OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbriviate Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.12666871

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Article

Volume-04|Issue-03|2024

Enhancing Data Curation with LLaMA Model

B Pranesh[1], Hozefa Bohara[2], Mahalakshmi L*[3], Moon Jain[4], Roopashree S[5]

[1,2,3,4,5]Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, Karnataka, India.

Article History
Received: 20.05.2024
Accepted: 05.06.2024
Published: 30.06.2024

Citation
Pranesh, B., Bohara, H., Mahalakshmi, L., Jain, M. & Roopashree, S. (2024). Enhancing Data Curation with LLaMA Model. *Indiana Journal of Multidisciplinary Research*, *4*(3), 10-15.

Abstract: This paper introduces a novel data curation method leveraging the LLaMA. The methodology focuses on converting raw text inputs to proper information, and rectifying grammatical and spelling errors in the input text, using the LLaMA acquired model. This system seamlessly handles diverse input formats, ensuring efficient processing in the standardized docx format. The LLaMA model makes text better by fixing grammar mistakes and making sure words are spelled correctly, through a process called tokenization. The careful handling of the information keeps everything organized and works well with the docx format, creating structured results. This system is not limited to docx; it can also change the curated content into audio, image, text, and PDF formats, making it easier for people to use.

Keywords: Data Curation, LLaMA, Tokenization, Docx Format, Rouge Metrics

INTRODUCTION

In the ever-expanding realm of information management, data curation plays a crucial role in deriving meaningful insights from a wide range of raw data sources. This research paper introduces a novel approach designed to enhance data curation by integrating the LLaMA model. The primary goal is to address grammatical and spelling errors in input text while preserving its original meaning and vocabulary. Through leveraging the capabilities of the LLama model, this methodology provides a comprehensive strategy for handling various input formats, ensuring a smooth transition to a standardized docx format for streamlined processing.

As the landscape of information sources continues to evolve, the demand for advanced curation techniques becomes increasingly apparent. The LLama model plays a pivotal role in this process, facilitating contextual comprehension, content generation, and text refinement. However, due to the model's token limitations, a meticulous tokenization process is employed, initiating a nuanced approach to enhance the curated content's quality.

Beyond addressing grammatical errors, the methodology employs a structured approach to data processing, maintaining coherence and compatibility within the docx format. Notably, the system extends its functionality to support various output formats, such as audio, image, text, and PDF. This adaptability enhances accessibility across platforms, significantly improving the usability and applicability of the curated data.

This paper provides an overview of the Llama model methodology, offering detailed insights into its processes from data input to diverse output formats, with a focus on enhancing data quality and enabling robust analysis. The paper also describes the existing landscape of data curation methodologies and explains the steps involved in the data curation process, emphasizing the role of the LLama model. The evaluation of generated text utilizes Rouge metrics to compare the effectiveness of the LLama model with other language model-based curation outputs and the original curated input.

LITERATURE SURVEY

Yong Yu *et al.*,[1] proposed Recurrent neural networks (RNNs) have gained widespread adoption in research domains dealing with sequential data like text, audio, and video. However, traditional RNNs composed of sigma or tanh cells face challenges in capturing relevant information from input data with substantial time lags. The introduction of gate functions in the cell architecture, as seen in Long Short-Term Memory (LSTM) networks, effectively addresses the issue of long-term dependencies. Since its inception, LSTM has been pivotal in achieving remarkable results across various applications of RNNs, making it a focal point in deep learning.

Schuster *et al.*,[2] proposed a conventional recurrent neural network (RNN) is expanded into a bidirectional recurrent neural network (BRNN), which offers the advantage of being trained without being restricted to utilizing input information only up to a

***Corresponding Author:** Mahalakshmi L

predetermined future frame.

Kyunghyun Cho *et al.*,[3] proposed a novel neural network model known as RNN Encoder–Decoder is introduced. This model comprises two recurrent neural networks (RNNs), where one RNN is responsible for encoding a sequence of symbols into a fixed-length vector representation, while the other RNN decodes this representation into another sequence of symbols.

Kai Han *et al.*,[4] Transformer is a new kind of neural architecture that encodes the input data as powerful features via the attention mechanism. Transformers are large and heavy and require a lot of space. There have been extensive studies to reduce the size and weight of transformers.

