Comparative Evaluation of Large Language Models for Abstractive Summarization.

Team 5:

ARCHANAA.N
VIFERT JENUBEN DANIEL.V
MOHAMED FAHEEM M
SUWIN KUMAR JDT
KOUSIHIK K

AMRITA VISHWA VIDYAPEETHAM | School of Engineering
AMRITAPURI | BENGALURU | CHENNAI | COIMBATORE

Introduction:

In the era where information overflows from every digital corner, the ability to condense and crystallize knowledge becomes imperative. Our research embarks on this journey, dissecting the prowess of Large Language Models (LLMs) in the domain of abstractive text summarization. We meticulously compare state-of-the-art models such as Google's PEGASUS, Microsoft's ProphetNet, Facebook's BART, and the fine-tuned T5 to unveil their applications in generating concise, coherent summaries. This study not only underscores the models' significance in educational and informational sectors but also sets a new bar in NLP's application in diverse areas like journalism, academia, and beyond.

Abstract

Exploring the frontier of NLP, this study benchmarks the summarization prowess of four leading LLMs: Google's PEGASUS, Microsoft's ProphetNet, Facebook's BART, and T5 (fine-tuned). We scrutinize their performance using METEOR, BLEU, cosine similarity, BERTScore, and ROUGE metrics, revealing their strengths and areas for improvement. Our comprehensive analysis provides vital insights into the current and potential capabilities of LLMs in producing succinct, meaningful summaries, shaping the future of information synthesis.

Literature survey

S.No	Paper Considered	Description of Work Done	Remarks	Published Year
1	**Text Summarization for Big Data Analytics: A Comprehensive Review of GPT-2 and BERT Approaches** (G. Bharathi Mohan, R. Prasanna Kumar, S. Parathasarathy, S. Aravind, K.B. Hanish, G. Pavithria)	Explores text summarization techniques in big data, focusing on GPT-2 and BERT models.	Highlights the growing importance of summarization in managing large volumes of data.	2023
2	**Survey of Text Document Summarization Based on Ensemble Topic Vector Clustering Model** (G. Bharathi Mohan, R. Prasanna Kumar)	Discusses a novel ensemble topic vector clustering for text summarization using semantic analysis.	Emphasizes the value of topic modeling in summarization for clarity in large datasets.	2023
3	**A Comprehensive Survey on Topic Modeling in Text Summarization** (G.B. Mohan, R.P. Kumar)	Provides a thorough examination of topic modeling techniques and their applications in text summarization.	Aligns with the importance of understanding document themes for coherent summary generation.	2022

Literature survey

S.No	Paper Considered	Description of Work Done	Remarks	Published Year
4	**RankT5: Fine-Tuning T5 for Text Ranking with Ranking Losses** (H. Zhuang, Z. Qin, R. Jagerman, K. Hui, J. Ma, J. Lu, J. Ni, X. Wang, and M. Bendersky)	Introduces RankT5, a model leveraging T5's architecture for text ranking, using ranking losses.	Suggests potential for LLMs to adapt to various NLP tasks beyond traditional summarization.	2023
5	**PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization** (J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu)	Proposes a novel pre-training strategy for abstractive summarization using extracted gap-sentences.	Shows the effectiveness of innovative pre-training methods for improving summarization.	2023
6	**BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension** (M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer)	Introduces BART, focusing on denoising techniques for sequence-to-sequence pre-training.	Demonstrates the utility of denoising autoencoders in generating high-quality text.	2022

