variant introduces a refined fusion technique, merging MythoLogic-L2 and Huginn through an experimental tensor type merge. What sets ‘MythoMax 13B’ apart is its emphasis on enriched descriptions and roleplay capabilities, making it a go-to choose for narrative tasks. Notably, the model employs Alpaca formatting, ensuring a visually consistent and engaging output. The innovative approach of allowing more intermingling of Huginn with the model's tensors enhances overall coherence. In essence, ‘MythoMax 13B’ combines fantasy elements with structural finesse, offering a powerful tool for immersive storytelling and roleplaying experiences.

The ‘toppy-m-7b’ model [12], accessible at ‘undi95/toppy-m-7b’, boasting an impressive 7 billion parameters, represents a convergence of influential models facilitated by the innovative ‘task_arithmetic’ merge method from ‘mergekit’. Merged models include ‘NousResearch/Nous-Capybara-7B-V1.9’, ‘HuggingFaceH4/zephyr-7b-beta’, ‘lemonilia /AshhLimaRP-Mistral-7B’, ‘Vulkane/120-Days-of-Sodom-LoRA-Mistral-7b’, and ‘Undi95/Mistral-pippa-sharegpt-7b-qlora’. This collaborative effort yields a powerful model, demonstrating the forefront of research and innovation in the field of NLP.

‘MythoMist 7B’, available at ‘gryphe/mythomist-7b’ [13], is a sophisticated chat-based language model designed to enhance roleplaying experiences. With an expansive context of 32,768 tokens, it offers a seamless conversational flow. Created by the mastermind behind ‘MythoMax’, this model skillfully merges several prominent models, including ‘Neural Chat 7B’, ‘Airoboros 7b’, ‘Toppy M 7B’, ‘Zephyr 7b beta’, ‘Nous Capybara 34B’, ‘OpenHeremes 2.5’, and more. The integration aims to minimize word anticipation, refine ministrations, and mitigate the presence of undesirable words, providing an enriched and tailored roleplaying environment.

As cited by Alpindale [14], ‘Goliath 120B’, is a formidable language model that leverages an extensive context of 6,144 tokens to provide a rich and nuanced chat experience. Created through the amalgamation of two finely-tuned Llama 70B models, this large LLM boasts an impressive parameter count of 120 billion. The model seamlessly integrates the capabilities of Xwin and Euryale, resulting in a powerful and versatile language generation model.

‘PaLM 2’, tailored for chatbot interactions, excels in assisting with inquiries related to coding. As a cutting-edge language model, ‘PaLM 2’ boasts enhanced multilingual proficiency, advanced reasoning abilities, and an adept understanding of coding concepts. Developed by Google, ‘PaLM 2’ is a chat-bison model designed to handle code-related questions effectively. Notably, it supports a substantial context window of 8,000 tokens, facilitating comprehensive and contextually rich conversations [15].

Anthropic introduced Claude models, as cited in [16], [17], often work well for writing, editing, summarizing, searching, and general, open-ended conversations. Constitutional artificial intelligence (AI) and unsupervised learning are used in the training of Claude models, which are general-purpose LLMs that employ transformer architecture. Claude models are corporate application-specific models. The chat completion model, Claude v1, is perfect for condensing, examining, and searching lengthy texts and discussions in order to gain a sophisticated grasp of intricate subjects and their connections throughout extremely long text segments. The flagship model, ‘Claude v2.0’, has longer reflexes and performs better. It has an astounding 100k token capacity for a context window. The advanced LLM ‘Claude v2.1’ has a 200K token context window and a 2x reduction in token usage.

As referenced in [18], ‘Nous-Hermes-Llama2-13B’ is an advanced language model meticulously fine-tuned on a large dataset of over 300,000 instructions. It is distinguished by its ability to generate extended responses, minimize hallucinations, and operate without OpenAI censorship mechanisms in its synthetic training data. Primarily trained on synthetic GPT-4 outputs, the model benefits from high-quality GPT-4 datasets, enhancing its proficiency in knowledge delivery, task execution, and stylistic generation.

