Grenze International Journal of Engineering and Technology, January Issue

Neural Network based Approach for Content Generation: Comprehensive Analysis of LLM Models

Spandan Marathe[1], Vijayendra S. Gaikwad[2], Atharva Khodke[3], Karan Shardul[4] and Rohit Kuvar[5]
[1-5]Dept. of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, India
Email: spandanmarathe95@gmail.com, vij711, atharva.j.khodke, karanshardul54, rohit.k.204086@gmail.com

***Abstract*— This paper presents a detailed comparative analysis of Large Language Models (LLMs) for automated content creation on social media, focusing on GPT-3.5, LLaMA 3.1, Mixtral 8x7B, Mistral Nemo 12B, and Gemma 2 (9B). As social media continues to evolve, creating platform-specific, engaging, and scalable content has become critical for both personal and business application. The study examines the selection criteria for these models, including scalability, support for multiple languages, and ethical content moderation, while highlighting the importance of fine-tuning and prompt engineering. Furthermore, this paper evaluates the performance of LLMs on both short and long-form content, exploring their effect on user engagement, retention, and factual accuracy. Using evaluation metrics such as ROUGE, sentiment analysis, and readability scores, the paper aims to highlight the strengths and weaknesses of each model, providing a basis for optimizing content creation for social media platforms.**

***Index Terms*— Large Language Models, social media platforms, content creation, prompt engineering.**

I. INTRODUCTION

Social media is a vital communication tool, but manual content creation across platforms is time-consuming and inconsistent. Large Language Models (LLMs) like GPT-3.5, LLaMA 3.1, Mixtral, Mistral Nemo, and Gemma 2 offer a scalable solution for generating human-like text, though their performance varies by content type and platform. This paper analyzes these LLMs in creating platform-specific content—short, impactful posts for Twitter, professional discussions for LinkedIn, and long-form articles for Medium.

Prompt engineering optimizes LLM output using structured inputs without retraining. Techniques like zero-shot and few-shot prompting, as well as advanced methods such as Chain-of-Thought and Self-Consistency, enhance contextual flow, especially in longer content. Platform-specific prompting helps match tone and format—concise prompts for Twitter, formal structures for LinkedIn.

However, prompt engineering has limits in domain knowledge and consistency. Partial fine-tuning methods like Domain-Adaptive Fine-Tuning and LoRA provide targeted improvements using platform-specific data. This helps models learn tone, structure, and cultural norms—boosting professionalism on LinkedIn or virality on Twitter. Multilingual and emotional fine-tuning further support global and marketing needs.

This study evaluates LLM performance across formats, combining prompt engineering and fine-tuning to build a framework for content optimization.

Grenze ID: 01.GIJET.12.1.336

Insights support content creators, marketers, AI researchers, and businesses. The study also lays groundwork for future system-level research in adaptive models, platform integration, and personalized content generation.

II. Background and Related Work

Language model development has transformed natural language processing (NLP), moving from rule-based systems to advanced neural networks [1][2]. Early models used grammar rules and dictionaries but lacked flexibility. Statistical Language Models (SLMs) improved prediction using word probabilities but failed at long-range context [3][4]. Word embeddings like Word2Vec (2010s) introduced continuous vector representations, improving semantic understanding [5]. RNNs and LSTMs further enhanced sequential processing and context retention [6].

The 2017 Transformer architecture by Vaswani et al. revolutionized NLP through attention mechanisms, enabling better context handling [7]. This led to powerful models like GPT-3 with 175B parameters [8] and BERT, which used bidirectional training for improved comprehension [9]. GPT-4 added advancements in reasoning [10], while multimodal models like LLaMA and PaLM integrated diverse data for broader understanding [11][12]. These developments raised concerns around fairness and bias, necessitating ethical oversight [13].