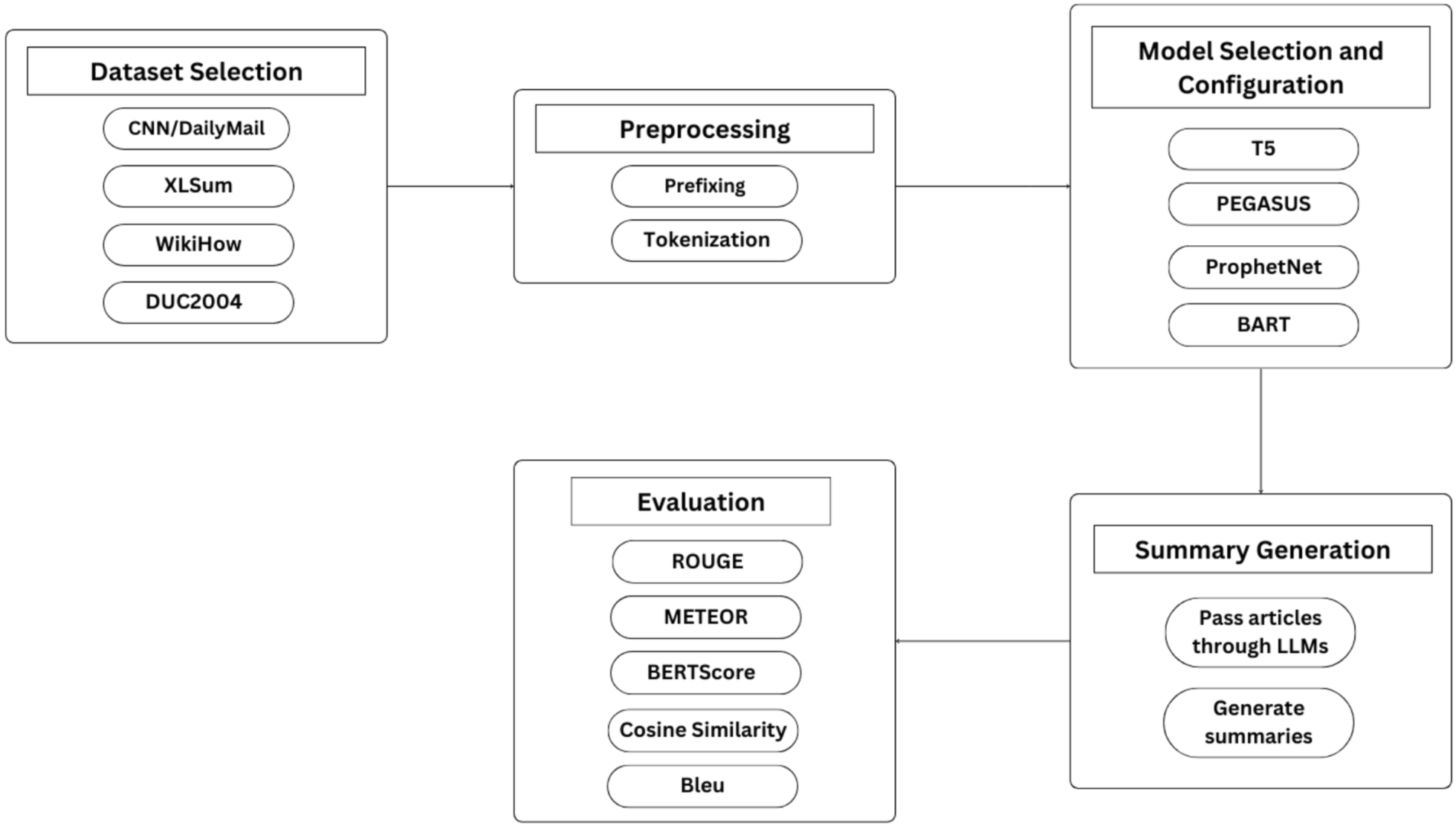
Existing System

In the realm of text summarization, Large Language Models (LLMs) have made remarkable strides, but they are not without limitations. The prevalent systems—Google's PEGASUS, designed for summarization-specific tasks; Microsoft's ProphetNet, with its novel future n-gram prediction; and Facebook's BART, which excels in text generation tasks—have set a strong foundation. Yet, these systems often grapple with challenges like maintaining narrative flow, ensuring factual consistency, and customizing outputs to varied contextual demands. Our study meticulously evaluates these existing models alongside the fine-tuned T5, which has been optimized for summarization, to unravel their capabilities and provide a clear benchmark in the field.