The PPLX models, exemplified by ‘PPLX-7B-Online’ and ‘PPLX-70B-Online’, redefine the landscape of LLMs by specifically tackling two prevalent challenges. Unlike many LLMs, these models prioritize delivering responses that are not only helpful and factual but also up-to-date, overcoming the limitation of outdated information. Additionally, they address inaccuracies commonly associated with LLMs, minimizing hallucinations and ensuring responses are accurate and reliable. Through these advancements, PPLX models set a new standard by providing a unique blend of real-time relevance, precision, and helpfulness in their language generation capabilities [19], [20].

This paper offers an in-depth exploration of text summarization, analyzing a diverse set of 25 LLMs, from conventional models to the latest innovations. It thoroughly examines the capabilities and limitations of state-of-the-art LLMs by experimenting with various hyperparameters and assessing the generated summaries using established metrics such as bilingual evaluation understudy (BLEU) score, recall-oriented understudy for gisting evaluation (ROUGE) score, and BERT score. As a valuable reference, this

study provides essential insights for those seeking to harness LLMs in NLP applications and paves the way for developing advanced generative AI solutions to tackle a wide range of business challenges.

Basyal and Sanghvi [2] in their study used three LLMs MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT with two datasets CNNDaily Mail and XSum. Their focus is on extractive summarization of news articles, primarily evaluating performance using precision. Extractive summaries, however, often result in a disjointed sequence of sentences due to the simple concatenation of sentences from various parts of the text, leading to a lack of natural flow and coherence. Additionally, because extractive summarization cannot rephrase or condense information and relies solely on the original wording, it fails to convey key ideas concisely. In contrast, our research focuses on generating abstractive summaries for scientific scholarly documents, incorporating a diverse set of 25 advanced LLMs. Our primary focus is on recall-oriented evaluation to ensure comprehensive inclusion of relevant details. The key objectives of our research work are:

- To improve natural flow and coherence in summaries. Our research aims to address the issue of extractive summaries often resulting in a disjointed sequence of sentences. By moving beyond simply concatenating sentences from different parts of the text, we seek to enhance the natural flow and coherence of the generated summaries. Additionally, our goal is to overcome the limitations of extractive summarization, which cannot rephrase or condense information and relies solely on the original wording, often failing to convey key ideas concisely.
- To ensure comprehensive inclusion of relevant details. Recognizing that scientific documents contain dense, critical information essential for understanding research, our research focuses on achieving high recall. This ensures that the summaries include as many relevant details and key points from the original document as possible, thereby providing comprehensive and accurate representations of the source material.

The remainder of the paper is structured as follows: section 2 introduces the dataset and evaluation metrics used to measure model performance. Section 3 details the experimental setup, while section 4 presents the inference results of various LLMs. Lastly, section 5 concludes the study and outlines directions for future enhancements.

2. DATASET AND EVALUATION METRICS

In this study, we conducted experiments and evaluations using the SciTLDR dataset, originally introduced by Cachola *et al*. [21] in "TLDR: extreme summarization of scientific documents". The SciTLDR dataset is designed to facilitate extreme summarization tasks in the context of scientific documents. Our study aimed to assess the performance of specific models in extreme summarization, leveraging this dataset as a benchmark. We followed a methodology that involved implementing various algorithms and assessing their summarization capabilities, using metrics relevant to extreme summarization tasks. The choice of the SciTLDR dataset provided a standardized and challenging set of scientific documents, allowing for a comprehensive evaluation of the summarization models.

2.1. Dataset

SciTLDR: this dataset, comprising multiple targets, encompasses 5.4K TLDRs extracted from 3.2K publications. The dataset incorporates both TLDRs written by authors and those derived by experts. The expert-derived summaries are obtained through a distinctive annotation process designed to reduce the annotation workload while ensuring the production of high-quality summaries.

