Impact of LLM on Urdu Text Summarization

Ayaz Mehmood
School of Electrical Engineering and Computer Science (SEECS) National University of Science and Technology (NUST),
Islamabad, Pakistan
ayaz.msds23seecs@seecs.edu.pk

Zainab Bibi
School of Electrical Engineering and Computer Science (SEECS) National University of Science and Technology (NUST),
Islamabad, Pakistan
zbibi.msds23seecs@seecs.edu.pk

Dr. Seemab Latif
School of Electrical Engineering and Computer Science (SEECS) National University of Science and Technology (NUST),
Islamabad, Pakistan
seemab.latif@seecs.edu.pk

***Abstract*—To evaluate the effectiveness of large language models (LLMs) on low-resource languages, this project utilizes quantized versions of LLaMA2, LLaMA3, and Mistral7B to study the effects of large language models on Urdu text summarization. This project uses the Parameter-Efficient Fine-Tuning (PEFT) technique to fine-tune these models on a variety of online sourced Urdu abstraction summary data. To determine the quality of the summaries, the models are evaluated using ROUGE and BLUE scores. These results demonstrate how well-quantized LLMs and PEFT handle Urdu summarization.**

I. Introduction

A shortage of complete datasets and the low efficiency of current summarizing techniques make Urdu text summarization extremely challenging. By using advanced language models more particularly, quantized versions of LLaMA2, LLaMA3, and Mistral7B, this research aims to overcome these issues. We use the Parameter-Efficient Fine-Tuning (PEFT) approach with Low-Rank Adaptation (LoRA) to improve these models for Urdu text summarization, taking into account the limitations of computing resources. Our process includes using the dataset of text passages in Urdu and the abstractive summaries that go along with it, quantizing language models, creating efficient prompt forms, and integrating PEFT with LoRA.

II. Problem Statement

The summarization of Urdu, a low-resource language, is currently inadequate due to the limited availability of annotated data and the linguistic complexities. So, there is a significant need for more robust models capable of generating high-quality abstract summaries in Urdu. This study aims to evaluate the effectiveness of large language models such as LLaMA2, LLaMA3, and Mistral7B in enhancing the quality of summarization for Urdu texts.

III. Literature Review

The emergence of large language models has significantly impacted the field of natural language processing specifically in the domain of translation, summarization and text generation. LLM has shown promising results in summarization of resource rich language. However, the application of LLMs in resource-scarce languages like Urdu remains relatively unexplored within the academic literature. Urdu, characterized by its linguistic complexity and limited digital resources, presents unique challenges for LLM-driven summarization techniques. The challenges and creative solutions in converting pre-trained LLM to low-resource languages were examined in this research. The authors address how a lack of training data and enormous computational requirements make it difficult to train models for low-resource languages from the start. The tokenizer has poor tokenization efficiency if it was not trained in a specific language, and the author also considers the devastating forgetting of LLM when adapting to a new language. These issues are handled in the context of low-resource languages. To reduce the impact of catastrophic forgetting, a data-combining strategy was implemented throughout the two distinct pre-training and fine-tuning stages. This approach ensures that the adapted models retain their proficiency in the original language while demonstrating improved performance in the new language. [1] This paper demonstrated the informative comparisons of the given large language models, particularly MPT-7b-instruct, Falcon-7b-instruct, and OpenAI ChatGPT text-davinci-003, and various interpreting summarization datasets namely CNN/Daily Mail 3.0.0 and XSum. Additionally, the paper features the ideas and tactics that are applied by means of extractive and abstractive text summing-up techniques. We have also deliberated the supervised and unsupervised text summing-up which are illustrated with manifestations of the pros and cons and of the area where both of them are used. When these LLMs were judged in terms of multiple performance measures, the text summarization task gave the highest-performing results on the Davinci text model and had a Rough-1 model score of 0.272. [2] This study mainly surveys the latest advances in neural networks. Abstracted text summarization predicts the context of their key elements, challenges, and potential solutions. We further suggest integrating neural network architectures with pre-trained language models to enhance summarization outcomes. The analysis of various encoder-decoder architectures, such as recurrent neural networks (RNNs), transformers, long short-term memory (LSTM), gated recurrent units (GRU), bidirectional RNNs/LSTMs/GRUs, and pre-trained language models like BERT and GPT, is one of the key aspects of the paper. The suitability of every structure for sequence-to-sequence learning tasks and its ability to efficiently capture contextual