Proposed System

Moving beyond traditional metrics, our proposed system introduces a comprehensive evaluation framework for LLMs in abstractive summarization. We leverage a multi-metric analysis that includes METEOR, BLEU, cosine similarity, BERTScore, and ROUGE to assess the models' performance on four diverse datasets: CNN/DailyMail, XLSum, WikiHow, and DUC2004. This system not only examines the linguistic quality of summaries but also their semantic fidelity, providing a holistic view of model efficiency. Our methodology underscores the nuanced capabilities and identifies potential enhancements in LLMs, propelling forward the field of NLP.

System Architecture



Methodology - Preparation Phase

System Preparation:

- Installation of essential Python libraries for NLP and LLM operations: langchain, sentence-transformers, transformers.
- Configuration of computational environment ensuring compatibility with model requirements.

Dataset Selection:

- Inclusion of varied datasets: CNN/DailyMail, XLSum, WikiHow, DUC2004.
- Rationale for selection based on dataset diversity to challenge the summarization breadth of LLMs.

Methodology - Preparation Phase

Preprocessing:

- Uniform data cleaning steps to standardize input text.
- Tokenization and encoding processes adapted for each LLM's architecture.

Model Configuration:

- Selection of four LLMs: Google's PEGASUS, Microsoft's ProphetNet, Facebook's BART, T5 (fine-tuned).
- Customization of hyperparameters and training setups to align with summarization tasks.

Methodology - Model Training & Summarization Process

- **Model Selection & Rationale**: Detailing the choice of Google's PEGASUS, Microsoft's ProphetNet, Facebook's BART, and T5 (fine-tuned) based on their unique architectural strengths in NLP tasks.
- **Data Preprocessing Steps**: Outlining the preprocessing steps for each dataset, including tokenization and encoding tailored to the input requirements of each LLM.
- **Training Process**: Discussing the training phase, which involves fine-tuning each LLM with the selected datasets to ensure the models are well-adapted for the summarization task.
- **Summary Generation**: Highlighting the process of generating summaries from the input data using each LLM, followed by the procedures for cleaning and standardizing the output for evaluation.
- **Preliminary Evaluation**: Initial assessment using established NLP metrics to measure the linguistic and semantic quality of the generated summaries.

Methodology - In-Depth Evaluation Framework

Comprehensive Metric Application: Implementation of a robust evaluation strategy using a variety of metrics, each providing insights into different aspects of summarization quality:

- **ROUGE Metrics: Assessing overlap with reference summaries to gauge information capture.**
- **METEOR: Going beyond n-gram overlap to account for synonymy, stemming, and word order for a nuanced evaluation.**
- **Cosine Similarity: Measuring the closeness of vector representations between generated and reference summaries, reflecting semantic similarity.**
- **BERTScore: Utilizing contextual embeddings to capture semantic coherence that n-gram metrics may miss.**

Cross-Model Evaluation: Synchronizing the evaluation across different models to compare performance on the same datasets and using the same metrics, ensuring an objective comparison.

Results and Discussion

Metric / Model	T5 Finetuned	Google PEGASUS	Microsoft ProphetNet	Facebook BART
ROUGE-1	0.232	0.189	0.140	0.157
ROUGE-2	0.064	0.035	0.026	0.044
ROUGE-L	0.149	0.122	0.087	0.101
ROUGE-Lsum	0.199	0.163	0.131	0.146
METEOR	0.170	0.185	0.113	0.184
BERT Precision	0.319	0.199	0.162	0.065
BERT Recall	0.215	0.247	0.106	0.227
BERT F1	0.250	0.208	0.126	0.118
Cosine Similarity	0.244	0.237	0.158	0.242
BLEU	3.067	1.848	1.281	2.195

Our evaluation reveals that the fine-tuned T5 model leads in most metrics, showcasing its strong summarization capabilities, particularly in terms of content overlap and precision. PEGASUS and BART exhibit commendable performance in understanding and generating semantically coherent summaries, as indicated by METEOR and Cosine Similarity scores. ProphetNet, while lagging slightly, contributes valuable insights into the role of future n-gram prediction in summarization tasks. These results underscore the importance of tailored model training and the selection of appropriate evaluation metrics to truly capture model efficacy in practical applications.