2.2. Evaluation metrics

To assess the effectiveness and accuracy of summaries generated by various LLMs, we utilized a set of well-established evaluation metrics. The BLEU score is a widely used metric for evaluating machine-generated text across different NLP tasks, including text summarization [22]. It measures how closely a generated summary aligns with one or more reference summaries, providing a quantitative assessment of precision and textual overlap. The BLEU score is calculated by comparing n-grams (sequences of consecutive words or tokens) in the generated summary to those in the reference summaries. Precision is determined by the proportion of matching n-grams, while a brevity penalty is applied to prevent the overvaluation of excessively short summaries. A higher BLEU score, ranging from 0 to 1, indicates a stronger correspondence between the generated and reference summaries, reflecting improved content accuracy and structural coherence.

The ROUGE score evaluates the similarity between a generated text and one or more reference texts by analyzing the overlap of n-grams and word sequences [23]. It comprises several metrics, including ROUGE-N (which considers unigrams and bigrams), ROUGE-L (which focuses on the longest common subsequence), and ROUGE-W (which measures word overlap). A higher ROUGE score, typically ranging

from 0 to 1, signifies a greater alignment between the generated and reference summaries, offering valuable insights into the performance of the summarization model [24].

BERT score utilizes contextual embeddings from the BERT model to measure the similarity between a generated summary and its reference summaries. Designed to capture the nuances of language and context, this metric provides a powerful approach for assessing the quality and relevance of the generated content [25], [26]. By computing these metrics for summaries generated by various LLMs, our goal is to provide a comprehensive assessment of their performance. This evaluation equips researchers and practitioners with valuable insights to make informed decisions when selecting an LLM. Additionally, it serves as a reference for fine-tuning summarization models to better suit specific tasks and datasets.

3. EXPERIMENTAL SETUP

Experiments were conducted for each LLM using a fixed temperature setting of 0.8 and a maximum token length of 80. The study involved summarizing 50 scientific documents. To generate text summaries, LangChain and Hugging Face pipelines were utilized for prompt engineering, ensuring accuracy and efficiency throughout the summarization process. The experiments were carried out utilizing a Google Colab Notebook, which was equipped with T4 GPUs. Additionally, a Kaggle Notebook with a GPU P100 accelerator was employed for the experiments. The execution involved the utilization of an OpenAI API key and the OpenRouter playground for recently launched models.

4. INFERENCE WITH DIVERSE LLMS

In our study on abstractive scientific document summarization, we observed a strong correlation between recall and the model's effectiveness in capturing key points and essential information. Recall, also known as sensitivity, evaluates the model's ability to accurately identify and incorporate all relevant details from the original document into the generated summary. A higher recall indicates that the model is proficient in capturing important content, even if it means including some non-essential details. Given our focus on extracting critical concepts and insights from scientific documents, prioritizing recall emerges as a key objective, ensuring comprehensive coverage of relevant information throughout the summarization process.

However, for a well-rounded evaluation, it is crucial to also consider precision and the F1 score. Precision assesses the accuracy of the generated summary by determining the proportion of correctly identified relevant information relative to the total predicted relevant content. The F1 score, calculated as the harmonic mean of precision and recall, offers a balanced measure by accounting for both false positives and false negatives. By incorporating recall, precision, and the F1 score, this comprehensive approach aligns with our objective of extracting key concepts and insights from scientific documents while ensuring the accuracy and relevance of the summaries.

We found that models such as 'Claude v2.1,' 'PPLX 70B Online,' and 'Mistral 7B Instruct' demonstrate exceptionally high recall values, indicating their proficiency in effectively extracting a substantial amount of relevant information from the source text. This proficiency is supported by remarkable word overlap accuracy, as shown in Figure 1, highlighting their precision in summarization through substantial word alignment.

These models are effective at minimizing the omission of important details in their summaries. Conversely, models with moderate recall, like 'gpt2-xl,' 'falcon-7b-instruct,' and 'Claude v2.0,' capture relevant information reasonably well but do not achieve the highest recall values. These models balance precision and recall in the summarization process. Models with low recall, such as 't5-small,' 'pegasus-xsum,' and 'mpt-7b-instruct,' may have limitations in capturing a substantial portion of relevant information, potentially missing important details in their summaries. Figures 2 to 4 illustrate the recall and F1 scores of these models for ROUGE-1, ROUGE-2, and ROUGE-L.