Jun Li *et al.*,[5] The fundamental concept behind contemporary deep learning approaches for dense prediction involves applying a model to a regular patch centered on each pixel, facilitating pixel-wise predictions. However, a limitation of these methods is that the patches are predefined by the network architecture rather than being learned from the data. In this study, the researchers introduce dense transformer networks. These networks can learn the shapes and sizes of patches from the data itself. The architecture of the dense transformer networks comprises an encoder-decoder structure, with a pair of dense transformer modules integrated into both the encoder and decoder paths.

Mohammed Sabry *et al.*,[6] Over the past few years, there has been a significant increase in the size of pre-trained language models (PLMs) such as GPT3, OPT BLOOM, and PaLM, which have billions of parameters. This increase in size has been accompanied by a commensurate increase in the cost of training and deploying large PLMs, with substantial financial and environmental implications. Reusing PLMs via adaptation to downstream tasks, rather than training new language models for new tasks, mitigates this cost significantly. However, full finetuning, the default task adaptation approach, is still very costly as it retrains, and subsequently stores, the entire model. As an increasing number of PEFT techniques are reported, it is becoming harder to compare them in terms of efficiency improvements and performance at different tasks, in particular, which aspects of their structure and functionality are linked to better efficiency and performance.

Jun-Ping Ng *et al.*,[7] ROUGE stands as a widely embraced automated evaluation metric utilized for assessing the quality of text summarization. Despite its proven strong correlation with human judgments, ROUGE demonstrates a bias towards surface lexical similarities. The limitation of ROUGE in evaluating abstractive summarization or summaries involving significant paraphrasing is acknowledged. The effectiveness of word embeddings is explored to address this bias. In this study, the researchers investigate the use of word embeddings to overcome ROUGE's bias towards lexical overlaps. Instead of assessing surface-level lexical similarities, word embeddings are employed to calculate the semantic similarity of words used in summaries. The experimental results indicate that this approach achieves improved correlations with human judgments when evaluated using the Spearman and Kendall rank coefficients. The findings suggest that the proposed method mitigates the bias of ROUGE in identifying lexical similarity when assessing the quality of generated summaries.

Elias Frantar *et al.*,[8] Generative Pre-trained Transformer models, known as GPT or OPT, set themselves apart through breakthrough performance across complex language modeling tasks, but also by their extremely high computational and storage costs. Specifically, due to their massive size, even inference for large, highly accurate GPT models may require multiple performant GPUs, which limits the usability of such models. While emerging work is on relieving this pressure via model compression, the applicability and performance of existing compression techniques is limited by the scale and complexity of GPT models. In this paper, we address this challenge and propose GPTQ, a new one-shot weight quantization method based on approximate second-order information, that is both highly accurate and highly efficient.

Hugo Touvron *et al.*,[9] The LLaMA and LLaMA 2 models are instances of Generative Pretrained Transformer (GPT) models, built upon the original Transformers architecture. The LLaMA models employ GPT-3- like pre-normalization, utilizing the RMS Norm normalizing function at the input of each transformer sub-layer. This approach enhances training stability by rescaling the invariance property and implicit learning rate adaptation ability. Additionally, LLaMA benefits from the SwiGLU activation function, replacing the conventional ReLU non-linearity activation function, leading to improved training performance. Incorporating insights from the GPT-Neo-X project, LLaMA incorporates rotary positional embeddings (RoPE) at each layer, contributing to its overall performance. Notably, LLaMA 2 introduces essential architectural differences, . These differences include an increased context length, doubling the context window size from 2048 to 4096 tokens. This extension enables the model to handle more extensive information, proving beneficial for tasks involving long documents, chat histories, and summarization. Furthermore, LLaMA 2 implements a grouped-query attention (GQA) format with eight key value projections, addressing the complexity concerns associated with the original Multi-Head attention baseline. This modification proves effective in managing the increased context windows or batch sizes. As a result of these updates, LLaMA demonstrates significantly improved performance across various tasks, surpassing

***Corresponding Author:** Mahalakshmi L

or closely matching other specialized GPT models such as Falcon and MPT. The model's promising performance paves the way for further research, anticipating future comparisons with prominent closed-source models like GPT-4 and Bard.

