IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Quality Assurance Framework For Evaluation Of Text Generation Models

[1] Pushpanathan G, [2] Nithin N, [3] Santhosh Adavala, [4] Shafath H Khan,[5] Syed Mohammed Maaz

[1] Assistant Professor, [2] Student, [3]Student, [4] Student, [5] Student

Department of Computer Science and Engineering,

Cambridge Institute of Technology, Bangalore, India

Abstract: This paper presents a comprehensive Quality Assurance Framework designed specifically for text generation models. Our approach combines automated metrics such as BLEU, ROUGE, and perplexity scores with novel techniques for coherence and factuality assessment. We integrate human evaluation methodologies to ensure a balanced assessment of linguistic quality, coherence, factual accuracy, and diversity in generated texts. Through extensive experimentation across different text generation tasks, our framework demonstrates improved evaluation accuracy and provides valuable insights for model refinement and optimization, contributing to the advancement of trustworthy text generation models.

I. INTRODUCTION

Text generation, a pivotal aspect of modern natural language processing, plays a crucial role in diverse applications and projects. The process involves developing models that can understand and generate human-like text, enabling automation of content creation, translation, summarization, and even dialogue systems. In the context of our project report, text generation serves as a powerful tool for producing coherent and contextually relevant content, enhancing the efficiency of data interpretation and communication. By leveraging advanced models, such as text-to-text generation, we aim to improve the quality of generated text, ensuring that it aligns with project objectives and meets industry standards. The implementation of text generation within our project report not only facilitates information dissemination but also showcases a commitment to staying at the forefront of technological advancements in natural language processing.

This paper introduces a robust Quality Assurance Framework tailored for text generation models. Our method merges automated metrics like BLEU, ROUGE, and perplexity scores with innovative approaches for coherence and factuality evaluation. Human assessment techniques are incorporated to achieve a comprehensive evaluation of linguistic quality, coherence, factual accuracy, and diversity in generated texts. Our framework, tested extensively across various text generation tasks, enhances evaluation precision and

offers valuable insights for refining and optimizing models, advancing the development of reliable text generation models.

II. LITERATURE SURVEY

There are multiple techniques employed for text generation models. Some approaches involve **[1]** This paper presents BART, a Transformer-based denoising autoencoder for sequence-to-sequence pretraining, demonstrating superior performance in various text tasks. It utilizes advanced noising strategies and a standard Transformer architecture for optimal outcomes. **[2]** This paper introduces BERTSCORE, a robust evaluation metric for text generation tasks that leverages contextual embeddings for token similarity computation. It outperforms existing metrics, showing higher correlation with human judgments and better model selection accuracy, especially in challenging contexts. **[3]** This paper presents an automated, cost-effective method for machine translation evaluation, offering a rapid and language-independent alternative to time-consuming human assessments. It correlates well with human judgments, making it suitable for quick or frequent evaluations at reduced costs. **[4]** This paper proposes prefix-tuning, an efficient alternative to fine-tuning for language models, optimizing a small task-specific vector while keeping the model parameters fixed. Results on GPT-2 and BART demonstrate comparable performance on full data, superior performance with limited data, and improved generalization to unseen topics. **[5]** This paper explores leveraging language models trained on WebText for zero-shot learning, achieving competitive performance across NLP tasks without task-specific data. It emphasizes the role of model capacity in facilitating zero-shot task transfer and highlights the potential of large, diverse datasets for training language models. **[6]** This paper introduces METEOR, an automatic metric for machine translation based on generalized unigram matching. It shows improved correlation with human judgments on Arabic-to-English and Chinese-to-English translations. **[7]** This paper introduces ROUGE, a metric for automatically evaluating summary quality by comparing it to ideal human-generated summaries. It details four ROUGE measures and their applications in large-scale summarization evaluations like DUC 2004. **[8]** This paper introduces Texygen, a benchmarking platform designed for evaluating open-domain text generation models. It includes various text generation models and metrics to assess text diversity, quality, and consistency, aiming to standardize and improve the reproducibility of text generation research. **[9]** This paper explores recent progress in text generation, showcasing its impact in diverse domains. Through a systematic review, it identifies research gaps and offers insights for future exploration and practitioner guidance. **[10]** This paper introduces the concept of stylized data-to-text generation to address style variation challenges. It presents StyleD2T, a model designed to generate coherent text from non-linguistic data while adhering to specific styles, showcasing its superiority through extensive experiments. **[11]** This paper discusses the rising popularity of online learning, particularly during the Covid-19 pandemic, and the challenges in assessment. It introduces a transfer learning-based AQG model using T5 and other techniques to automatically generate subjective and objective questions from text documents, improving context awareness and performance. **[12]** This paper delves into the utilization of pre-trained language models (PLMs) for text generation, addressing key aspects such as input encoding, PLM design,