The evaluation metrics presented in Table 1 reveal that 'Claude v2.1,' 'PPLX 70B Online,' and 'Mistral 7B Instruct' exhibit exceptional BLEU scores, reflecting their proficiency in generating summaries that closely match reference texts across various n-gram orders. Additionally, BERT scores highlight the balance between precision, recall, and F1, with some models showing exceptional overall performance. Notably, 'Claude v1' presents a comparatively lower BERT score, suggesting a potential trade-off between precision and recall. Overall, these metrics provide a nuanced understanding of each LLM's strengths and weaknesses in text summarization tasks.

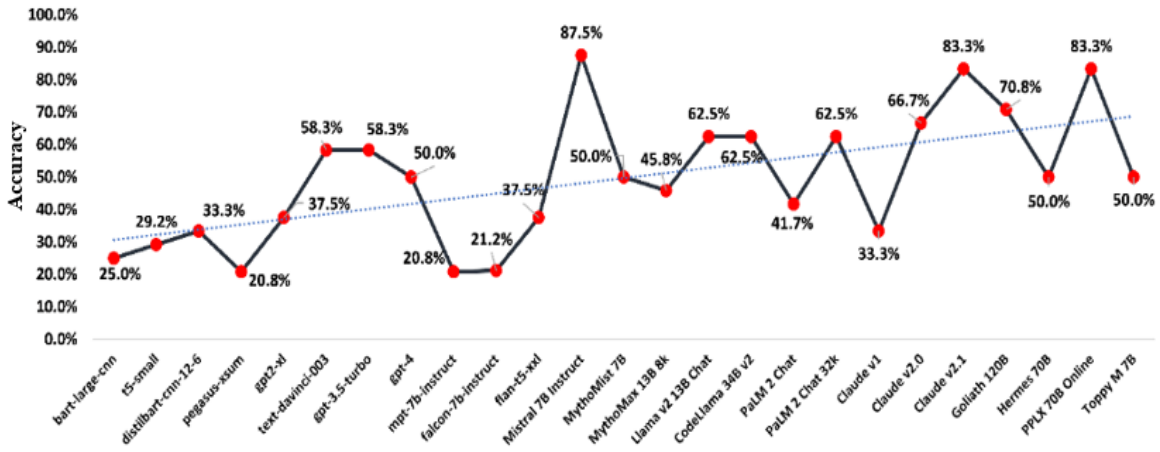

Figure 1. Word overlap accuracy across different LLMs

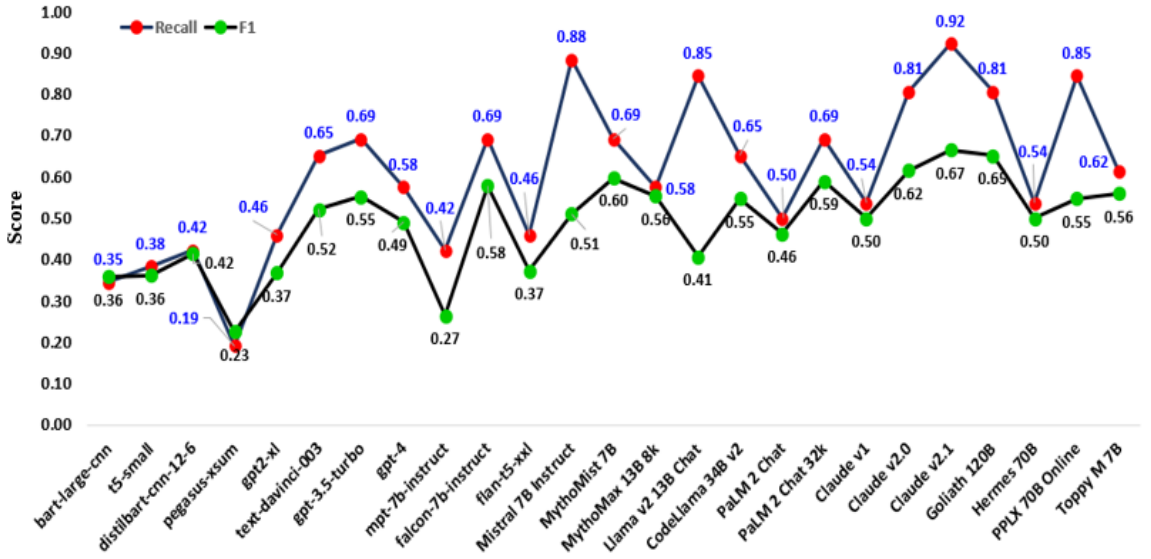

Figure 2. ROUGE-1 recall and F1 score values among different LLMs

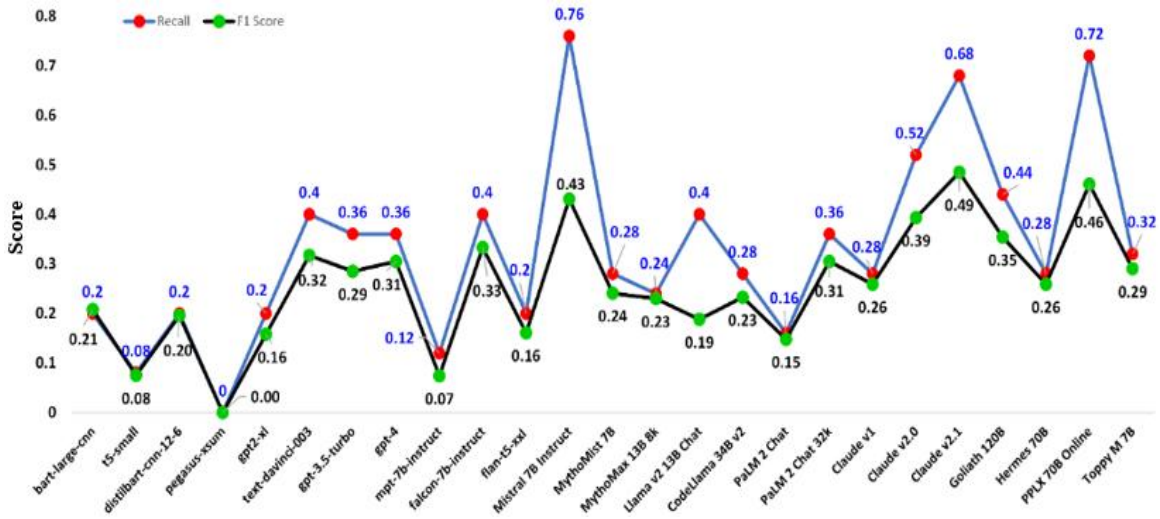

Figure 3. ROUGE-2 recall and F1 score values among different LLMs

