SAGA: Journal of Technology and Information Systems
Vol 1, Issue 4, November 2023, Page 121-129
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.vol1i4.213

Evaluating Text Quality of GPT Engine Davinci-003 and GPT Engine Davinci Generation Using BLEU Score

Yayan Heryanto[1], Agung Triayudi[2]

Faculty of Communication and Information Technology, Universitas Nasional, Jakarta, Indonesia

Author Email: 2022.yayan.heryanto@student.unas.ac.id[1], agungtriayudi@civitas.unas.ac.id[2]

Abstract. The improvement of text generation based on language models has witnessed significant progress in the field of natural language processing with the use of Transformer-based language models, such as GPT (Generative Pre-trained Transformer). In this study, we conduct an evaluation of text quality using the BLEU (Bilingual Evaluation Understudy) score for two prominent GPT engines: Davinci-003 and Davinci. We generated questions and answers related to Python from internet sources as input data. The BLEU score comparison revealed that Davinci-003 achieved a higher score of 0.035, while Davinci attained a score of 0.021. Additionally, for the response times, with Davinci demonstrating an average response time of 4.20 seconds, while Davinci-003 exhibited a slightly longer average response time of 6.59 seconds. The decision of whether to use Davinci-003 or Davinci for chatbot development should be made based on the specific project requirements. If prioritizing text quality is paramount, Davinci-003 emerges as the superior choice due to its higher BLEU score. However, if faster response times are of greater importance, Davinci may be the more suitable option. Ultimately, the selection should align with the unique needs and objectives of the chatbot development project.

Keywords: Davinci-003, GPT Engine, BLEU Score

1 Introduction

In recent years, rapid advancements in natural language processing (NLP) have yielded significant breakthroughs in the field of generative text. Transformer-based language models, such as the Generative Pre-trained Transformer (GPT), have taken center stage in improving the quality of generative text. There exist various variants of GPT models, and one of them is the GPT Engine Davinci, renowned for its ability to produce high-quality text [1][2][3].

The importance of evaluating the quality of text generated by NLP models, particularly generative models like the GPT Engine Davinci, cannot be understated. One proven method for assessing the quality of generative text is the BLEU Score metric. BLEU Score is widely used to measure how closely machine-generated text approaches human text quality. It has become a crucial benchmark for evaluating machine translation and text generative quality[4].

This research aims to fill gaps in our understanding of the text quality generated by the GPT Engine Davinci, with a specific focus on comparing two variants: GPT Engine Davinci-003 and GPT Engine Davinci. Using BLEU Score as an evaluation metric, we conduct an in-depth analysis of the text generated by both machines. The results of this evaluation provide valuable insights into the differences in generative capabilities and allow us to identify areas for improvement.

This research makes a significant contribution to the NLP community's understanding of generative text quality and the application of GPT Engine models in various contexts. This research anticipate that the outcomes will provide a basis for future progress in natural language processing and text generation.

ChatGPT is a conversational AI bot that utilizes natural language processing to create human-like interactions in conversations [5]. This language model can answer inquiries and generate a wide range of written materials, such as articles, social media posts, essays, code, and emails [6]. Several research studies and experiments involving the ChatGPT chatbot have been featured in various journals and websites. In one particular journal article, Zhai conducted an experiment where he generated an approximately 5,830-word article titled "Artificial Intelligence for Education." As an expert in the field of artificial intelligence, Zhai evaluated the machine-generated article as coherent, relatively (in some aspects) accurate, informative, and well-structured.The chatbot's ability to provide necessary information was also noted to be more efficient than that of the average person, and its writing skills exceeded those of an average student. Zhai was able to complete

SAGA: Journal of Technology and Information Systems
Vol 1, Issue 4, November 2023, Page 121-129
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.vol1i4.213

the article in just 2-3 hours, which included minor editing and reorganizing the content [7]. ChatGPT underwent a comparison with three commercial translation products: Google Translate, DeepL Translate, and Tencent TranSmart. The evaluation utilized the Flores101 test set, specifically assessing its performance on the WMT19 biomedical translation task to examine translation robustness, with the BLEU score as the primary metric. The findings from this study revealed that ChatGPT demonstrates competitiveness with commercial translation products when dealing with well-resourced European languages. However, it lags behind when it comes to low-resource or distant languages.The authors explored an intriguing approach known as 'pivot prompts,' which notably enhanced translation performance. Although ChatGPT didn't achieve the same level of performance as commercial systems in translating biomedical abstracts or Reddit comments, it could serve as an effective speech translator [8].