LLMs such as GPT-4, BERT, and LLaMA are reshaping social media by enabling human-like text generation [14][15]. They streamline post creation, summarize content [16][17], perform real-time sentiment analysis [18], and power responsive chatbots [19]. LLMs support social listening [20], enhance user engagement through personalized content, assist in moderation, and optimize influencer-brand pairing. These capabilities demonstrate how LLMs are redefining content strategy and interaction across platforms.

III. Model Selection for Social Media Content Generation

Choosing the right Large Language Model (LLM) for social media content involves evaluating fluency, speed, cost-efficiency, and adaptability. High-performing models like GPT-4, LLaMA, and PaLM generate coherent, context-aware content suitable across platforms. Cost is a major consideration—cloud solutions offer scalability, while open-source models such as LLaMA, GPT-Neo, and Bloom reduce licensing costs. For real-time responsiveness, GPU-optimized models like GPT-3 Turbo are effective in fast-paced environments. Platform-specific constraints, such as Twitter's brevity or LinkedIn's formal tone, demand tailored outputs, which customizable models help achieve. GPT-4 supports multilingual generation and scales well during traffic surges, while API-based integration and built-in moderation features make adoption easier and safer.

Among the evaluated models, GPT-3.5 Turbo performs well across formats with a conversational tone but can be overly verbose for short-form posts unless fine-tuned. Mixtral 8x7B, a Sparse Mixture of Experts model, balances performance and efficiency with a large context window, though it may encounter latency during high-volume use. Mistral Nemo 12B excels at scalable, multilingual generation but may require fine-tuning for personalized tasks and can face inference bottlenecks. Gemma 2 9B offers fast, structured content generation ideal for platforms like LinkedIn and Medium, though it may struggle with deeper contextual tasks. LLaMA 3.1 8B enables quick multilingual content creation for platforms like Twitter but has limited performance in generating complex, threaded content and may need additional tuning for trend-specific alignment.

IV. Methodology

For analyzing, training, and evaluating language models used for content generation, we developed a structured methodology. A flowchart (see Fig. 1) illustrates the key steps, which are detailed below. Each step's importance in achieving high-performing models is highlighted.

Step 1: Data Collection. High-quality and diverse training data is crucial for model performance. We sourced data from YouTube videos, blogs, poetry, and summaries of academic papers and news articles. This variety helped the models learn from conversational, structured, creative, and technical content, enhancing their ability to respond to different user inputs.

Step 2: Data Pre-processing. To ensure consistent input, we transcribed and normalized audio, removed noise and offensive content, and applied tokenization and lemmatization. Context window optimization and data augmentation further refined the dataset, improving model adaptability in both creative and technical domains.

Step 3: Model Training. Using a blend of general and domain-specific datasets, we fine-tuned models for contextual versatility. Transfer learning from pre-trained models reduced training time while maintaining performance. This step ensured efficient and high-quality content generation.

Step 4: Performance Evaluation. Model performance was assessed using benchmark datasets like SQuAD, QuAC, CommonSenseQA, and GSM8K, focusing on accuracy, coherence, and logical reasoning. A comparative analysis of models like LLaMA 3.1 (8B), Gemma 2 (9B), Mixtral 8x7B, and Mistral Nemo 12B is presented (see Table 1, Fig. 2).

Step 5: Feedback and Iteration. Evaluation results guided iterative fine-tuning. Feedback loops revealed weaknesses in areas like commonsense reasoning and creativity, which were addressed through targeted retraining. This ensured continuous performance improvement across iterations.

This methodology ensures that the language models are systematically developed, optimized, and refined for diverse real-world content generation tasks.

