e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

EXPLORING THE IMPACT OF PROMPT ENGINEERING ON CHATGPT 3.5 TEXT SUMMARIZATION: A BERT SCORE EVALUATION

Kartik Ashok Hawelikar*1

*1Student, Department Of Information Technology, B. K. Birla College Of Arts, Science & Commerce (Autonomous), Kalyan, Maharashtra, India.

DOI : https://www.doi.org/10.56726/IRJMETS45268

ABSTRACT

In the domain of Natural Language Processing (NLP), the technique of prompt engineering is a strategic method utilized to guide the responses of models such as ChatGPT. This research explores the intricacies of prompt engineering, with a specific focus on its effects on the quality of summaries generated by ChatGPT 3.5, an openly accessible chatbot developed by OpenAI. The study encompasses a comprehensive examination of 110 summaries produced from ten diverse paragraphs, employing eleven distinct summarization prompts under zero-shot setting. Evaluation is conducted using the BERT Score, a metric that offers a more contextually relevant assessment of summary quality. This study introduces an innovative approach to appraising the quality of summaries, setting it apart from prior investigations and delivering valuable insights into the nuances of prompt engineering's role within the NLP landscape. Ultimately, this inquiry illuminates the strengths and weaknesses associated with various prompts and their influence on ChatGPT 3.5's summarization capabilities, thereby making a significant contribution to the constantly evolving field of NLP and automated text summarization.

Keywords: Prompt Engineering, Chatgpt 3.5, Summarization Prompts, BERT Score, Zero-Shot Prompting, NLP Research.

I. INTRODUCTION

In recent years, natural language processing (NLP) models have made significant advancements in the field of text summarization, offering the potential to enhance the accessibility and comprehensibility of complex medical reports for a wider audience, including patients. Prompt engineering, a natural language processing (NLP) concept is the deliberate construction of specific instructions or queries to elicit desired responses from natural language processing models like GPT-3.5, tailored to a particular task or outcome. In generative AI, prompts are inputs or queries that a user or program gives to an AI model. The model analyzes the prompt and generates a response based on the patterns it has learned through its training. The quality of input will determine the quality of output. Prompt engineering has emerged as an essential technique to influence the quality of generated summaries. The choice of summarization prompts is a crucial factor in determining the quality and effectiveness of generated summaries. Prompting methods such as zero-shot and few-shot prompting have emerged as pivotal techniques in the field of natural language processing, revolutionizing the way we interact with language models. Zero-shot prompting is a technique in natural language processing where a model generates text or responses without any prior training on specific examples, relying solely on its pre-existing knowledge and understanding of language, whereas few-shot prompting takes advantage of a limited amount of training data, typically a small number of examples or prompts, to enhance the model's ability to generate coherent and contextually relevant text or responses, enabling it to perform better in tasks that require specific knowledge or context.

ChatGPT, short for Chat Generative Pre-trained Transformer, is a cutting-edge chatbot developed by OpenAI and introduced on November 30, 2022. It empowers users to guide conversations to their preferred length, style, level of detail, and language through a technique called prompt engineering. Initially launched as a freely accessible research preview, its immense popularity led OpenAI to adopt a freemium model, providing GPT-3.5-based functionality for free, while offering a more advanced GPT-4-based version and exclusive access to new features through a paid subscription service known as "ChatGPT Plus." By January 2023, ChatGPT had achieved unprecedented success, amassed over 100 million users, and significantly boosting OpenAI's valuation to $29 billion. This rapid rise prompted competitors like Google, Baidu, and Meta to expedite the development of their own conversational AI products, including Bard, Ernie Bot, and LLaMA. Microsoft also entered the arena with

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

Bing Chat, built on OpenAI's GPT-4. ChatGPT [3] is a modified GPT-3 model, GPT-3.5, with 6.7 billion parameters (compared to GPT-3's 175 billion). It excels in various natural language tasks, including summarization, thanks to training on a large text corpus and fine-tuning for conversational responses.

Automatic text summarization is a fundamental Natural Language Processing (NLP) task that aims to condense lengthy text documents into shorter, coherent versions while retaining essential information. Evaluating the quality of generated summaries is a crucial step in assessing the performance of summarization algorithms. BERT Score [1], a metric based on contextual embeddings, has gained popularity for this purpose due to its ability to capture semantic similarity between texts. It is a metric for evaluating the quality of text generation models, such as machine translation or summarization. It calculates the similarity between two sentences as a sum of cosine similarities between their tokens' embeddings. It was invented as an improvement on n-gram-based metrics like BLEU. It focuses on computing semantic similarity between tokens of reference and hypothesis. It correlates with human judgment on sentence-level and system-level evaluation.