and optimization for desired text properties. It offers insights into challenges, solutions, resources, and future research directions in PLM-based text generation.

III. METHODOLOGY

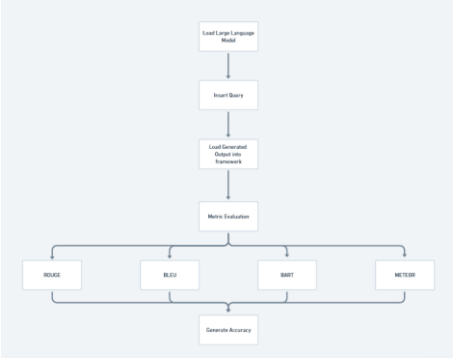

Figure 1: System architecture

The above figure shows how our project actually works. Initially, it loads the Large Language Model of our choice (ex. Mistral 7B, GPT etc.). Next, we submit a query to the LLM and wait for it to generate an answer to the requested query. The LLM generates the output and loads this generated output into the framework.

This framework is responsible for metric evaluation. The metrics included in this framework are ROUGE, BLEU, BART, METEOR and other standardized metrics. Each of these metrics are used to evaluate certain criteria such as translation, summarization, question-answering, logical reasoning etc. Accordingly, the framework generates accuracy of the model and based on the accuracy, changes can be done to the model to increase the accuracy significantly by finetuning.

The methodology is broken down into 5 main crucial steps which represent the flow of execution of the system and the relationship between the individual components. These are:

1. Data Acquisition.
2. Data Ingestion.
3. Aggregation and scoring.
4. Analysis.
5. Improvement

The different methods employed for quality assurance framework for text generation models are as follows

3.1 Data Acquisition

The initial phase entails Data Acquisition, which encompasses gathering information from diverse sources, including multiple users and their requirements for the large language model. Users submit queries or provide data, and the large language model processes this input to deliver the corresponding output. This foundational step is crucial for evaluating the large language model's accuracy and efficiency in generating responses.

The data acquisition phase in text generation models involves several key steps to gather and prepare the necessary data for training. Initially, data is collected from diverse sources such as books, articles, or websites to create a representative dataset. Following this, the raw text undergoes cleaning to remove noise, punctuation, and other unwanted elements, ensuring a consistent format suitable for training. Tokenization is then applied to break down the text into smaller units like words or subwords, facilitating processing by the model. Depending on requirements, data augmentation techniques may be employed to enhance dataset diversity and volume. Subsequently, the dataset is split into training, validation, and test sets, each serving specific purposes in model development and evaluation. Additional language-specific preprocessing steps such as stemming or lemmatization may be applied based on the language or domain of the text. Overall, the data acquisition phase is foundational in establishing a clean and relevant dataset that enables effective training of text generation models for producing high-quality outputs.

3.2 Data Ingestion

In The next stage involves Data Ingestion, during which user queries or data are incorporated into the large language model to produce the corresponding output. In this step, the large language model processes the input query, generating summaries, translations, or answers tailored to user requirements. The resulting output from the large language model is then utilized to assess the model's performance using standardized metrics within the evaluation framework.