Conclusion and Future Work

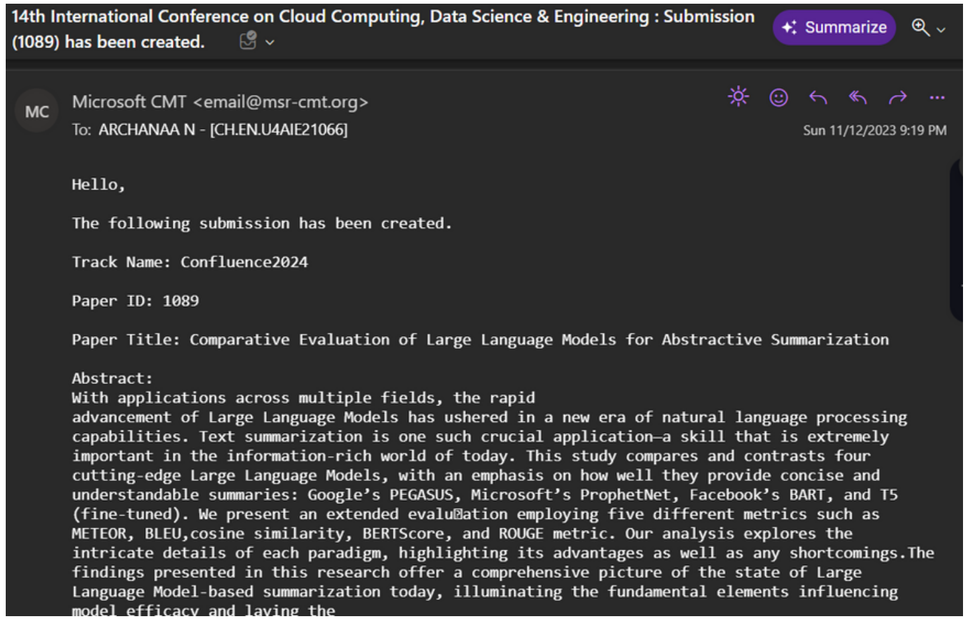
Our investigation provides a detailed analysis of LLMs in text summarization, illustrating T5's superior performance with fine-tuning and the proficiency of PEGASUS and BART in semantic coherence. This study aids in informed model selection, tailored to specific summarization needs.

Key Takeaways:

- T5 excels across metrics, validating the impact of fine-tuning.
- PEGASUS and BART show strengths in semantic tasks.
- ProphetNet's results highlight the complexities of LLM design choices.

Future Directions:

- We envision expanding this research to explore cross-dataset performance and developing hybrid models that integrate the strengths of individual LLMs, driving advancements in summarization technology.



Important Dates

- Paper Submission Deadline :
  15th November 2023
- Acceptance Notification
  20th December 2023
- Camera Ready Paper Submission Deadline :
  30th December 2023
- Last Date of Registration :
  31th December 2023

Confluence-2024:14th International Conference on Cloud Computing, Data Science & Engineering