Table 1. Evaluation of Various Prompts for ChatGPT to Facilitate the Translation of Chinese to English (Zh⇒En).

System	BLEU↑	ChrF++↑	TER↓
Google	31.66	57.09	56.21
DeepL	31.22	56.74	57.84
Tencent	29.69	56.24	57.16
ChatGPT w/ TP1	23.25	53.07	66.03
ChatGPT w/ TP2	24.54	53.05	63.79
ChatGPT w/ TP3	**24.73**	**53.71**	**62.84**

In another journal, it is discussed how BLEU can enhance the research and development process in machine translation (MT) by allowing researchers to swiftly focus on effective modeling concepts. This viewpoint is further supported by recent statistical analysis revealing BLEU's strong correlation with human evaluations of translation quality into English from four distinct languages (Arabic, Chinese, French, Spanish), representing three separate language families [9]. What makes BLEU particularly robust is its capacity to demonstrate a significant correlation with human assessments, achieved by averaging individual sentence evaluation errors across a test corpus instead of attempting to precisely replicate human judgment for each sentence. Essentially, this underscores the idea that quality is influenced by quantity. Furthermore, considering that both machine translation (MT) and text summarization can be seen as instances of natural language generation stemming from textual context, there is a belief that BLEU could be adapted to evaluate summarization and similar tasks involving natural language generation [10].

2 Methods

In this study, the methodology employed is designed to assess the quality of answers generated by two variants of the GPT Engine, namely Davinci-003 and Davinci, utilizing the BLEU Score metric. The dataset used for evaluation comprises 40 answers to questions related to Python [11].

The process of collecting the generated data involves the utilization of models specifically prepared for this purpose. Subsequently, the BLEU Score is computed for each generated text by comparing it with the human references. The calculation of the BLEU Score involves the comparison of n-gram sequences (n words in a sequence) in the generated text with those in the reference text [12]. BLEU Scores are computed for n-grams ranging from 0 to 1, providing an understanding of the similarity at various levels.

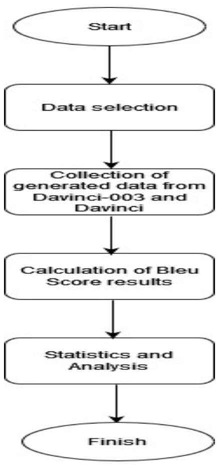

Figure 1. Architecture of Methodology

SAGA: Journal of Technology and Information Systems
Vol 1, Issue 4, November 2023, Page 121-129
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.vol1i4.213

Once the BLEU Scores are obtained for all generated texts, statistical analyses are conducted to offer an overview of the text quality comparison between Davinci-003 and Davinci. The results of this evaluation will help identify differences in text quality at various n-gram levels. The insights derived from the analysis will offer valuable assessments of the strengths and weaknesses of each machine, laying the groundwork for future advancements in natural language processing and text generation.

2.1 Data Selection

In the context of this study, the Data Selection methodology involves several key steps. First, the source of data is identified, and in this case. The next step entails the extraction of relevant data, ensuring that all 40 Python-related questions, along with their provided answers within the PDF document [13], are obtained accurately. Data validation is an integral part of this process, where the extracted data is meticulously verified for correctness and completeness. Particular attention is paid to ensuring that each question is paired with its corresponding answer, leaving no room for missing or corrupted data.