information are evaluated. Furthermore, the study considers a number of operations, such as pointer-generator networks, different attention mechanisms, and coverage mechanisms. Each approach handles different problems like identifying key concepts, regulating terms that are not frequently used, reducing duplication, and improving the accuracy of the facts in the summaries that are generated. Furthermore, the study assesses the performance of multiple models with basic evaluation metrics like as BLEU and ROUGE scores. [3] Having a particular focus on challenges that low-resource languages like Urdu have in natural language processing (NLP), the paper examines the use of Language Model Models (LLMs) for Urdu summarization. With the focus on updating pre-trained models for low-resource languages such as Urdu, utilizing a smaller version of the mT5 model, the urT5 model performs significantly in obtaining contextual data in Urdu which was derived from mT5. This study showed more importance beyond Urdu and also builds larger datasets for the future to improve low-resource summarizing calls for building stronger multilingual models. [4] The main motive of this paper is to compare the performance of four large language models (LLMs) PEGASUS, Prophet-Net, BART, and T5 for text summarization. These models were assessed using different evaluation metrics on four datasets (CNN/DailyMail, XLSum, WikiHow, and DUC2004). The aim of this study by using four different models is to check which model gives the best concise summarization. The T5 model gives the best performance out of all other models and Prophet-Net performs poorly on all datasets. [5] This paper's primary goal is to address a lack of research, despite the huge amount of digital data available, on abstractive text summarization in the Urdu language. The study evaluated several deep learning and large language models (LSTM, Bi-LSTM, LSTM with attention, GRU, Bi-GRU, and GRU with attention, BART, and GPT-3.5) using a benchmark corpus of 2,067,784 Urdu news articles. The GRU with attention model outperforms the other models with ROUGE-1 = 46.7, ROUGE-2 = 24.1, and ROUGE-L = 48.7, according to the evaluation conducted on 20,000 test instances. [6] This study aims to examine and review the algorithms and methods used for the summarization of Urdu text and to select the optimal technique focused on the extractive abstract generation for summarizing the text of Urdu news. We evaluate different algorithms and methods for text summarization with a focus on the quality of the solution based on the set of headlines. So, a comparison of the following methods of summary extraction for Urdu text shows that Luhn's algorithm for text summarization demonstrated the highest result with preprocessed data. [7] This study explores the capabilities of abstractive summarization techniques for the Urdu language, to be precise, the strictly correlated transformer encoder-decoder architecture approach for the abstractive summarization of the Urdu language. The usage of the transformer-based model has enabled generating summaries that were not only semantically aligned but also grammatically correct with self-attention mechanisms. The model performs quite well on the ROUGE measure scoring an average F1 of 0.43 for ROUGE-1, 0.25 for ROUGE-2, and 0.23 for ROUGE-L. This work is a significant contribution to the development of abstractive summarization methodologies for low-resource languages, including the focus on Urdu. [8] The primary goals of this study are, to construct an automatic system to summarize the texts in Urdu, integrate extractive and abstractive algorithms, word frequency and sentence weight methods, TF-I D F, use the B E R T model's ability to generate better summaries and seek help from Urdu language specialists. This paper is similar to the qualitative one that employs diverse kinds of Urdu text from several areas. Thus, the results indicate that the hybrid summarizing approach is more efficient in comparison to the traditional extractive methods by generating summaries of better quality concerning the content integrity across documents and enhancing coherence and readability. Finally, their evaluation by experts ensures the accuracy and quality of the produced summaries while providing valuable information on the system's performance and potential areas of enhancement. [9] This study shows that there has been a significant gap highlighted in the domain of natural language processing for the Urdu language, specifically in text summarization. As noted, the abstractive summary for Urdu has not been researched primarily, while a considerable degree of work has been done on extractive approaches for summarization. In short, this study reveals a new route to enhance the contextual understanding of Urdu text by deploying Roberta embedding. Unlike traditional one-hot encodings that provide a less definitive representation of the textual data, this approach provides better summarization results. The study's dataset, which consists of more than 19,000 annotated samples, significantly strengthens the model's generalization abilities. Therefore, the study forms a basis for potentially enhanced techniques in Urdu language processing due to the introduced labeled dataset, proposed innovative methods, and query measurements. [10]

IV. Methodology

The objective of this project is to evaluate how well LLMs can summarize texts written in Urdu. We focused on using LLaMA2, Mistral7B, and LLaMA3, using these advanced techniques like Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (Lora) and 4-bit quantization.