The data ingestion phase in text generation models encompasses several critical steps to prepare and utilize data effectively for model training. Initially, data collection involves gathering a diverse and representative dataset relevant to the desired text generation task, which may involve web scraping or utilizing existing corpora. Following this, data cleaning and preprocessing are essential to ensure that the collected data is in a usable format by removing noise, correcting errors, and normalizing text. Tokenization and embedding then transform the text into numerical representations (word embeddings) suitable for model input. Additionally, feature extraction may be employed to capture relevant linguistic features like part-of-speech tags or named entities. Dataset splitting is performed to segregate the data into training, validation, and test sets for model training and evaluation purposes. Optionally, data augmentation techniques can be applied to increase dataset diversity and size. Ultimately, effective data formatting organizes the data into batches or sequences as required by the specific text generation model, optimizing input quality and contributing to the overall performance and robustness of the model.

3.3 Aggregation and Scoring

In the third phase, the output from the large language model, comprising the generated text, is fed into the framework. This generated text is then compared against the Reference text, which has been previously fed and trained using datasets. The evaluation of the large language model's accuracy and efficiency is determined by how closely the generated text aligns with the Reference text. The Reference text represents the real-time expectations of users from the large language model, while the generated text is the model's response to a

given query. The evaluation involves assessing the consistency of the generated text and assigning a score, with 0 indicating lower accuracy and 1 indicating higher accuracy.

3.4 Analysis

In the fourth stage, the large language model undergoes analysis based on the metrics or scores provided by the framework, offering a comprehensive assessment of its accuracy when generating text for a given query. This analysis serves to recapitulate the model's performance, identifying both its strengths and weaknesses. Insights gained from this phase are instrumental in refining and enhancing the model. By understanding the specific areas where the model excels or falls short, adjustments can be made, such as increasing the size of training datasets or addressing deficiencies in particular aspects of the model. Ultimately, the analysis phase is pivotal in guiding improvements to meet user expectations more effectively.

3.5 Improvement

In the fifth and concluding phase, the focus shifts to the improvement of the large language model, guided by user requirements and the accuracy score obtained from the framework. This phase is integral for the model to remain relevant and responsive to user needs in the market. The framework serves as a crucial tool in gauging the accuracy with which the large language model generates text according to user expectations. The enhancement process primarily involves extensive training of the model with large datasets, refining its ability to produce outputs that align closely with user preferences for various queries. This iterative improvement cycle occurs at significant time intervals, ensuring the model evolves to meet changing user demands effectively.

IV. Results and Discussion

The effectiveness of the Quality Assurance Framework for text generation models involved a range of metrics. The key results are as follows. The score for question-answering (Figure 2) was found to be 1.0, i.e. it is precise. The Summarization accuracy (Figure 3) was found to be 0.892. The obtained accuracy for translation (Figure 4) is 0.99. This analysis offers a perspective on the model used for the generation and its accuracy. The accuracy values indicate the model's ability to accurately generate the text according to the user’s query.