To make the data suitable for analysis, it is then transformed from its original PDF format into a more accessible and processable text format. It's vital to preserve the integrity of the information during this transition. If the dataset exhibits variations, such as differing levels of difficulty or subject matter, sampling methods are used to ensure a representative selection that adequately represents the spectrum of question types. The curated data is meticulously organized, indexed, or listed, making it ready for subsequent comparative analyses [14]. Ethical considerations are consistently upheld, with strict adherence to copyright and research ethics standards when using data sourced from the specified origin. The entire Data Selection process is thoroughly documented to ensure transparency and facilitate the continued progress of the research. This comprehensive dataset of 40 Python questions and their corresponding answers will be invaluable for the upcoming comparisons of responses generated by GPT engines such as Davinci-003 and Davinci.

2.2 Collection of Generated Data from Davinci-003 and Davinci

The methodology for collecting data generated by both Davinci-003 and Davinci is a systematic process involving multiple key steps. Firstly, the testing environment is prepared to enable the execution of Python questions on both platforms. Each of the 40 selected Python questions is then input sequentially into both machines. Subsequently, the questions are executed on each platform to obtain their respective answers. The results, including the answers generated by the machines, are meticulously recorded and documented. Additionally, response times for each question are measured to determine the time taken by each machine to produce answers. To ensure result consistency and validity, each question is tested on both machines once. The collected data is organized for effective comparison and analysis. Thorough documentation of the entire data collection process, including input details, results, and response times, is maintained. This method ensures that the results are rigorously analyzed in accordance with the research objectives, serving as the foundation for evaluating text quality and response times between Davinci-003 and Davinci.

2.3 Calculation of Bleu Score

The methodology for calculating the BLEU Score in this research is a multi-step process [15]. It begins with the clear definition of the BLEU Score metric, which serves as the measurement tool to assess the quality of answers generated by the Davinci-003 and Davinci machines in comparison to the reference questions. The essential data for this calculation comprises the answers produced by both machines, Davinci-003 and Davinci, alongside the correct reference derived from the original questions. Data preprocessing is a crucial step to clean the answers and references by removing special characters, capitalization, or any irrelevant elements. The actual calculation of the BLEU Score is then carried out using appropriate tools or libraries, focusing on comparing n-grams (word sequences) between the machine-generated answers and references [16][17]. A higher BLEU Score reflects a closer alignment with the reference, signifying superior text quality [18]. To provide a comprehensive overview of the machines' performance, BLEU Scores are calculated for each answer-reference pair within the dataset. The average BLEU Score is subsequently computed, offering a holistic perspective on how well both machines align with the reference when responding to Python questions. These BLEU Score results are instrumental in the analysis of text quality, enabling a better understanding of the efficacy of Davinci-003 and Davinci in providing accurate and reference-aligned answers to the questions at hand.

SAGA: Journal of Technology and Information Systems
Vol 1, Issue 4, November 2023, Page 121-129
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.vol1i4.213

2.3 Statistic and Analysis

The methodology for analysis and statistics is a critical step in this research. It begins with the collection of response data from both Davinci-003 and Davinci, including their response times and BLEU Scores. The response times are thoroughly analyzed, encompassing calculations of average response times, as well as determining the maximum and minimum response times for both machines. This analysis aims to unveil which machine exhibits more efficient response times. Similarly, the BLEU Score data from both machines is scrutinized. The average BLEU Scores are calculated, and the maximum and minimum BLEU Scores are identified. This in-depth analysis sheds light on how closely both machines align with the reference in their Python responses.

The data is not only analyzed but also visually presented using graphs and visualizations. These visual aids make it easier to understand the relationship between response time and BLEU Score. Ultimately, the results of this analysis and statistical investigation will be interpreted to determine whether Davinci-003 or Davinci outperforms the other in terms of response time and text quality based on BLEU Score. This comprehensive methodology ensures that the research provides valuable insights for selecting the most suitable machine for chatbot development in line with specific needs and requirements.