This research investigates the impact of different summarization prompts on the quality of summaries generated using ChatGPT 3.5(freely available chatbot by open ai), shedding light on the strengths and weaknesses of various prompts. The zero-shot prompting method is used to generate a total of 110 summaries using 11 summarization prompts from 10 different paragraphs from diverse sources such as Wikipedia, News articles, Amazon product reviews, Research papers and Reddit posts. These summaries are then assessed using BERT Score to provide a comprehensive evaluation of summary quality. This study distinguishes itself from prior research by opting to assess summary quality using BERTScore rather than the conventional use of ROUGE metrics.

II. LITERATURE REVIEW

The evolution of natural language processing (NLP) technologies, particularly those employing zero-shot prompting for natural language generation, has witnessed significant strides in recent years. [7] Radford et al. (2019) laid the foundation for this field with the introduction of GPT-2, a groundbreaking model showcasing the potential to generate coherent and diverse texts from simple prompts. This pioneering work marked the initial foray into zero-shot prompting for NLP tasks.

Building upon Radford's work, [6] Brown et al. (2020) elevated the field further with the development of GPT-3, a model that exhibited remarkable performance in natural language understanding and generation tasks. It demonstrated its versatility in zero-shot and few-shot prompting, showcasing the immense potential of AI models in enhancing human-computer interactions. Since then, numerous researchers have embarked on exploring the effectiveness and limitations of zero-shot prompting across diverse NLP tasks, such as sentiment analysis, text summarization, question answering, and content generation.

[5] Chung et al. (2023) illustrated how ChatGPT can distill complex prostate MRI reports into easily comprehensible language suitable for patients. This innovation holds immense promise for improving patient-doctor communication and underscores the growing role of AI-driven tools in healthcare. Beyond summarization, [2] Luo et al. (2023) revealed another groundbreaking aspect of ChatGPT – its proficiency in evaluating the factual correctness, clarity, and completeness of generated summaries under a zero-shot setting. This unique capability distinguishes ChatGPT from prior state-of-the-art evaluation methods, signifying a significant advancement in the assessment of automated text summarization.

The findings of [4] Yang et al. (2023) provide further insights into ChatGPT's capabilities. Their experiments demonstrated that ChatGPT's performance in terms of Rouge scores rivals that of traditional fine-tuning techniques. Moreover, these experiments unveiled striking disparities between summaries generated by ChatGPT and those crafted by humans.

III. METHODOLOGY

To investigate the impact of prompt engineering on ChatGPT 3.5 text summarization, the approach encompasses data collection, prompt design, prompts classification, summarization process, evaluation process and data analysis.

Data Collection

A heterogeneous dataset comprising ten paragraphs was assembled from diverse sources.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

The dataset consists of two paragraphs taken from Wikipedia, followed by two paragraphs from news articles, then two from Amazon product reviews, and finally, two from research papers, all in the exact sequence as described.

Prompt design

For guiding ChatGPT 3.5 in generating summaries of the collected paragraphs, eleven prompts were crafted. These prompts ranged from the most common and basic to advanced and instructive. The prompts designed were as follows:

1) Summarize the given paragraph.

2) Summarize the following paragraph in a concise manner, capturing the main points and key information.

3) I want you to act as a Summary Writer. Summarize the given paragraph.

4) I want you to act as a Summary Writer. Summarize the given paragraph in one sentence.

5) I want you to act as a Summary Writer. Summarize the given paragraph in one sentence without losing its meaning.

6) I want you to act as a Summary Writer. Summarize the given paragraph without losing its meaning.

7) Summarize the given paragraph without losing its meaning

8) Summarize the given paragraph in one sentence.

9) Summarize the given paragraph in one sentence losing its meaning

10) I want you to act as a Summary Writer. Summarize the given paragraph in a concise manner, capturing the main points and key information without losing its meaning.

11) I want you to act as a Summary Writer. Summarize the given paragraph in a concise manner, capturing the main points and key information without losing its meaning in one sentence.

Prompts Classification

The eleven prompts were categorized into two distinct groups. The first group comprises prompts designed to generate one-sentence summaries (Group 1-One sentence Summary Prompts: 4,5,8,9,11), while the second group includes prompts for generating summaries without a specified word limit (Group 2 -Summary prompts without words limit:1,2,3,6,7,10).