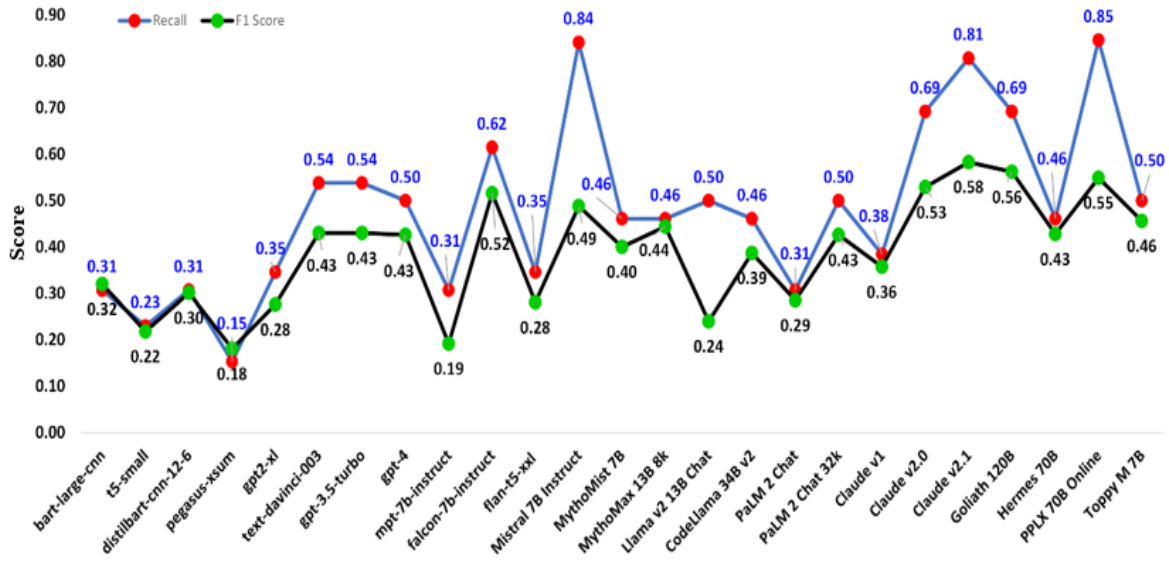

Figure 4. ROUGE-L recall and F1 score values among different LLMs

Table 1. BLEU and BERT scores compared across different LLMs

LLM model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU	BERT score (*P/R/F1)
Bart-large-cnn	0.37	0.123	0.041	0.013	8.84E-232	0.895 / 0.901 / 0.898
T5-small	0.4	0.133	0.044	0.014	1.02E-231	0.860 / 0.857 / 0.859
Distilbart-cnn-12-6	0.4	0.133	0.044	0.014	1.02E-231	0.892 / 0.885 / 0.889
Pegasus-xsum	0.372	0.124	0.041	0.013	8.97E-232	0.865 / 0.881 / 0.873
Gpt2-xl	0.355	0.118	0.039	0.013	7.78E-232	0.901 / 0.875 / 0.888
Text-davinci-003	0.352	0.117	0.039	0.013	7.54E-232	0.931 / 0.901 / 0.916
Gpt-3.5-turbo	0.351	0.116	0.037	0.012	6.85E-232	0.922 / 0.897 / 0.909
Gpt-4	0.355	0.118	0.395	0.013	7.78E-232	0.925 / 0.898 / 0.911
Mpt-7b-instruct	0.346	0.115	0.038	0.012	6.85E-232	0.890 / 0.84 / 0.865
Falcon-7b-instruct	0.353	0.117	0.039	0.013	7.60E-232	0.933 / 0.907 / 0.920
Flan-t5-xxl	0.355	0.118	0.039	0.013	7.78E-232	0.901 / 0.877 / 0.889