A. Large Language Model

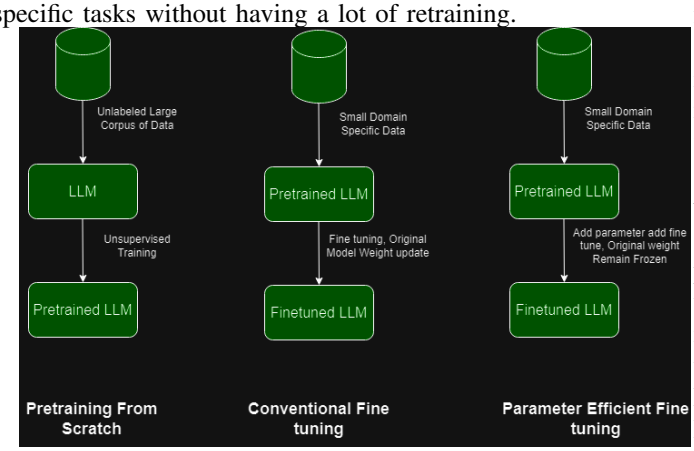
LLM is trained on large text datasets to comprehend and produce language similar to that of humans. LLM replace the Translation, summarizers, and question answerers that are capable of capturing complex language-based patterns by the use of transformer architectures and deep learning. Due to a large number of parameters, these models become strong tools for NLP, which allows them to produce meaningful and appropriate results.

1) Pre-training From Scratch: Unsupervised learning techniques are used to train an LLM on a huge amount of unlabeled data in training from scratch. Via this procedure, the model can choose language structures and patterns even in the absence of

labeled data, making a pre-trained LLM with a wide linguistic knowledge. It is computationally expensive.

2) Conventional Fine-Tuning: Fine-tuning of LLM involves training the pre-trained model again on small and specific data to modify it for particular tasks. The initial weights are modified during this process to improve the model's performance on specific tasks. It is computationally expensive because all model parameters must need to be updated.

3) Parameter Efficient Fine-Tuning: The aim of Parameter-efficient fine-tuning is to modify only a small portion of a pre-trained LLM's parameters or add new ones while preserving most of the initial weights constant, to match it to unknown tasks. PEFT especially reduces the computation power and memory needs. The fine-tuning process is optimized by providing the model to achieve excellent performance on domain-specific tasks without having a lot of retraining.



B. Dataset

1) Data collection: The Urdu-based Abstractive text summarization dataset is gathered from publicly available open-source datasets, which are essential for training and evaluating our models for producing abstractive summaries of Urdu texts. The use of publicly available data guarantees openness reproducibility, and accessibility of the research, promoting collaboration and future developments in Urdu natural language processing

C. Prompt Format

To fine-tune LLM to achieve the best results and deliver accurate responses catered to certain activities or objectives, they must use the appropriate prompt structure. Every large language model has unique prompt forms that must be thoughtfully planned and constructed for the model to produce excellent results.

D. Model Selection

Based on the efficiency and feasibility, we chose the following key language models for our text summarizing task:

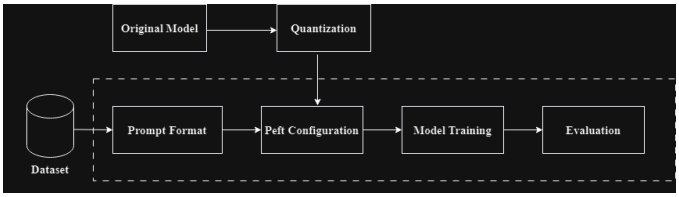
1) LLaMA2: Due to LLaMA2's outstanding ability to handle difficult language issues and its comprehensive text generation and comprehension capabilities, it was chosen for use. This model provides organized, relevant to-context summaries, which makes it a great choice for abstractive text summaries. It can understand and provide excellent summaries in Urdu because of its thorough pre-training on a variety of datasets, which is crucial for our project.

2) LLaMA3: A higher number of features and improved performance are offered by the Llama 3 model, the most modern in the Llama series. Even more accurate and fluid summaries could be produced by LLaMA3 with the use of more parameters and sophisticated training techniques. To push the boundaries of text summarizing and guarantee exceptional performance for our Urdu summarization requirements, this model makes use of the most recent advancements in language modeling.

3) Mistral7B: Mistral7B is famous for its efficiency and balanced performance because it offers a balance between processing requirements and output quality. When there are few resources available but a requirement for high-quality summaries, this approach works especially well. Because of its architecture, which puts efficiency optimization first while maintaining the standard of the output summaries, it is a good fit for our objective.