Microsoft - cmt

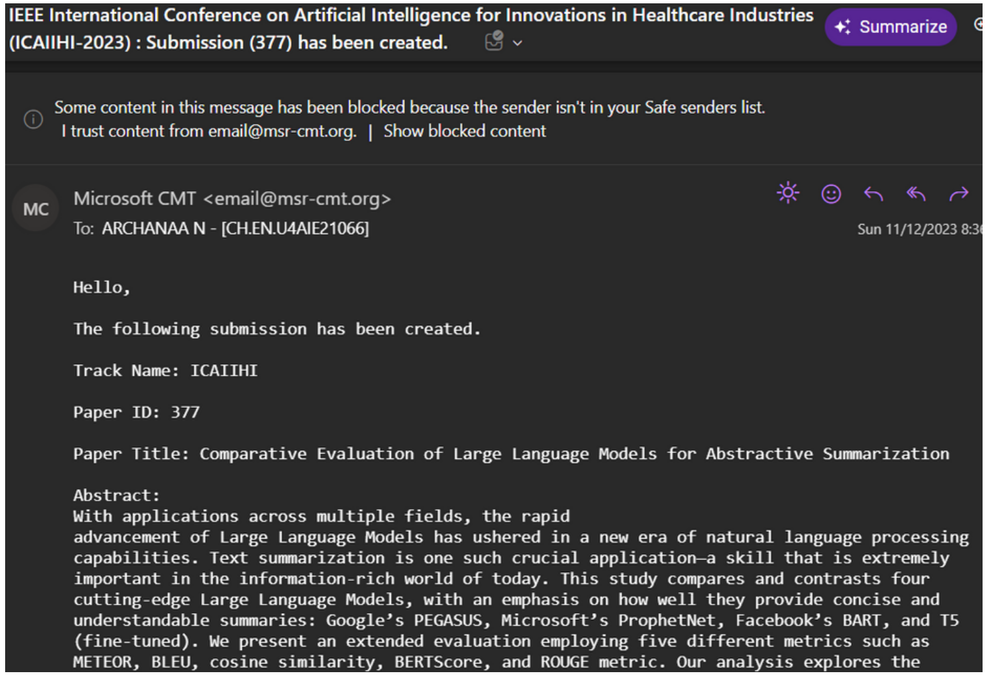
Submission Summary

Conference Name
14th International Conference on Cloud Computing, Data Science & Engineering

Paper ID
1089

Paper Title
Comparative Evaluation of Large Language Models for Abstractive Summarization

Abstract
With applications across multiple fields, the rapid
advancement of Large Language Models has ushered in a new era of natural language processing capabilities. Text summarization is one such crucial application—a skill that is extremely important in the information-rich world of today. This study compares and contrasts four cutting-edge Large Language Models, with an emphasis on how well they provide concise and understandable summaries: Google's PEGASUS, Microsoft's ProphetNet, Facebook's BART, and T5 (fine-tuned). We present an extended evalu□ation employing five different metri such as METEOR, BLEU,cosine similarity, BERTScore, and ROUGE metric. Our analysis explores the intricate details of each paradigm, highlighting its advantages as well as any shortcomings.The findings presented in this research offer a comprehensive picture of the stat of Large Language Model-based summarization today, illuminating the fundamental elements influencing model efficacy and laying the
groundwork for further developments in the area.



IEEE International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI-2023) : Submission (377) has been created.

Summarize

Some content in this message has been blocked because the sender isn't in your Safe senders list.
I trust content from email@msr-cmt.org. | Show blocked content

MC Microsoft CMT <email@msr-cmt.org>
To: ARCHANAA N - [CH.EN.U4AIE21066]
Sun 11/12/2023 8:3

Hello,

The following submission has been created.

Track Name: ICAIIHI

Paper ID: 377

Paper Title: Comparative Evaluation of Large Language Models for Abstractive Summarization

Abstract:
With applications across multiple fields, the rapid
advancement of Large Language Models has ushered in a new era of natural language processing capabilities. Text summarization is one such crucial application—a skill that is extremely important in the information-rich world of today. This study compares and contrasts four cutting-edge Large Language Models, with an emphasis on how well they provide concise and understandable summaries: Google's PEGASUS, Microsoft's ProphetNet, Facebook's BART, and T5 (fine-tuned). We present an extended evaluation employing five different metrics such as METEOR, BLEU, cosine similarity, BERTScore, and ROUGE metric. Our analysis explores the