Mistral 7B Instruct	0.347	0.115	0.038	0.012	6.92E-232	0.949 / 0.888 / 0.917
MythoMist 7B	0.352	0.116	0.037	0.013	7.83E-229	0.911 / 0.892 / 0.901
MythoMax 13B 8k	0.356	0.115	0.037	0.012	7.88E-232	0.909 / 0.899 / 0.904
Llama v2 13B Chat	0.342	0.11	0.038	0.012	6.23E-232	0.916 / 0.833 / 0.873
CodeLlama 34B v2	0.352	0.117	0.039	0.013	7.49E-232	0.92 / 0.898 / 0.912
PaLM 2 Chat	0.358	0.119	0.039	0.013	7.99E-232	0.898 / 0.882 / 0.890
PaLM 2 Chat 32k	0.354	0.118	0.039	0.013	7.72E-232	0.932 / 0.901 / 0.916
Claude v1	0.354	0.117	0.039	0.013	7.74E-232	0.908 / 0.577 / 0.894
Claude v2.0	0.351	0.117	0.039	0.013	7.36E-232	0.938 / 0.893 / 0.915
Claude v2.1	0.35	0.116	0.038	0.012	7.20E-232	**0.968 / 0.902 / 0.934**
Goliath 120B	0.354	0.118	0.039	0.013	7.65E-232	0.957 / 0.920 / 0.938
Hermes 70B	0.358	0.119	0.039	0.013	7.89E-232	0.911 / 0.887 / 0.899
PPLX 70B Online	0.348	0.116	0.038	0.012	7.07E-232	**0.955 / 0.894 / 0.925**
Toppy M 7B	0.346	0.117	0.037	0.013	7.37E-232	0.916 / 0.901 / 0.909