Summarization Process

In the summarization process, ChatGPT, the open AI chatbot (free version), was employed to generate summaries under the zero-shot setting. Each of the ten paragraphs was prompted using the designated prompts as shown in Fig.1 in the following format:

<Summarization prompt>

<Paragraph enclosed within double quotation marks>

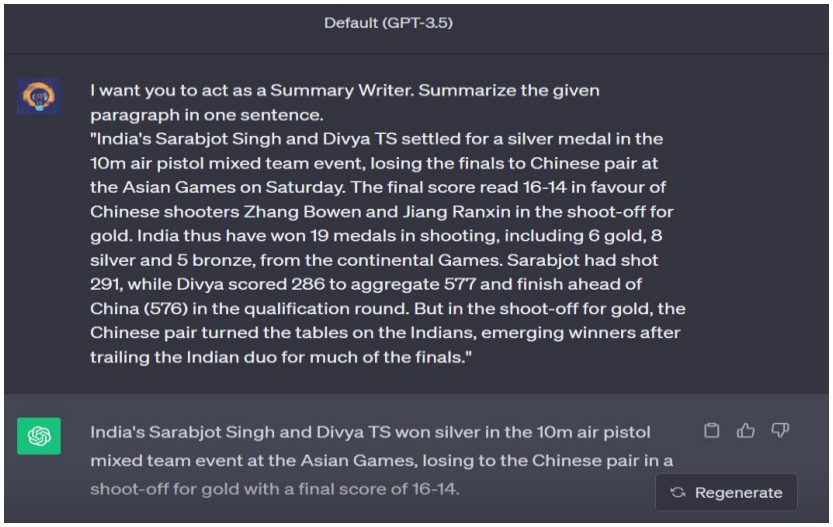

Figure 1: Example of prompt format and output.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

This approach resulted in the generation of a total of 110 summaries, encompassing all ten paragraphs and utilizing the eleven prompts (i.e., 10 paragraphs * 11 prompts = 110 summaries). No regenerated output summary was considered in the analysis.

Evaluation process

To assess the quality of these 110 summaries, BertScore, a state-of-the-art metric, is employed for evaluation.

In subsequent analyses, the prompts were segregated into two distinct groups: Group 1, consisting of one-sentence summary prompts, and Group 2, comprising prompts without word limits as mentioned in the Prompt classification step. These groupings allow us to comprehensively evaluate and compare the performance of different prompt types.

Data Analysis

Separate analyses were conducted for Group 1 and Group 2 prompts. BertScore was calculated for each summary, based on its semantic similarity to the corresponding reference paragraph. This approach ensured that the evaluation was contextually relevant and meaningful, assessing how effectively each summary captured the main points and key information contained within its respective paragraph.

IV. RESULTS AND DISCUSSION

The statistical data for the BERTScore evaluation across all summaries are presented in Table 1. The overall mean BERTScore across 110 summaries was 0.8739, with a standard deviation of 0.0318. The minimum BERTScore observed was 0.8218, while the maximum was 0.9441, indicating a variation in the quality of generated summaries. Also on average, the generated summaries have a good level of semantic similarity to the reference paragraphs.

Table 1. Overall Statistics

count (No. of summaries)	110
Mean	0.873992666
Std	0.031792186
Min	0.821787953
25%	0.852032587
50%	0.866524071
75%	0.898677856
max	0.944123924

Paragraph-wise analysis revealed that the quality of generated summaries varied across different paragraphs. The mean BERTScore for each paragraph is shown in Table 2. It indicates the quality of summaries generated for different source materials. Paragraph 3 has the highest mean BERTScore (0.925761109), indicating that the summaries generated for this news article are generally of high quality. On the other hand, Paragraph 5 has the lowest mean BERTScore (0.842), indicating that generating quality summaries for product reviews is more challenging.

Table 2. Overall Statistics of Group 1 and Group 2

Paragraph	Mean BERT Score
1	0.885931644
2	0.86029225
3	0.925761109
4	0.876919801
5	0.842283677
6	0.868397902
7	0.898277088

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

8	0.898963115
9	0.84263425
10	0.840465827

The mean BERT Scores and summary statistics for both Group 1 and Group 2 prompts are presented in Table 3. For Group 1 (one-sentence summary prompts), the mean BERT Score was 0.8608, with a standard deviation of 0.0275. Group 2 (summary prompts without word limits) achieved a higher mean BERT Score of 0.8850, with a standard deviation of 0.0312. This indicates that the absence of word limits (Group 2) allows for better quality summaries on average. It suggests that not constraining the length of the summary may lead to more comprehensive and contextually relevant results.