1st International Conference on Artificial Intelligence for Innovations in Healthcare Industries
(ICAIIHI-2023)
Date: December 29th & 30th, 2023 | Raipur, Chhattisgarh, India
Organized by: Shri Shankaracharya Institute of Professional Management and Technology Raipur, Chhattisgarh, India
In Association with: Inence Publications Pvt Ltd

Important Dates

Paper submission starts from:	20th July 2023
Last date for paper submission (Extended date):	15th November 2023
Notification of acceptance:	25th November 2023
Camera ready paper submission:	1st December 2023
Registration Deadline:	5th December 2023
Release of conference schedule:	10th December 2023
Workshop registration deadline:	10th December 2023
Date of workshops:	28th December 2023
Conference date:	29th & 30th December 2023

Submission Summary

Conference Name
IEEE International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI-2023)

Paper ID
377

Paper Title
Comparative Evaluation of Large Language Models for Abstractive Summarization

Abstract
With applications across multiple fields, the rapid
advancement of Large Language Models has ushered in a new era of natural language processing capabilities. Text summarization is one such crucial application—a skill that is extremely important in the information-rich world of today. This study compares and contrasts four cutting-edge Large Language Models, with an emphasis on how well they provide concise and understandable summaries: Google's PEGASUS, Microsoft's ProphetNet, Facebook's BART, and T5 (fine-tuned). We present an extended evaluation employing five different metrics such as METEOR, BLEU, cosine similarity, BERTScore, and ROUGE metric. Our analysis explores the intricate details of each paradigm, highlighting its advantages as well as any shortcomings.The findings presented in this research offer a comprehensive picture of the stat of Large Language Model-based summarization today, illuminating the fundamental elements influencing model efficacy and laying the groundwork for further developments in the area.

Submission of Research Paper for ICAECT 2024 - Comparative Evaluation of LLMs for Abstractive Summarization

Archanaa N <nkanniga04@gmail.com> Sun, Nov 12, 4:33 PM (3 days ago)
to ieeeicaect

Dear Conference Committee,

I hope you are well. I am excited to submit our research paper for the upcoming ICAECT 2024 conference.

Title: Comparative Evaluation of Large Language Models for Abstractive Summarization

Authors:
Vifert Jenuben Daniel. V
Archanaa.N
Mohammed Faheem
Suwin Kumar. J.D.T
Kousihik. K
Bharathi Mohan G
Prasanna Kumar R

I have attached the full-length research paper in PDF format, following the IEEE double-column format guidelines.I kindly request acknowledgement of the submission, and I eagerly await the peer review process. Thank you for considering our submission, and we look forward to the opportunity to present our research at ICAECT 2024.

2024 Fourth International Conference on Advances in Electrical, Computing, Communications and Sustainable Technologies (ICAECT 2024) scheduled to be held at Shri Shankaracharya Technical Campus (SSTC), Bhilai, Chhattisgarh, India during 11 – 12, January 2024.

Advances in Electrical, Computing, Communications and and Sustainable Technologies (ICAECT 2024)

Important Dates

Last date for paper submission	20, November 2023 (Open)
Notification of acceptance	10, December 2023
Registration deadline	15, December 2023
Camera ready paper submission	20, December 2023
Release of conference schedule	20, December 2023
Workshop registration deadline	25, December 2023
Date of Workshops	10, January 2024
Conference date(s)	11 – 12, January 2024

References

1. G. Bharathi Mohan, R. Prasanna Kumar, S. Parathasarathy, S. Aravind, K.B. Han ish, G. Pavithria, "Text Summarization for Big Data Analytics: A Comprehen sive Review of GPT 2 and BERT Approaches," in Data Analytics for Internet of Things Infrastructure, R. Sharma, G. Jeon, Y. Zhang, Eds. Springer, Cham, 2023, https://doi.org/10.1007/978-3-031-33808-3_14