*P-Precision, R-Recall, F1-F1 score

5. CONCLUSION AND FUTURE WORK

This study investigated the use of various LLMs for summarizing scientific documents. While previous research has explored different summarization techniques, it has not specifically addressed their impact on summary coherence and the thorough inclusion of key details. Our findings reveal a strong correlation between recall and the quality of the generated summaries. Advanced LLMs such as 'Claude v2.1,' 'PPLX 70B Online,' and 'Mistral 7B Instruct' demonstrated greater effectiveness than earlier models in capturing essential information. Additionally, recent observations indicate that improved summarization quality is associated with enhanced model recall, rather than merely an increase in the number of generated tokens. Our results provide compelling evidence that higher recall metrics contribute to more effective summaries. Future research could explore ensemble approaches to harness the strengths of multiple models for even better summarization outcomes.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Surabhi Anuradha	✓	✓			✓	✓			✓	✓	✓		✓	
Martha Sheshikala					✓		✓			✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [Surabhi], upon reasonable request.

REFERENCES

[1] A. Ghadimi and H. Beigy, "Hybrid multi-document summarization using pre-trained language models," *Expert Systems with Applications*, vol. 192, Apr. 2022, pp. 116292, doi: 10.1016/j.eswa.2021.116292.
[2] L. Basyal and M. Sanghvi, "Text summarization using large language models: a comparative study of MPT-7b-instruct, Falcon-7b-instruct, and OpenAI Chat-GPT Models," *arXiv preprint*, Oct. 2023, arXiv:2310.10449.
[3] M. Lewis, *et al.*, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint*, Oct. 2019, arXiv:1910.13461.
[4] OpenAI, GPT-3.5, text-davinci-003. https://platform.openai.com/docs/models/gpt-3-5. (Accessed: Dec, 4 2023).
[5] K. Wiggers, "OpenAI releases GPT-4, a multimodal AI that it claims is state-of-the-art," *TechCrunch*, 2023, Accessed 15 March 2023.
[6] MosaicML NLP Team (2023), "Introducing MPT-7B: a new standard for open-source, commercially usable LLMs," https://www.mosaicml.com/blog/mpt-7b. Accessed 4 December 2023
[7] G. Penedo *et al.*, "The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data only," *Advances in Neural Information Processing Systems*, vol. 36, Dec. 2023, pp. 79155-79172.
[8] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, and W. E. Sayed, "Mistral 7B," *arXiv preprint*, Oct. 2023, arXiv:2310.06825.
[9] H. Touvron *et al.*, "Llama 2: open foundation and fine-tuned chat models," *arXiv preprint*, Jul. 2023, arXiv:2307.09288.
[10] O. Khattab *et al.*, "DSPy: compiling declarative language model calls into self-improving pipelines," *arXiv preprint*, 2023, doi: 10.48550/arXiv.2310.03714.
[11] T. S. Wang and A. S. Gordon, "Playing story creation games with large language models: experiments with GPT-3.5," *International Conference on Interactive Digital Storytelling*, Cham: Springer Nature Switzerland, pp. 297-305, 2023, doi: 10.1007/978-3-031-47658-7_28.
[12] J. Cusidó, L. Solé-Vilaró, P. Marti-Puig, and J. Solé-Casals, "Assessing the capability of advanced AI models in cardiovascular symptom recognition: a comparative study," *Applied Sciences*, vol. 14, no. 18, p.8440, 2024, doi: 10.3390/app14188440.
[13] Grephy, MythoMist 7B, grephy/mythomyst-7b (2023). [Online]. Available: https://openrouter.ai/models/gryphe/mythomist-7b. (Accessed: Dec, 4 2023)
[14] M. R. C. Qazani, H. Asadi, S. Mohamed, S. Nahavandi, J. Winter and K. Rosario, "A real-time motion control tracking mechanism for satellite tracking antenna using serial robot," *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Melbourne, Australia, 2021, pp. 1049-1055, doi: 10.1109/SMC52423.2021.9658909.
[15] Google Team, PaLM 2 Code Chat 32k, google/palm-2-codechat-bison-32k (2023). [Online]. Available: https://openrouter.ai/models/google/palm-2-codechat-bison-32k. (Accessed: Dec, 4 2023)
[16] L. Deng *et al.*, "Evaluation of large language models in breast cancer clinical scenarios: a comparative analysis based on ChatGPT-3.5, ChatGPT-4.0, and Claude2," *International Journal of Surgery*, vol. 110, no. 4, Apr. 2024, pp. 1941-1950, doi: 10.1097/JS9.0000000000001066.
[17] Anthropic Team, Anthropic Completion Models (2023). [Online]. Available: https://clarifai.com/anthropic/completion/models. Accessed 4 December 2023
[18] Y. Ou, Z. Hui, T. Zhou, Y. Cai, and J. Li, "Llama2-13b-based NEFT fine-tuning for financial sentiment classification," *Proceedings of the 2024 Guangdong-Hong Kong-Macao Greater Bay Area International Conference on Digital Economy and Artificial Intelligence*, Jan. 2024, pp. 641-644, doi: 10.1145/3675417.3675523.
[19] W.L. Chiang *et al.*, "Chatbot arena: an open platform for evaluating LLMs by human preference," *Forty-first International Conference on Machine Learning*, 2024, doi: 10.48550/arXiv.2403.04132.