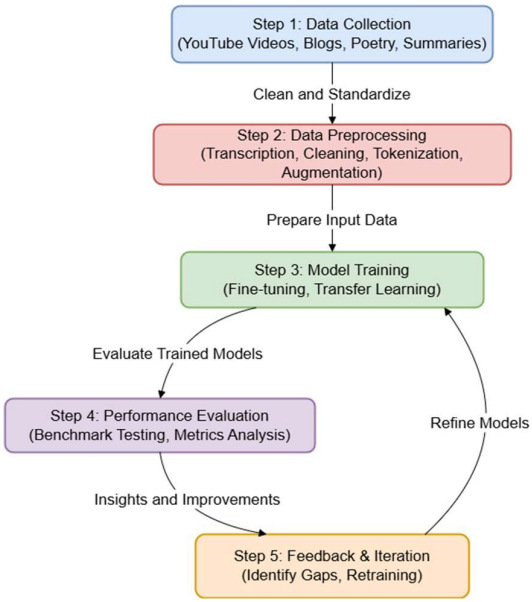


Figure 1. Methodology Flow Diagram

TABLE I. PERFORMANCE COMPARISON OF LANGUAGE MODELS ACROSS NLP BENCHMARKS AND CONTEXT WINDOW SIZES

Dataset	LLaMA 3.1 8B	Gemma 2 9B	Mixtral 8x7B	Mistral Nemo 12B
SQuAD	77.0	81.8	84.1	73.2
QuAC	44.9	42.4	44.9	44.7
RACE	54.3	48.8	59.2	53.0
CommonSenseQA	75.0	74.4	82.4	71.2
PiQA	81.0	81.5	85.5	83.0
OpenBookQA	45.0	52.8	50.8	47.8
Winogrande	75.7	80.6	84.7	78.1
GSM8K (8-shot, CoT)	66.7	64.3	70.6	62.5
MBPP (Pass@1)	37.2	47.6	60.7	47.5
Context Window	8k	8k	13k	128k

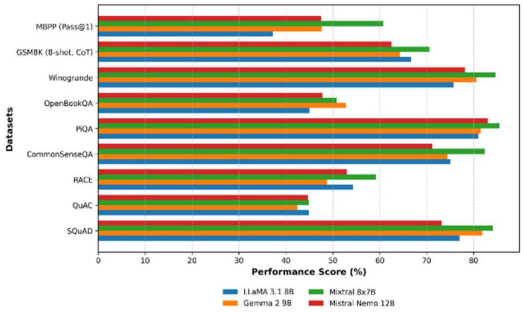


Figure 2. Performance of different Models on Various Datasets

VI. Results and Discussions

A. Performance Metrics for Content Generation

Evaluating content generation for formats like social media, blogs, and opinion pieces involves metrics such as creativity, sentiment, and engagement. Social media values emotional impact and quick engagement, while blogs focus on readability and relevance [13][14]. Meeting notes require clarity and accuracy; opinion pieces balance sentiment with engagement. Real-time generation, especially for social media, demands low latency and memory efficiency. For longer content [15], ROUGE-1 and ROUGE-L measure summarization quality. Mistral Nemo scored the highest ROUGE-1 F-measure at 0.40685, indicating strong alignment with reference texts (see Table 3, Fig. 3).

Table II. Content Generation Priorities Across Various Content Types

Metric	Social Media Posts	Blogs	Meeting Notes	Opinion Pieces
Creativity & Novelty	High priority	Medium priority	Low priority	Medium priority
Sentiment/Emotion	High priority	Medium priority	Low priority	High priority
Engagement Simulation	High priority	N/A	N/A	N/A
ROUGE (Relevance)	Low priority	High priority	High priority	Medium priority
Factual Consistency	Low priority	High priority	Medium priority	High priority
Readability	Low priority	High priority	Medium priority	Medium priority
Latency & Memory Footprint	High priority	Medium priority	Medium priority	Medium priority

Table III. Performance of Language Models on 'ROUGE' Metrics

Model	ROUGE-1 Precision	ROUGE-1 Recall	ROUGE-1 F-measure	ROUGE-L Precision	ROUGE-L Recall	ROUGE-L F-measure
Llama 3.1 8B	0.75543	0.21576	0.37468	0.55312	0.07732	0.13998
Gemma 2 9B	0.74522	0.19988	0.36301	0.50894	0.06854	0.12804
Mistral Nemo	0.79258	0.24912	0.40685	0.61320	0.09642	0.17530
Mixtral 8x7B	0.73912	0.19576	0.35218	0.49051	0.06510	0.12078