Konstantinos I Roumeliotis *et al.*,[10] The rapidly evolving field of artificial intelligence (AI) continues to witness the introduction of innovative open-source pre-trained models, fostering advancements in various applications. One such model is Llama 2, an open-source pre-trained model released by Meta, which has garnered significant attention among early adopters. In addition to exploring the foundational elements of the Llama v2 model, this paper investigates how these early adopters leverage the capabilities of Llama 2 in their AI projects. Through a qualitative study, we delve into the perspectives, experiences, and strategies employed by early adopters to leverage Llama 2's capabilities. For the purpose of data analysis, the capabilities inherent in the Llama 2 model were employed to conduct keyword extraction from the context of the early adopters' case studies. The findings shed light on the model's strengths, weaknesses, and areas of improvement, offering valuable insights for the AI community and Meta to enhance future model iterations. Additionally, we discuss the implications of Llama 2's adoption on the broader open-source AI landscape, addressing challenges and opportunities for developers and researchers in the pursuit of cutting-edge AI solutions. The present study constitutes an early exploration of the Llama 2 pre-trained model, holding promise as a foundational basis for forthcoming research investigations.

Thomas M. Breuel,[11] LSTM networks have become very popular for many sequence classification tasks. This note presents the results of large-scale benchmarking with a wide range of parameters to determine the effects of learning rates, batch sizes, momentum, different non-linearities, and peepholes. The two benchmark datasets are MNIST and the UW3 text line OCR task. Numerous other issues remain to be explored experimentally. For example, we do not know what effect different choices of weight initialization have. Also, several other LSTM-like architectures have been proposed.

Yang Liu *et al.*,[12] proposed the latest iteration of pre-trained language models, known as Bidirectional Encoder Representations from Transformers (BERT; Devlin *et al.*, 2019), is presented. These models have recently made significant strides in various natural language processing tasks. The paper demonstrates the practical application of BERT in text summarization and introduces a comprehensive framework for both extractive and abstractive summarization models.

Mahmood Yousefi Azar *et al.*,[13] This paper presents methods for extractive query-oriented single-document summarization using a deep autoencoder (AE) to compute a feature space from the term-frequency (tf) input. The researchers explore both local and global vocabularies in their experiments. They examine the impact of introducing small random noise to the local term frequency (tf) as the input representation of the autoencoder and propose an ensemble of such noisy autoencoders, referred to as the Ensemble Noisy Auto-Encoder (ENAE). A potential application of semi-supervised learning techniques is suggested due to the limited availability of manually annotated data, or noisy labeling could be employed with the unlabeled data. The authors recommend exploring other ensemble approaches to enhance performance and accuracy.

Tim Dettmers *et al.*,[14] Finetuning large language models (LLMs) is a highly effective way to improve their performance, and to add desirable or remove undesirable behaviors. However, finetuning very large models is prohibitively expensive; regular 16-bit finetuning of a LLaMA 65B parameter model requires more than 780 GB of GPU memory. While recent quantization methods can reduce the memory footprint of LLMs, such techniques only work for inference and break down during training. We demonstrate for the first time that it is possible to finetune a quantized 4-bit model without any performance degradation. Our method, QLORA, uses a novel high-precision technique to quantize a pre-trained model to 4-bit, then adds a small set of learnable Low-rank Adapter weights.

Adaku Uchendu et al,[15][Advances in Large Language Models (e.g., GPT-4, LLaMA) have improved the generation of coherent sentences resembling human writing on a large scale, resulting in the creation of so-called deep fake texts However, this progress poses security and privacy concerns, necessitating effective solutions for distinguishing deep fake texts from human-written ones. Although prior works studied humans' ability to detect deep fake texts, none has examined whether "collaboration" among humans improves the detection of deep fake texts.

METHODOLOGY

The data curation process within our work revolves around rectifying grammatical and spelling errors in the provided input text while preserving its original meaning and vocabulary. Our chosen tool for this task is the Llama model.

The method adopts a comprehensive strategy to manage diverse input formats, ensuring smooth conversion to and from the docx format. Initially, the system is engineered to accommodate a wide array of input types, ensuring a seamless transition to the standardized docx format for efficient processing.

Following this, fetching the LLama model is performed. This pre-trained and quantized model serves as the pivotal component of our methodology, facilitating the curation process.