[20] M. Shafique, G. Mumtaz, S.Z. Ahmad, and S. Iqbal, "A comparative analysis of AI chatbot performance in IoT environments," *Journal of Computing & Biomedical Informatics*, vol. 7, no. 2, 2024.
[21] I. Cachola, K. Lo, A. Cohan, and D. S. Weld, "TLDR: extreme summarization of scientific documents," *arXiv preprint*, Apr. 2020, arXiv:2004.15011.
[22] 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 of the Association for Computational Linguistics*, Jul. 2002, pp. 311-318.
[23] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "Summeval: re-evaluating summarization evaluation," *Transactions of the Association for Computational Linguistics*, vol. 9, Apr. 2021, pp. 391-409.
[24] C. Y. Lin, "Rouge: a package for automatic evaluation of summaries," in *Text summarization branches out*, Jul. 2004, pp. 74-81.
[25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTscore: evaluating text generation with BERT," *arXiv preprint*, Apr. 2019, arXiv:1904.09675.
[26] W. Yuan, G. Neubig, and P. Liu, "BARTscore: evaluating generated text as text generation," in *Adv. Neural Information Processing Systems Conference (NeurIPS)*, vol. 34, Apr. 2021, pp. 27263-27277, doi: 10.48550/arXiv.1904.09675.

BIOGRAPHIES OF AUTHORS

Surabhi Anuradha is a research scholar at the School of Computer Science and Artificial Intelligence, SR University, Warangal. She also serves as an Associate Professor in Department of CSE (AIML) at Keshav Memorial Institute of Technology, Hyderabad, India. With over two decades of experience in education and administration, she specializes in artificial intelligence, machine learning, deep learning, natural language processing, and generative AI. Her research focuses on generative AI, large language models (LLMs), and visual language models (VLMs). She can be contacted at email: anuradha@kmit.in.

Dr. Martha Sheshikala is currently holding the position of Head and Professor at the School of Computer Science and Artificial Intelligence at SR University in Warangal, India. She earned her Ph.D. in computer science and engineering from K L Educational Foundation, Andhra Pradesh, in March 2018. Dr. Sheshikala's research focuses on areas such as data mining, machine learning, and natural language processing. Her extensive academic contributions include over 50 publications in various national and international journals, conferences, and proceedings. She can be contacted at email: marthakala08@gmail.com.