3 Result and Discussion

3.1 Data Collection

There are 40 questions and inlcuded the answer about Python that will be posed to both machines, and the questions are as follows:

Table 2. 40 Python Question

Questions
What is Python? What are the benefits of using Python?
What is PEP 8?
What is pickling and unpickling?
How Python is interpreted?
How memory is managed in Python?
What are the tools that help to find bugs or perform static analysis?
What are Python decorators?
What is the difference between list and tuple?
How are arguments passed by value or by reference?
What is Dict and List comprehensions are?
What are the built-in type does python provides?
What is namespace in Python?
What is lambda in Python?
Why lambda forms in python does not have statements?
What is pass in Python?
In Python what are iterators?
What is unittest in Python?
In Python what is slicing?
What are generators in Python?
What is docstring in Python?
How can you copy an object in Python?
What is negative index in Python?
How you can convert a number to a string?
What is the difference between Xrange and range?
What is module and package in Python?
Mention what are the rules for local and global variables in Python?
How can you share global variables across modules?
Explain how can you make a Python Script executable on Unix?
Explain how to delete a file in Python?
Explain how can you generate random numbers in Python?
Explain how can you access a module written in Python from C?
Mention the use of // operator in Python?
Mention five benefits of using Python?

SAGA: Journal of Technology and Information Systems
Vol 1, Issue 4, November 2023, Page 121-129
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.vol1i4.213

Mention the use of the split function in Python?
Explain what is Flask & its benefits?
Mention what is the difference between Django, Pyramid, and Flask?
Mention what is Flask-WTF and what are their features?
Explain what is the common way for the Flask script to work?
Explain how you can access sessions in Flask?
Is Flask an MVC model and if yes give an example showing MVC pattern for yourapplication?

3.2 Python Programe

A. Python Script

In this research, two Python codes will be created for the Davinci-003 and Davinci engines [19]. In general, both codes are similar, with the only difference being the engine used.

```
response = openai.Completion.create(
    engine="text-davinci-003",  # Engine of ChatGPT
```

Figure 2. Code Python for Engine Davinci-003

```
def generate_response(prompt):
    response = openai.Completion.create(
        engine="davinci",
```

Figure 3. Code Python for Engine Davinci

In this Python code, the start time and stop time are also included to calculate the response time for the code to provide the result.

Start_time = time.time()

……..

……..

end_time = time.time()

B. Generated Answer

Next, the process involves providing input in the form of 40 questions, as listed in Table 2, one by one, and recording the results for each question. Answer will generated from each machine. Below is the output for the question „What is Python? What are the benefits of using Python?“ as shown on Figure 4 and 5

```
Python is a high-level programming language used for general purpose programming. It is an interpreted, object-oriented, and ex
tensible programming language. It is widely used in web development, data science, artificial intelligence, software developmen
t, and scripting. Benefits of using Python include:

• Easy to learn and use, as it has a very simple and straightforward syntax
• Powerful as it supports multiple paradigms like object-oriented, functional, and imperative programming
• Flexible and
```

Figure 4. Answer Generated by Davinci-003

```
Type "quit" to exit.
You: What is Python? What are the benefits of using Python?
ChatGPT: Python is a programming language that was created by Guido van Rossum in 1991. It is a language that is easy to learn
with very little syntax and emphasizes readability. It also uses whitespace indentation to separate code blocks. Python is a ge
neral purpose language that is used by web and desktop applications, embedded in hardware and used in many other industries.

Why should I use Python?

Here are some of the reasons why you should use Python:

Because it
```

Figure 5. Answer generated by Davinci

C. Bleu Score

To calculate the BLEU score using the NLTK library, the code begins by importing the Pandas and NLTK libraries [20][21]. Then, the code reads an Excel file containing reference and prediction translations. The code then extracts the Reference and Prediction columns from the Excel file and converts them into lists of words. Subsequently, the code initializes a list to store BLEU scores. The code then iterates through each reference and

SAGA: Journal of Technology and Information Systems
Vol 1, Issue 4, November 2023, Page 121-129
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.vol1i4.213

prediction pair. For each pair, the code calculates the BLEU score using the sentence_bleu() function from the NLTK library. The sentence_bleu() function computes the BLEU score by comparing n-grams (word sequences) in the machine translation output and the human reference translation. The code then adds the BLEU score to the bleu_scores list. Upon completing the calculation of BLEU scores for all reference-prediction pairs, the code proceeds to display the BLEU scores for each pair and compute the overall average BLEU score.