***Corresponding Author:** Mahalakshmi L

An important consideration for the LLama model is its token limitations. Once the model is obtained, we commence a tokenization process for data formatted in docx. This crucial step initiates contextual comprehension and content generation using the LLama model. The LLama 2 model adeptly refines the text by automatically rectifying grammatical errors and ensuring precise spelling, thereby enhancing the quality of the provided text for the given prompt.

Following execution, the output data undergoes meticulous processing to maintain coherence and compatibility within the docx format. This structured approach guarantees that the generated outputs are well-structured, facilitating effective analysis and interpretation. Additionally, the removal of extraneous or superfluous text generated by the model is carried out.

Moreover, this system extends its functionality beyond the confines of the docx format, enabling the conversion of processed outputs into various formats, such as audio, image, text, and PDF. This diverse compatibility enhances accessibility across multiple platforms and user preferences, significantly elevating the overall usability and applicability of our methodology.

Through these integrated steps, our primary goal is to ensure the efficient handling and transformation of data across a spectrum of formats, thereby fostering robust analysis and ensuring accessible information.

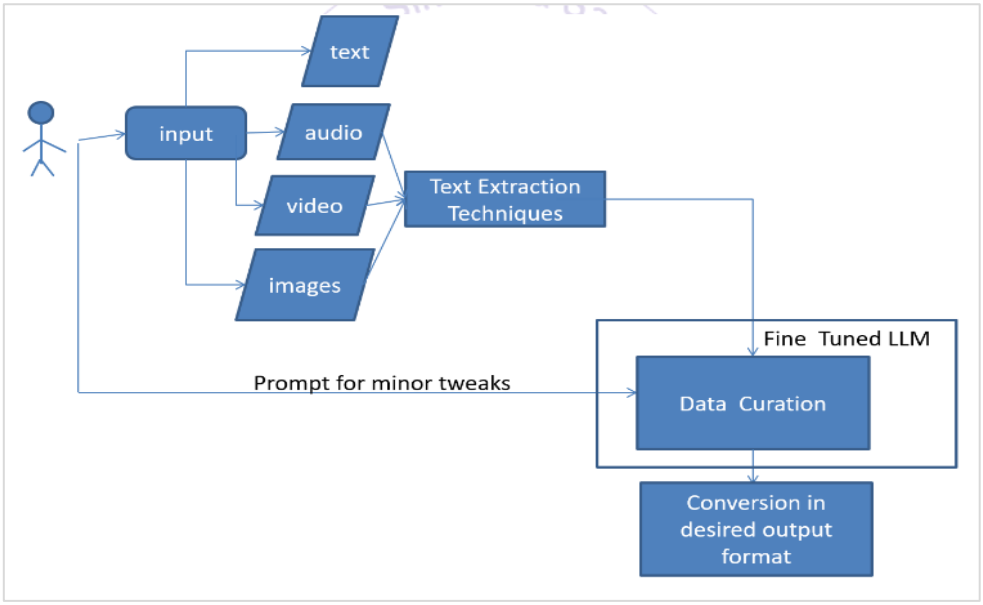

Figure 1. Data Flow Diagram

Figure 1 depicts the data flow diagram of the implemented methodology.

RESULTS

The Rouge metrics, including recall, precision, and F1 score, are used to compare the result obtained from this LLM with various other LLM outputs and with the actual input that is curated.

Rouge metrics are used to compare the generated text with the target or reference text and return a value from 0 to 1. The value 0 represents the least similarity and 1 represents the highest similarity. There are 3 different rouge metrics used i.e. Rouge 1, Rouge 2, and Rouge L. Rouge 1 measures the overlap of unigrams or single words between the generated text and the target/reference text. Rouge 2 measures the overlap of bigrams (pairs of words) between the generated text and the reference text. Rouge-L is similar to Rouge-1 but uses the longest common subsequence between the generated and reference texts, considering sentence-level structure more than just word overlap.

Each Rouge metric assesses the model's performance by calculating recall, precision, and F1 scores across different text lengths. These metrics determine how effectively the model generates text that aligns with the reference text, evaluating content, grammatical accuracy, and structural coherence.