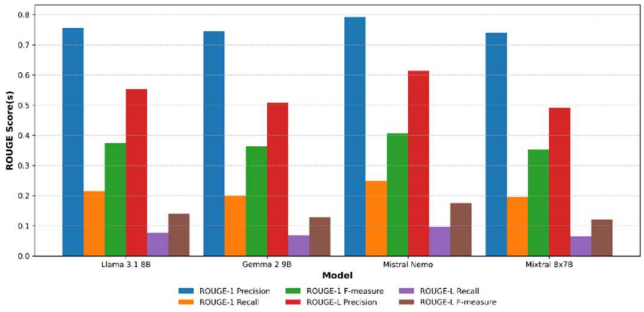


Figure 3. 'ROUGE' Score Comparison of Language Models: Precision, Recall, and F-Measure for 'ROUGE-1' and 'ROUGE-L'.

B. Engagement, Readability, and Multilingual Content Metrics

- Engagement Metrics: Social media engagement is driven by likes, shares, and comments. Posts with positive sentiment and higher compound scores see more interaction [16]. Sentiment analysis reveals emotional impact beyond basic metrics (see Table 4, Fig. 4). Mistral Nemo scores highest (0.82310), indicating strong emotional appeal, while LLaMA 3.1 (8B) scores lowest (0.68231). This underscores the role of sentiment in boosting engagement.
- Readability Metrics: Readability ensures content is accessible to diverse audiences. Metrics like Flesch Reading Ease and Flesch-Kincaid Grade Level assess complexity and appropriate reading level. Scores between 90–100 indicate easy readability, while lower scores imply greater complexity [17]. Model comparison (see Table 5, Table 6) shows Mistral Nemo (12B) with the highest readability (Flesch: 36.66, Grade Level: 12.76). Gemma 2 (9B) scores lowest (Flesch: 18.80), indicating more complex output.
- Factual Consistency and Linguistic Richness: Factual accuracy is measured using cosine similarity between the embeddings of generated and reference texts. This metric ensures that the generated text remains factually consistent, with scores closer to 1 indicating higher levels of accuracy. The performance of various models in terms of factual consistency is displayed in Table 7 and Figure 4. Scores that surpass a certain threshold, such as 0.7, suggest greater factual consistency in the generated content [18].

The formula for cosine similarity is as follows:

$$Cosine\ Similarity\ (A,B) = \frac{A\ .\ B}{||A||.||B||} \qquad (1)$$

- Linguistic Richness: Linguistic richness, measured by Type-Token Ratio (TTR), compares unique words to total word count. Mistral Nemo has the highest TTR (0.56987), followed by Gemma 2 (0.55112), indicating more diverse vocabulary (see Table 8). Mixtral, despite generating the most words, has the lowest TTR, showing less lexical variety. This relationship is illustrated in Fig. 7 [19].

TABLE IV. SENTIMENT ANALYSIS OF LANGUAGE MODELS BASED ON NEGATIVE, NEUTRAL, POSITIVE SENTIMENT SCORES AND COMPOUND SENTIMENT VALUES

Model	Negative	Neutral	Positive	Compound
Mistral Nemo	0.05443	0.75734	0.18823	0.82310
Mixtral 8x7B	0.08546	0.69425	0.22029	0.73650
Llama 3.1 8B	0.11687	0.69422	0.18891	0.68231
Gemma 2 9B	0.09136	0.68643	0.22221	0.72652