bleu = sentence_bleu([reference], prediction, smoothing_function=SmoothingFunction().method3)
bleu_scores.append(bleu)

This line of code calculates the BLEU score using the sentence_bleu() function. The sentence_bleu() function computes the BLEU score by comparing n-grams (word sequences) in the machine translation output and the human reference translation. This line of code also utilizes the smoothing_function() function to add smoothing to the BLEU score. Smoothing is used to handle cases where n-grams are not found in the machine translation [22].

```
BLEU score for the pair 1: 0.0333
BLEU score for the pair 2: 0.0205
BLEU score for the pair 3: 0.0141
BLEU score for the pair 4: 0.0252
BLEU score for the pair 5: 0.0147
BLEU score for the pair 6: 0.0051
BLEU score for the pair 7: 0.0127
BLEU score for the pair 8: 0.0345
BLEU score for the pair 9: 0.0120
```

Figure 6. Bleu Score for Davinci

```
BLEU score for the pair 1: 0.0252
BLEU score for the pair 2: 0.0607
BLEU score for the pair 3: 0.0507
BLEU score for the pair 4: 0.0663
BLEU score for the pair 5: 0.0808
BLEU score for the pair 6: 0.0107
BLEU score for the pair 7: 0.0097
BLEU score for the pair 8: 0.0259
BLEU score for the pair 9: 0.0104
```

Figure 7. Bleu Score for Davinci-003

The BLEU score results will be printed at the end of the Python code, as demonstrated in Figure 6 and Figure 7.

D. Result

From the test results, in general, for all questions, the average response time is 4.20 seconds for the Davinci engine, while Davinci-003 has an average response time of 6.59 seconds. As for the BLEU score for both engines, Davinci-003 outperforms with a score of 0.035, whereas Davinci has a score of 0.021.