***Corresponding Author:** Mahalakshmi L

Table 1. ROUGE scores for Bard, ChatGPT, and LLama models

Metric	Bard	ChatGPT	LLama
ROUGE-1 Prec.	0.5726	0.6099	0.5882
ROUGE-1 Recall	0.6158	0.5903	0.5667
ROUGE-1 F1	0.5936	0.6000	0.5773
ROUGE-2 Prec.	0.375	0.3188	0.4031
ROUGE-2 Recall	0.4073	0.3061	0.3866
ROUGE-2 F1	0.3907	0.3123	0.3948
ROUGE-L Prec.	0.5227	0.4978	0.5477
ROUGE-L Recall	0.5632	0.4809	0.5273
ROUGE-L F1	0.5424	0.4892	0.5373

The values of the ROUGE scores for Bard, ChatGPT, and Llama models are displayed in Table 1.

DISCUSSION

The comparison of various Language Model Models (LLMs) against the Llama model in terms of Rouge metrics sheds light on crucial performance differences. In comparison to the LLama model, "Bard" exhibits competitive performance across various ROUGE metrics. While "Bard" demonstrates slightly higher precision and recall scores in ROUGE-1 metrics, LLama model performs marginally better in ROUGE-2 F1 score. However, "Bard" maintains an advantage in capturing long-range dependencies, as evidenced by its higher ROUGE-L recall and F1 scores. Overall, while "Bard" shows a slight edge over Llama model concerning precision, recall, and F1 score, the differences between the two models indicate comparable performance in summarization tasks.

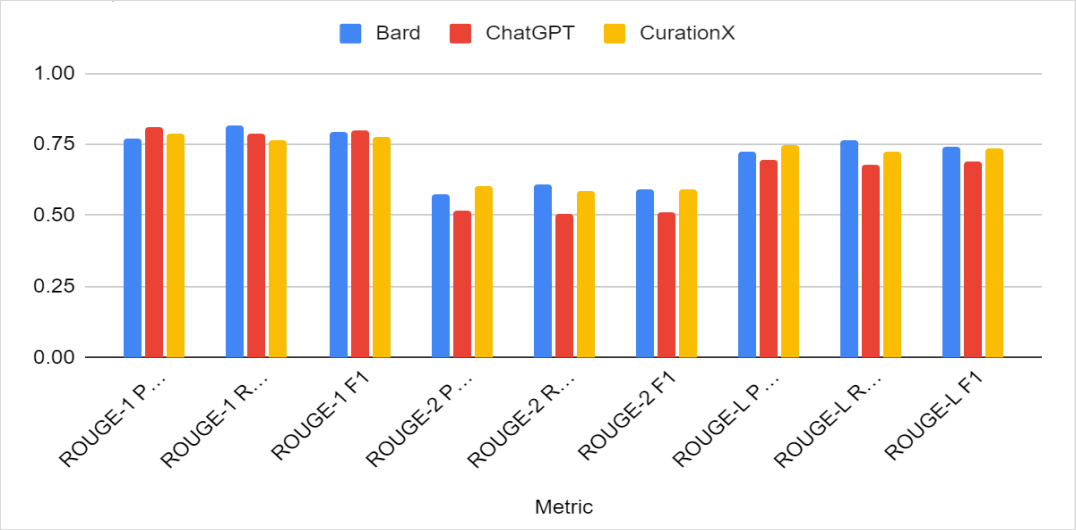

Figure 2. Comparing ROUGE scores of Bard, ChatGPT, and CurationX models

Figure 2 Displays a bar graph that shows the comparison of ROUGE scores for Bard, ChatGPT, and LLama models.

CONCLUSION

The methodology, powered by the LLaMA model, represents a significant stride in the field of data curation. By successfully addressing grammatical and spelling errors while preserving the text's original meaning and vocabulary, this approach marks a notable improvement over existing methodologies. The use of Rouge metrics for performance evaluation underscores the model's effectiveness in producing high-quality, coherent, and contextually accurate text outputs. The adaptability of model to various input and output formats further enhances its utility across different platforms and use cases. Its ability to maintain coherence and compatibility within these formats ensures that the curated data is not only of high quality but also versatile in its applications.