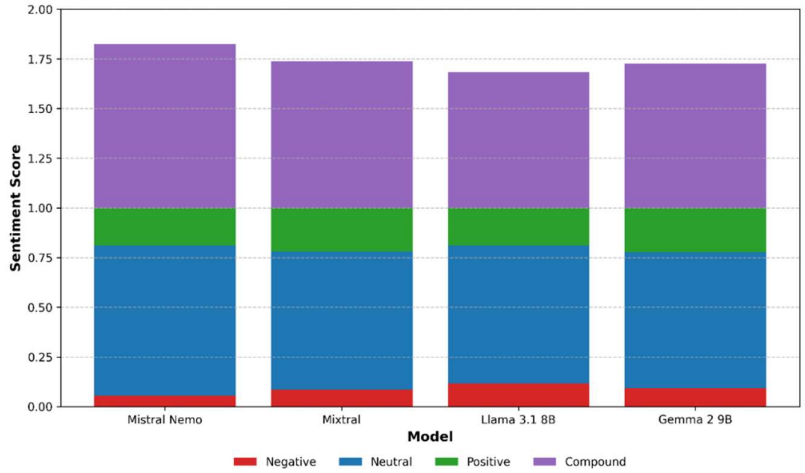


Figure 4. Sentiment Analysis Scores for Language Models – Graphical Representation

TABLE V. FLESCH-KINCAID READABILITY LEVELS

Flesch-Kincaid Score	Reading Level	School Level	Age Range	Example Book
0 - 3	Basic	Kindergarten / Elementary	5 - 8 years	Hooray for Fish!
3 - 6	Basic	Elementary	8 - 11 years	The Gruffalo
6 - 9	Average	Middle School	11 - 14 years	Harry Potter
9 - 12	Average	High School	14 - 17 years	Jurassic Park
12 - 15	Advanced	College	17 - 20 years	A Brief History of Time
15 - 18	Advanced	Post-graduate / Academic	20+ years	Academic papers

TABLE VI. READABILITY METRICS COMPARISON FOR LANGUAGE MODELS

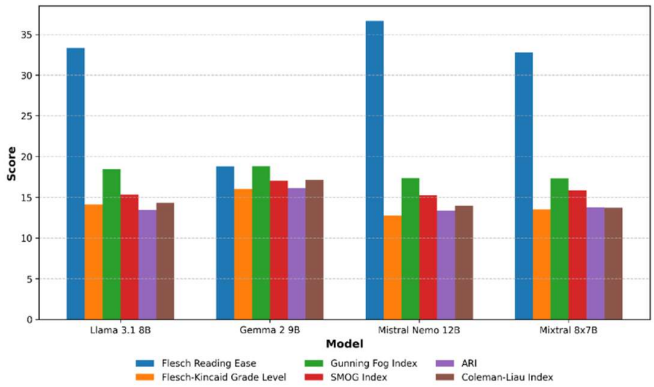
Model	Flesch Reading Ease	Flesch-Kincaid Grade Level	Gunning Fog Index	SMOG Index	ARI	Coleman-Liau Index
Llama 3.1 8B	33.36	14.11	18.45	15.33	13.45	14.33
Gemma 2 9B	18.80	16.03	18.83	17.02	16.14	17.11
Mistral Nemo 12B	36.66	12.76	17.36	15.23	13.38	13.98
Mixtral 8x7B	32.79	13.48	17.29	15.85	13.77	13.73

C. Computational Efficiency and Scalability

For large-scale content generation, computational efficiency and scalability are vital. These factors determine a model's practicality for real-world, high-volume use, especially in organizational settings.

Computational Efficiency Metrics: Computational efficiency measures resource use during content generation. Key metrics include inference time, memory usage, and latency. Inference time reflects how long a model takes

to generate content, while memory usage shows the resources used during inference. Latency measures the delay between input and output. Mistral Nemo (12B) strikes a balance between efficiency and quality, while Mixtral (8x7B) is optimized for low latency, ideal for high-throughput tasks like automated tweets or real-time updates.



Readability Metrics Comparison for Language Models – Plots.