E. Interpretation of the Model Training Process

Quantizing the original model minimizes its size as well as processing specifications, and this is the first step in the model training process. After quantization, the dataset is converted into an appropriate prompt format. The structured data and the quantization model are sent to the Parameter-Efficient Fine-Tuning (PEFT) setup step. Following the setup of the PEFT configuration, the model is trained by applying the provided dataset to teach it new concepts. In the last stage, the effectiveness and performance of the trained model in the given task is evaluated. This process ensures the best possible performance while making effective use of the available resources.



F. PEFT with Lora

PEFT with Lora is a technique that increases the training of LLM by decreasing the number of parameters that need to be adjusted and, thus, lowering computational requirements. This method enhances training efficiency and allows enormous models to be fine-tuned even with limited computer resources. Important steps are involved in this procedure,

1) Base Model Initialization: The base model which we called a pre-trained language model it must be loaded with the pre-trained weights. Fine-tuning is a better way to train on domain-specific data as the model is already trained on a vast amount of data and has a general knowledge of language which can be more easily applied to particular tasks than if it were trained from scratch.

2) Low-Rank Adaptation: It is a technique used to fine-tune LLM in a much more efficient way instead of the traditional fine-tuning method Low-Rank Adaptation (Lora) is used to concentrate on a smaller number of parameters instead of fine-tuning the original pre-trained model weight. This is achieved by Lora by using low-rank matrices to estimate the weight matrices of the model, resulting in a major decrease in the number of parameters that must be changed during training. This speeds up the process of training and it requires less computational resource

3) 4-bit Quantization: In the PEFT with the Lora procedure, 4-bit quantization is the last step that lowers the model's weights to 4-bit precision. This lowers memory use and computational costs even more, allowing the model to run on GPUs with constrained resources without noticeably sacrificing performance. Notwithstanding their reduced precision, models quantized to 4 bits are able to generate good results.

G. Evaluation Metrics

For the summarization task, ROUGH and BLEU Score and the commonly used matrix that is used to evaluate the performance of the Summarization model. ROUGE measures how well the generated summaries convey the important details, whereas BLEU focuses on the precision of word selections and sentence structures.

1) Rough Score: This score measures how much the summary matches the reference summaries. It looks at word pairs, word sequences, and n-grams to do this. A higher score means the summary holds onto more important info and context from the original text. This shows the summary keeps more of the main structure and content from the reference summary.

2) Bleu Score: The BLEU score is a machine translation statistic that may also be used for summary. It evaluates how accurate the generated text is in relation to reference summaries. It focuses on getting the right words and order in the generated summary to match the content and syntax of the references.

V. Results

The models' initial performance was quite low and required computational resources. Initially, we used only half of the data and fine-tuned a limited number of parameters due to computational resources. By increasing the number of fine-tuned parameters using LORA parameters r and alpha, increasing the training epochs, and utilizing Colab Pro+ for GPU resources, we observed significant improvements in results. Furthermore, using the full dataset further enhanced the outcomes.

Model	ROUGE Score	BLEU Score
LLaMA3	12.43	15.27
LLaMA2	5.32	7.41
Mistral7B	1.63	4.17

A. LLaMA3

None of the models give good results but LLaMA3's predictions were a bit better than LLaMA2's and Mistral7B's. The higher ROUGE and BLEU scores attained by LLaMA3 show this improvement.

Predicted Summary	Summary
پاکستان میں محدود سرمایہ کاری مواقع اور کم منافع کی وجہ سے غیر ملکی کرنسی کی غیر قانونی منتقلی خلیجی ریاستوں کو ہو رہی ہے، جس سے روپے کی قیمت میں کمی اور تجارتی خسارہ بڑھ رہا ہے۔	پاکستان کے زرمبادلہ کے بارہ ارب کے ذخائر کئی سالوں بعد پہلی دفعہ دباؤ کا شکار ہیں۔ پیسہ تیزی کے ساتھ ملک سے باہر جا رہا ہے جس کی وجہ سے روپے کی قیمت میں کمی جبکہ افراط زر کی شرح میں مسلسل اضافہ ہو رہا ہے۔ کیپیٹل فلائیٹ کا سب سے بڑا سبب خلیجی ممالک میں سرمایہ کاری ہے۔ متعدد کمپنیاں جائیداد میں سرمایہ کاری کی سکیموں کے ذریعے خلیجی ممالک میں پیسہ کھینچ رہی ہیں

B. LLaMA2

When it came to predictions, LLaMA2 performed a better job than Mistral7B, but not as well as LLaMA3. LLaMA2 is still behind LLaMA3 in overall accuracy even if it shows a slight improvement in coherence.