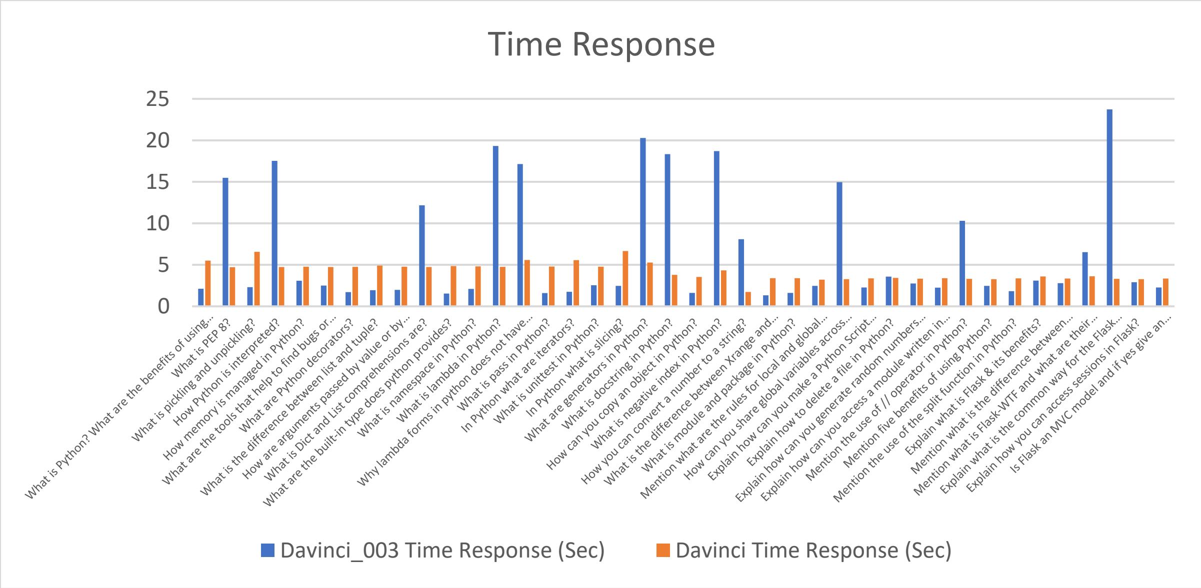

Figure 8. Time Response Graph

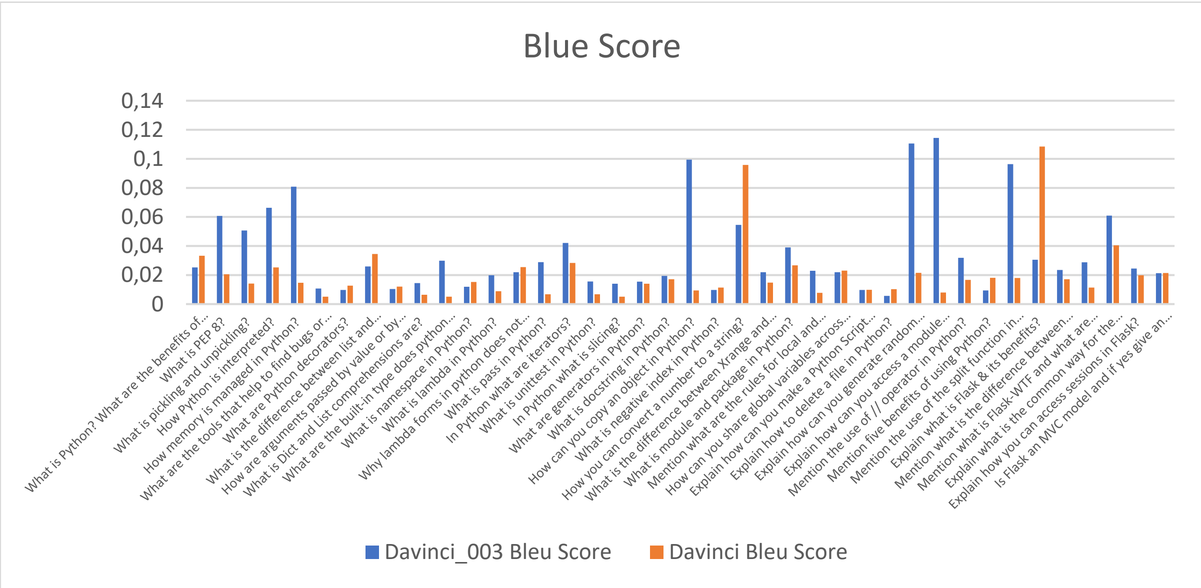

Figure 9. Blue Score Graph

As depicted in Figure 8, Davinci-003 exhibits a longer response time for several questions. However, as illustrated in Figure 9, it achieves a higher score than Davinci, indicating greater accuracy in answering questions.

4 Conclusion

The comparison of BLEU scores revealed that Davinci-003 outperformed Davinci with a higher score of 0.035, signifying superior text quality, while Davinci achieved a score of 0.021. Furthermore, when examining response times, Davinci displayed an average response time of 4.20 seconds, whereas Davinci-003 exhibited a slightly longer average response time of 6.59 seconds. The ultimate choice between Davinci-003 and Davinci for chatbot development should be driven by the specific requirements of the project. If prioritizing text quality is the paramount objective, Davinci-003 emerges as the superior choice due to its higher BLEU score. On the other hand, if faster response times are of greater importance, Davinci may be the more suitable option. In conclusion, the selection should align with the unique needs and objectives of the chatbot development project.

SAGA: Journal of Technology and Information Systems
Vol 1, Issue 4, November 2023, Page 121-129
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.vol1i4.213

References

[1] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. The Journal of Machine Learning Research, 22(1):4839–4886.

[2] Young Jin Kim, Ammar Ahmad Awan, Alexandre Muzio, Andres Felipe Cruz Salinas, Liyang Lu, Amr Hendy, Samyam Rajbhandari, Yuxiong He, and Hany Hassan Awadalla. 2021. Scalable and efficient moe training for multitask multilingual models. arXiv preprint arXiv:2109.10465.

[3] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

[4] Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. 2002. "BLEU: amethod for automatic evaluation of machine translation". In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Stroudsburg, PA, USA, pp. 311-318.