TABLE VII. AVERAGE COSINE SIMILARITY FOR LANGUAGE MODELS

Model	Average Cosine Similarity
Llama 3.1 (8B)	0.6724
Gemma 2 (9B)	0.6469
Mistral Nemo (12B)	0.7745
Mixtral (8x7B)	0.6610

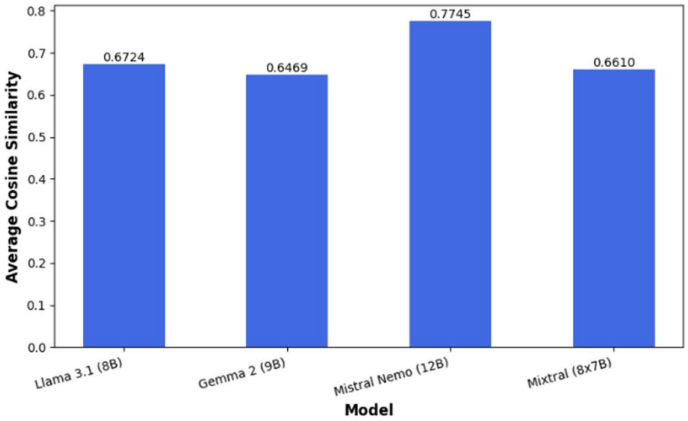


Figure 6. Average Cosine Similarity of Models – Plots.

Scalability Metrics: Scalability measures a model's performance with increased workload or data volume, essential for organizations generating high volumes of content.
Metrics include throughput (posts per second), model parallelization (distributing computation), and content quality retention (how well quality is maintained under heavy load).
Mixtral (8x7B) excels in throughput, while Mistral Nemo (12B) better retains content quality during scaling, making it suitable for diverse applications.

TABLE VIII. AVERAGE WORD COUNT AND TTR SCORE FOR LANGUAGE MODELS

Model	Average Word Count	Average TTR Score
Llama 3.1 8B	482.2	0.47019
Gemma 2 9B	412.2	0.55112
Mistral Nemo 12B	425.6	0.56987
Mixtral 8x7B	491.8	0.44613

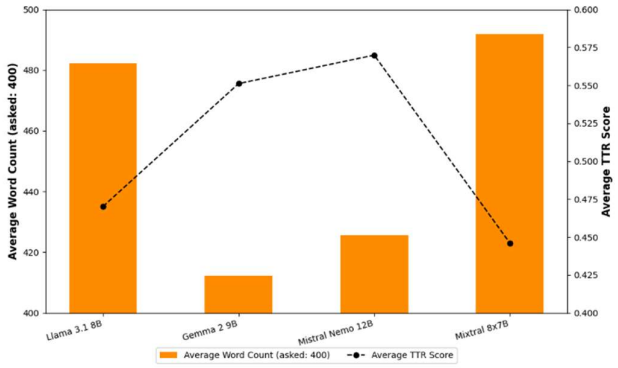


Figure 7. Linguistic Richness: Type-Token Ratio (TTR) Scores and Word Count of Selected Language Models.

VII. Conclusion

This study evaluates Large Language Models (LLMs) for automated social media content generation, comparing GPT-3.5, LLaMA 3.1 (8B), Mixtral 8x7B, Mistral Nemo 12B, and Gemma 2 (9B). While each model has platform-specific strengths—such as Mistral Nemo 12B's multilingual capabilities and LLaMA 3.1's effectiveness in concise posting—Mixtral 8x7B stands out for its consistent performance in content diversity, coherence, and computational efficiency. Fine-tuning improves model alignment with platform objectives, while prompt engineering offers a flexible, cost-effective alternative for customization. Based on these findings, Mixtral 8x7B will be adopted for future development of a real-time AI content generation system focused on adaptability, ethical moderation, and multilingual reach.

References

[1] S. Pokhrel and S. R. Banjade, AI Content Generation Technology based on Open AI Language Model, Journal of Artificial Intelligence and Capsule Networks, vol. 5, no. 4, pp. 534-548, 2023. Available: http://dx.doi.org/10.36548/jaicn.2023.4.006.

[2] C. Jensen and A. Højmark, Formalizing content creation and evaluation methods for AI-generated social media content, in Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), Singapore, Dec. 2023, pp. 22–41. Association for Computational Linguistics. Available: https://aclanthology.org/2023.gem-1.3.