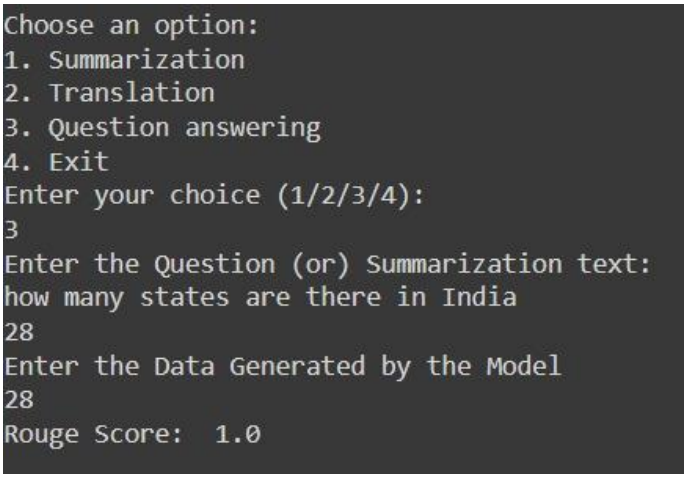

Figure 2: Question-answering Accuracy

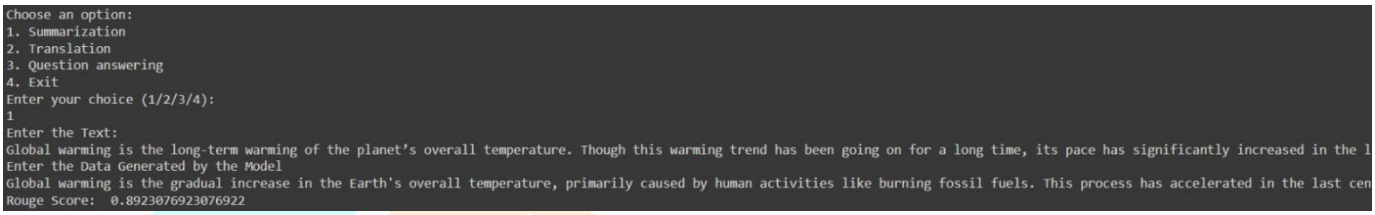

Figure 3: Summarization Accuracy

Figure 4: Translation Accuracy

V. CONCLUSION

In conclusion, our proposed Quality Assurance Framework provides a comprehensive and systematic approach to evaluating text generation models. By combining automated metrics like BLEU, ROUGE, and perplexity scores with innovative techniques for coherence and factuality assessment, we ensure a balanced evaluation of linguistic quality and factual accuracy in generated texts. The integration of human evaluation methodologies further enhances the robustness of our framework, capturing aspects like coherence and diversity that automated metrics alone may overlook. Through extensive experimentation across various text generation tasks, our framework demonstrates improved evaluation accuracy and offers valuable insights for model refinement and optimization. Moving forward, the framework holds promise for advancing the development of trustworthy and high-performing text generation models by providing a standardized methodology for evaluating their outputs. Future work could focus on expanding the framework to accommodate emerging challenges in text generation and adapting it to different domains or languages for broader applicability and impact.

VI. ACKNOWLEDGMENT

We extend our heartfelt gratitude to Mr. Pushpanathan G, Assistant Professor in the Dept of CSE at CITech, for his invaluable guidance and impressive technical insights that significantly contributed to the successful completion of our project. Additionally, we wish to convey our deep appreciation to our friends and teachers who provided assistance in various technical aspects, enriching our project with their expertise and feedback. Lastly, we are grateful for our parents for their unwavering support and encouragement throughout this journey, serving as a constant source of motivation.

REFERENCES

[1] BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer.

[2] BERTSCORE: EVALUATING TEXT GENERATION WITH BERT, Tianyi Zhang , Varsha Kishore , Felix Wu , Kilian Q. Weinberger , and Yoav Artzi.

[3] BLEU: a Method for Automatic Evaluation of Machine Translation, Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu.

[4] Prefix-Tuning: Optimizing Continuous Prompts for Generation Xiang Lisa Li, Percy Liang.

[5] Language Models are Unsupervised Multitask Learners, Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever.

[6] METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, Satanjeev Banerjee, Alon Lavie.

[7] ROUGE: A Package for Automatic Evaluation of Summaries, Chin-Yew Lin.

[8] Texygen: A Benchmarking Platform for Text Generation Models, Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, Yong Yu.

[9] A Systematic Literature Review on Text Generation Using Deep Neural Network Models, Noureen Fatima, Sher Muhammaddaudpota, Ali Shariq Imran, (Member, Ieee), Zenun Kastrati And Abdullah Soomro.

[10] Stylized Data-to-Text Generation: A Case Study in the E-Commerce Domain, Liqiang Jing, Xuemeng Song, Xuming Lin, Zhongzhou Zhao, Wei Zhou, Liqiang Nie

[11] Context Aware Automatic Subjective and Objective Question Generation using Fast Text to Text Transfer Learning, Arpit Agrawal1, Pragya Shukla2.

[12] Pre-trained Language Models for Text Generation: A Survey, Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, Ji-Rong Wen.