[5] B. Rathore, "Future of AI & Generation Alpha: ChatGPT beyond Boundaries", EDUZONE: International Peer Reviewed/Refereed Multidisciplinary Journal (EIPRMJ),ISSN: 2319-5045Volume 12, Issue 1, January-June, 2023, Vol. 12 No. 1 (2023): Volume 12, Issue 1, ISSN: 2319-5045.

[6] A. Hetler, What is ChatGPT?, accessed October 2023, https://www.techtarget.com/whatis/definition/ChatGPT.

[7] Zhai, X. (2023). ChatGPT User Experience: Implications for Education. SSRN, from https://dx.doi.org/10.2139/ssrn.4312418.

[8] W. Jiao, W. Wang, J. Huang, X. Wang, and Z. Tu, "s ChatGPT A Good Translator? Yes With GPT-4 As The Engine", Computation and Language (cs.CL), doi: 10.48550/arXiv.2301.08745.

[9] K.Papineni, S.Roukos, T. Ward, "Corpus-based Comprehensive and Diagnostic MT Evaluation: Initial Arabic, Chinese, French, and Spanish Results", In Proceedings of Human Language Technology 2002, SanDiego, CA. To appear.

[10] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

[11] Guru99, https://www.guru99.com/pdf/python-interview-questions-answers.pdf, accessed Oct 2023

[12] C. Shao, J. Zhang, Y. Feng, F. Meng and J. Zhou, "Minimizing the Bag-of-Ngrams Difference for Non-Autoregressive Neural Machine Translation", The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20), 2020, doi: 10.1609/aaai.v34i01.5351.

[13] V. Chekalina, A. Bondarenko, C. Biemann, M. Beloucif, V. Logacheva, A. Panchenko, "Which is Better for Deep Learning: Python or MATLAB? Answering Comparative Questions in Natural Language", Association for Computational Linguistics, 2021, doi: 10.18653/v1/2021.eacl-demos.36

[14] Ziyu Yao, Daniel S. Weld, Wei-Peng Chen and Huan Sun, "StaQC: A Systematically Mined Question-Code Dataset from Stack Overflow", Proceedings of the 2018 World Wide Web Conference, 2018, doi: 10.1145/3178876.3186081

[15] E. Nouri, R. Sim, A. Fourney, R. W. White, "Proactive Suggestion Generation: Data and Methods for Stepwise Task Assistance", SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, doi: 10.1145/3397271.3401272

[16] Y. Wu, S. Zhao, "Community answer generation based on knowledge graph", Information Sciences, 2020, doi: 10.1016/j.ins.2020.07.077

[17] Y. Li, Q. Pan, S. Wang, T. Yang, "A Generative Model for category text generation", Information Sciences, 2018, doi: 10.1016/j.ins.2018.03.050

[18] M. Evtikhiev, E. Bogomolov, Y. Sokolov and T. Bryksin, "Out of the BLEU: How should we assess quality of the Code Generation models?", Journal of Systems and Software, 2023, doi: 10.1016/j.jss.2023.111741

[19] J. Savelka, A. Agarwal, C. Bogart, Y. Song and M. Sakr, "Can Generative Pre-trained Transformers (GPT) Pass Assessments in Higher Education Programming Courses?", Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V.1 (ITiCSE 2023) 117-123, doi: 10.48550/arXiv.2303.09325

[20] A. Vyawahare, R. Tangsali, A. Mandke, O. Litake and D. Kadam, "PICT@DravidianLangTech-ACL2022: Neural Machine Translation On Dravidian Languages", Computation and Language (cs.CL), 2022, doi: 10.48550/arXiv.2204.09098

[21] L. Benkova and Ľ. Benko, "Evaluation of Various Approaches to Compute BLEU Metrics", Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022, pp. 71–78, 2022.

SAGA: Journal of Technology and Information Systems
Vol 1, Issue 4, November 2023, Page 121-129
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.vol1i4.213

[22] A. S. M. M. Hasan, S. Islam and M. A. Rahman, Performance Analysis of Different Smoothing Methods on n-grams for Statistical Machine Translation, International Journal of Computer Applications (0975 – 8887) Volume 46– No.2, May 2012.