[3] OpenAI, J. Achiam, S. Adler, S. Agarwal, et al., GPT-4 Technical Report, arXiv. Available: https://doi.org/10.48550/arXiv.2303.08774.

[4] A. Dubey, A. Jauhri, A. Pandey, and A. Kadian, The Llama 3 Herd of Models, arXiv. Available: https://arxiv.org.

[5] S. Sands, C. L. Campbell, K. Plangger, and C. Ferraro, Unreal influence: leveraging AI in influencer marketing, European Journal of Marketing, vol. 56, pp. 1721–1747. Available: https://doi.org/10.1108/EJM-12-2019-0949.

[6] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, et al., Mixtral of Experts, arXiv. Available: https://doi.org/10.48550/arXiv.2401.04088.

[7] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, A Systematic Survey of Prompt Engineering in Large Language Models, arXiv preprint arXiv:2406.09184.

[8] D. Grabb, The Impact of Prompt Engineering in Large Language Model Performance: A Psychiatric Example, arXiv preprint arXiv:2307.10973.

[9] D. T. Chinh Nguyet, Adoption of Generative AI in content creation: A case study from the advertising industry, 2024 IEEE Conference on Artificial Intelligence, Singapore. Available: https://doi.org/10.1109/CAI59869.2024.00029.

[10] Z. Zhou et al., A Survey on Efficient Inference for Large Language Models, ArXiv, abs/2404.14294. Available: https://arxiv.org/abs/2404.14294.

[11] G. Getto and J. T. Labriola, How to Create Content Models That Invite User Participation, IEEE Transactions on Professional Communication, vol. 62, pp. 385–397. Available: https://doi.org/10.1109/TPC.2019.2946996.

[12] M. Wu, A. Waheed, C. Zhang, M. Abdul-Mageed, and A. Aji, LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions, ArXiv, abs/2304.14402. Available: https://arxiv.org/abs/2304.14402.

[13] S. Li, X. Ning, L. Wang, T. Liu, et al., Evaluating Quantized Large Language Models, ArXiv, abs/2402.18158. Available: https://arxiv.org/abs/2402.18158.

[14] Y. Tang, F. Liu, Y. Ni, Y. Tian, et al., Rethinking Optimization and Architecture for Tiny Language Models, ArXiv, abs/2402.02791. Available: https://arxiv.org/abs/2402.02791.

[15] S. R. Chowdhury and K. Sarkar, Abstractive Multi-Document Summarization Using Sentence Fusion, 2023 International Conference on Information Technology, Amman, Jordan. Available: https://doi.org/10.1109/ICIT58056.2023.10225941.
[16] A. Kaushik, S. H. Attri, and R. S. Jha, Exploring Text Summarization Techniques: A Review of Current Challenges and Future Directions, 2024 2nd International Conference on Disruptive Technologies, Greater Noida, India. Available: https://doi.org/10.1109/ICDT61202.2024.10489243.
[17] K. S., S. R., S. R., and T. S. V., Survey on Automatic Text Summarization using NLP and Deep Learning, 2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems, Bangalore, India. Available: https://doi.org/10.1109/ICAECIS58353.2023.10170660.
[18] P. Batra, S. Chaudhary, K. Bhatt, S. Varshney, and S. Verma, A Review: Abstractive Text Summarization Techniques using NLP, 2020 International Conference on Advances in Computing, Communication & Materials, Dehradun, India. Available: https://doi.org/10.1109/ICACCM50413.2020.9213079.
[19] B. Dalwadi, N. Patel, and S. Suthar, A Review Paper on Text Summarization for Indian Languages, IJSRD - International Journal for Scientific Research & Development, vol. 5.
[20] A. K. Chitturi and S. Kandaswamy, Survey on Abstractive Text Summarization using various approaches, International Journal of Advanced Trends in Computer Science and Engineering, vol. 8.