2. G. Bharathi Mohan, R. Prasanna Kumar, "Survey of Text Document Summariza tion Based on Ensemble Topic Vector Clustering Model," in IoT Based Control Net works and Intelligent Systems, P.P. Joby, V.E. Balas, R. Palanisamy, Eds. Springer, Singapore, 2023, https://doi.org/10.1007/978-981-19-5845-8_60

3. G.B. Mohan, R.P. Kumar, "A Comprehensive Survey on Topic Modeling in Text Summarization," in Micro-Electronics and Telecommunication Engineering, D.K. Sharma, SL. Peng, R. Sharma, D.A. Zaitsev, Eds. Springer, Singapore, 2022, https://doi.org/10.1007/978-981-16-8721-1_22

4. Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen, “A Survey of Large Language Models,” arXiv:2303.18223, 2023.

5. H. Zhuang, Z. Qin, R. Jagerman, K. Hui, J. Ma, J. Lu, J. Ni, X. Wang, and M. Bendersky, “RankT5: Fine-Tuning T5 for Text Ranking with Ranking Losses,” arXiv:2210.10634, 2022.

6. N. A. Kumar, N. Fernandez, Z. Wang, and A. Lan, “Improving Reading Comprehen sion Question Generation with Data Augmentation and Overgenerate-and-rank,” arXiv:2306.08847 , 2023.

7. J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” arXiv:1912.08777 , 2020.

References

8. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “BART: Denoising Sequence-to-Sequence Pre-training for Nat ural Language Generation, Translation, and Comprehension,” Facebook AI.

9. W. Qi, Y. Yan, Y. Gong, D. Liu, N. Duan, J. Chen, R. Zhang, and M. Zhou, “Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training,” arXiv:2001.04063 , 2020.

10. M. Post, “A Call for Clarity in Reporting BLEU Scores,” arXiv:1804.08771 , 2018.

11. A. Lavie and A. Agarwal, “METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments,” 2007.

12. A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, “Cosine similarity to determine similarity measure: Study case in online essay assessment,” 2016 4th In ternational Conference on Cyber and IT Service Management, Bandung, Indonesia, 2016.

13. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Roz ière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lam ple, “LLaMA: Open and Efficient Foundation Language Models,” arXiv:2302.13971 , 2023.

14. R. Zhang, J. Han, C. Liu, P. Gao, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, and Y. Qiao, “LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention,” arXiv:2303.16199 , 2023.

15. C. Xu, D. Guo, N. Duan, and J. McAuley, “Baize: An open-source chat model with parameter-efficient tuning on self-chat data,” arXiv:2304.01196 , 2023

References

16. W. Jiao, J.-t. Huang, W. Wang, X. Wang, S. Shi, and Z. Tu, “ParroT: Translating During Chat Using Large Language Models,” arXiv:2304.02426 , 2023.

17. X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang, “Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data,” arXiv:2307.14385 , 2023.

18. S. Wu, M. Koo, L. Blum, A. Black, L. Kao, F. Scalzo, and I. Kurtz, “A Comparative Study of Open-Source Large Language Models, GPT-4 and Claude 2: Multiple Choice Test Taking in Nephrology,” arXiv:2308.04709 , 2023.

19. H. Luo, Q. Sun, C. Xu, P. Zhao, J. Lou, C. Tao, X. Geng, Q. Lin, S. Chen, and D. Zhang, “WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct,” arXiv:2308.09583 , 2023.

20. A. Kaur, S. Singh, J. S. Chandan, T. Robbins, and V. Patel, “Qualitative ex ploration of digital chatbot use in medical education: A pilot study,” DIGITAL HEALTH, vol. 7, 2021, https://doi.org/10.1177/20552076211038151

21. V. Momodu, “Using Local Large Language Models to Simplify Requirement Engi neering Documents in the Automotive Industry,” Logistics Engineering and Tech nologies Group-Working Paper Series, vol. 4, 2023

Thank You