The major drawback is the disparity in performance which highlights a critical challenge: the

***Corresponding Author:** Mahalakshmi L

need for relatively high-end computing resources to leverage the full potential of advanced AI models like the LLaMA quantized model. The increased latency on older systems can hinder the widespread adoption and scalability of the model approach, particularly in environments where access to cutting-edge hardware is limited or cost-prohibitive.

FUTURE SCOPE

The future scope of the methodology, leveraging the LLaMA model, includes several promising directions. Key areas for advancement encompass enhancing the model to support multilingual processing, thereby broadening its global applicability. Additionally, integrating an automated quality assessment system would streamline the curation process by providing real-time feedback on content quality. Finally, addressing hardware dependency and optimizing the model for varying computational resources will also be crucial, ensuring wider accessibility and usability.

ACKNOWLEDGEMENTS

We would like to extend our heartfelt gratitude to our Institution for providing the necessary resources, facilities, and support that enabled the successful completion of this research paper. We are also deeply appreciative of Guide our esteemed guide, for their invaluable guidance, mentorship, and encouragement throughout this project. Their expertise, insights, and unwavering support have been instrumental in shaping the direction and outcomes of this research.

REFERENCES

1. Yu, Y., Si, X., Hu, C., & Zhang, J. (2019b). A review of Recurrent Neural networks: LSTM cells and network architectures. Neural Computation, 31(7), 1235–1270. https://doi.org/10.1162/neco_a_01199
2. Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673–2681. https://doi.org/10.1109/78.650093
3. Han, K., Xiao, A., Wu, E., Guo, J., Xu, C., & Wang, Y. (2021, February 27). Transformer in transformer. arXiv.org. https://arxiv.org/abs/2103.00112
4. Li, J., Chen, Y., Cai, L., Davidson, I., & Ji, S. (2017, May 24). Dense transformer networks. arXiv.org. https://arxiv.org/abs/1705.08881
5. Sabry, M., & Belz, A. (2023b). PEFT-ReF: A Modular Reference Architecture and Typology for Parameter-Efficient Finetuning Techniques. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2304.12410
6. Ng, J., & Abrecht, V. (2015, August 25). Better Summarization Evaluation with Word Embeddings for ROUGE. arXiv.org. https://arxiv.org/abs/1508.06034
7. Frantar, E., IST Austria, Ashkboos, S., ETH Zurich, Hoefler, T., ETH Zurich, Alistarh, D., & IST Austria & NeuralMagic. (2023). GPTQ: ACCURATE POST-TRAINING QUANTIZATION FOR GENERATIVE PRE-TRAINED TRANSFORMERS. In ICLR 2023. https://arxiv.org/pdf/2210.17323.pdf
8. Touvron, H., Martin, L., Stone, K. H., Albert, P. J., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., . . . Scialom, T. (2023). Llama 2: Open foundation and Fine-Tuned chat models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2307.09288
9. Konstantinos I Roumeliotis, Nikolaos D Tselikas and Dimitrios K Nasiopoulos (2023), "Llama 2: Early Adopters' Utilization of Meta's New Open-Source Pretrained Model".Preprints https://www.preprints.org/manuscript/202307.2142/v1
10. Breuel, T. M. (2015, August 11). Benchmarking of LSTM networks. arXiv.org. https://arxiv.org/abs/1508.02774
11. Liu, Y., & Lapata, M. (2019, August 22). Text Summarization with Pretrained Encoders. arXiv.org. https://arxiv.org/abs/1908.08345
12. Yousefi Azar, M., Jr., Sirts, K., Moll´a Aliod, D., Hamey, L., Department of Computing, & Macquarie University. (2015). Query-Based single document summarization using an ensemble noisy Auto-Encoder. In Proceedings of Australasian Language Technology Association Workshop (pp. 2–10). https://aclanthology.org/U15-1001.pdf
13. Uchendu, A., Lee, J., Su, H., Le, T. H., Huang, T., & Lee, D. (2023). Does human collaboration enhance the accuracy of identifying LLM-Generated deepfake texts? arXiv (Cornell University). https://doi.org/10.48550/arxiv.2304.01002
14. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L., & University of Washington. (2023). QLORA: Efficient Finetuning of Quantized LLMs. Preprint. https://arxiv.org/pdf/2305.14314.pdf

***Corresponding Author:** Mahalakshmi L