Predicted Summary	Summary
رفاہ میں جھڑپوں سے 12 فلسطینی اور 5 اسرائیلی فوجی ہلاک ہوئے، بفر زون کی توسیع کے لیے گھروں کی مسماری جاری ہے	اسرائیلی ریڈیو کے مطابق فوج کو غزہ کی پٹی میں رفاہ میں واقع سینکڑوں گھروں کو مسمار کرنے کے اختیارات دے دیئے گئے ہیں

C. Mistral7B

Mistral7B's predictions were highly inaccurate and showed a significant drop in generating reliable summaries as compared to LLaMA2 and LLaMA3.

Predicted Summary	Summary
پاکستان میں کرنسی کی سمگلنگ اور خلیجی ریاستوں میں جائیداد خریدنے کے باعث معیشت تباہ ہو رہی ہے۔	پاکستان کے زرمبادلہ کے بارہ ارب کے ذخائر کئی سالوں بعد پہلی دفعہ دباؤ کا شکار ہیں۔ پیسہ تیزی کے ساتھ ملک سے باہر جا رہا ہے جس کی وجہ سے روپے کی قیمت میں کمی جبکہ افراط زر کی شرح میں مسلسل اضافہ ہو رہا ہے۔ کیپیٹل فلائیٹ کا سب سے بڑا سبب خلیجی ممالک میں سرمایہ کاری ہے۔ متعدد کمپنیاں جائیداد میں سرمایہ کاری کی سکیموں کے ذریعے خلیجی ممالک میں پیسہ کھینچ رہی ہیں

VI. Discussion

The zero-shot performance of the models was zero and even after training the model on complete data, the results were not very good. LLaMA3 achieved quite good results than LLaMA2 and Mistral7B, and the evaluation matrix ROUGE and BLEU ratings of 12.43 and 15.27, respectively. However, by utilizing PEFT with LoRA, we were able to optimize our resource utilization and attain some level of improvement via hyperparameter tuning. This suggests that significant enhancements in Urdu text summarizing might be made with more thorough training, an increase in data, and improved resources.

VII. Future Work

Given the current limitation and low performance, there is a need to explore additional large language models and fine-tuning techniques to identify more effective approaches to improve the performance of LLM on Urdu summarization. To improve the quality and consistency of the generated summaries, it will also be essential to add more extensive and diverse datasets for Urdu text. Finally, further research into novel structures and language modeling methodologies could provide more effective solutions for Urdu text summarizing tasks.

VIII. Conclusion

This study evaluated the use of advanced language models for Urdu text summarization, including LLaMA2, LLaMA3, and Mistral7B. We optimized these models using Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (LoRA) due to computational limitations. While none of the models outperformed the others, LLaMA3 performed somewhat better than LLaMA2 and Mistral7B. Mistral7B's performance was noticeably poor. Our results demonstrate the promise of these models for Urdu summarization, but they also point to the need for more research and computing resources to increase coherence and accuracy.

IX. Details of Project

You can check our complete project here,
https://github.com/zainab-10/Impact-of-LLM-on-Urdu-Text-Summarization

References

[1] U. T. Q. X. Zoltan Csaki, Pian Pawakapan, "Efficiently adapting pretrained language models to new languages," *NLP*, vol. 2, p. 13, 2023.

[2] M. S. Lochan Basyal, "Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models," *NLP*, vol. 2, p. 4, 2023.

[3] A. A. Syed, "A survey of the state-of-the-art models in neural abstractive text summarization," vol. 9, pp. 13 248–13 265, 2021.

[4] N. I. K. M. Mubashir Munaf, Dr Hammad Afzal, "Low resource summarization using pre-trained language models," *NLP*, vol. 1, 2023.

[5] V. J. D. V. A. N. M. F. S. K. J. T. K. K. Bharathi Mohan G, Prasanna Kumar R, "Comparative evaluation of large language models for abstractive summarization," pp. 59–64, 2024.

[6] R. M. A. N. MUHAMMAD AWAIS1, "Abstractive text summarization for the urdu language: Data and methods," vol. 10, 2017.

[7] A. Farooq, S. Batool, and Z. Noreen, "Comparing different techniques of urdu text summarization," pp. 1–6, 2021.

[8] U. M. Ali Raza, Hadia Sultan Raja, "Abstractive summary generation for the urdu language," vol. 1, 2023.

[9] M. Asif, S. A. Raza, J. Iqbal, N. Perwaiz, T. Faiz, and S. Khan, "Bidirectional encoder approach for abstractive text summarization of urdu language," pp. 1–8, 2022.

[10] W. S. HASSAN RAZA, "End to end urdu abstractive text summarization with dataset and improvement in evaluation metric," vol. 10, 2024.