Table 3. Overall Statistics of Group 1 and Group 2

	Group 1 Summary Stats	**Group 2 Summary Stats**
count	50	60
mean	0.860834665	0.884957667
std	0.027481642	0.031162146
min	0.821787953	0.833335161
25%	0.835519835	0.859238848
50%	0.857342452	0.879438013
75%	0.877861932	0.910837352
max	0.920332909	0.944123924

Table 4 presents summary statistics for Group 1 prompts, which were designed for generating one-sentence summaries and Group 2 prompts, which allowed for generating summaries without specific word limits. The minimum BERT Score observed for group 1 prompts was 0.8218, while the maximum BERT Score reached 0.9203. The mean BERT Score for this group was approximately 0.8610, with some variability among individual prompts. Prompt 8 and Prompt 9 have the highest mean BERT Scores, both around 0.862. Prompt 11 has the lowest mean BERT Score at 0.859. This indicates that for one-sentence summaries, prompts 8 and 9 are particularly effective, while Prompt 11 performs slightly less well. There is a range of BERT Scores within this group, from 0.824 to 0.911, demonstrating that even within one-sentence summaries, there is variability in summary quality.

Table 4. Overall Statistics of Group 1 and Group 2

Group 1								
Summary no. (Prompt no.)	**count (No. of Paragraphs)**	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**
4	10	0.859594	0.026376	0.824909	0.83766	0.861193	0.87185	0.911158
5	10	0.861729	0.030675	0.821788	0.839651	0.857342	0.882096	0.912391
8	10	0.861735	0.027961	0.827587	0.837199	0.859116	0.877862	0.911962
9	10	0.86188	0.029069	0.827435	0.839097	0.85382	0.876807	0.920333
11	10	0.859236	0.028997	0.825424	0.839572	0.855472	0.87913	0.916113
Group 2								

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:10/October-2023 Impact Factor- 7.868 www.irjmets.com

1	10	0.881088	0.033663	0.842591	0.85292	0.872728	0.907753	0.944124
2	10	0.876285	0.032511	0.833335	0.857542	0.874263	0.895239	0.941669
3	10	0.889909	0.033054	0.845654	0.863447	0.880757	0.917032	0.934845
6	10	0.897608	0.031357	0.857744	0.866036	0.90155	0.924571	0.94143
7	10	0.884363	0.028886	0.84604	0.867346	0.877674	0.906985	0.935778
10	10	0.880493	0.030624	0.841299	0.856992	0.87849	0.904379	0.929487

The minimum BERT Score for group 2 was 0.8333, and the maximum BERT Score was 0.9414. Prompt 6 yields the highest mean BERT Score (0.898), while Prompt 1 has the lowest mean BERT Score (0.881). This group demonstrates a higher overall mean BERT Score in comparison to Group 1, suggesting that permitting more flexibility in the length of the summary generally results in better-quality summaries. BERT Scores within this group span from 0.833 to 0.944, indicating that even without word limits, there is still variability in the quality of the generated summaries.

The quality of generated summaries is notably influenced by the source material's content and complexity. Research papers and Wikipedia articles tend to yield higher mean BERT Scores, whereas Reddit posts and product reviews typically result in lower scores. This difference may be attributed to the level of complexity and informativeness of the source content. It is important to note that even within the same group (Group 1 or Group 2), there is variability in the quality of summaries generated by different prompts. This underscores the importance of choosing the right prompt wording as well as prompt format for achieving the desired summary quality. It's apparent that certain prompts perform better with specific paragraphs. Recognizing these prompt-paragraph combinations could lead to tailored strategies for improving summary quality on a per-content basis.

While BERT Score proves to be a valuable metric for assessing summary quality, other factors like coherence, fluency, relevance, and additional evaluation metrics such as ROUGE or human assessment should be explored to gain a more comprehensive understanding of the performance of different prompts and models. A comprehensive evaluation should encompass not only summary quality but also these critical aspects to assess the suitability of generated summaries for specific applications. This research offers a unique perspective on prompt engineering in the context of NLP and text summarization. It raises questions about the need for adaptive prompt selection and automated methods for prompt optimization to enhance summary quality consistently. The choice of prompts continues to play a significant role in determining the quality of summaries for each source. Different prompts, even when applied to the same source, may result in varying BERT Scores. This suggests that prompt engineering is crucial in achieving the desired quality of summarization.

V. CONCLUSION

This research focused on the critical role of prompt engineering in the field of Natural Language Processing, particularly in the context of text summarization using ChatGPT 3.5. By employing a diverse dataset and 11 distinct summarization prompts, the study evaluated the quality of 110 generated summaries using BERT Score, a contextually relevant metric. The results showed variations in summary quality based on source material and prompt type, highlighting the importance of prompt selection for achieving desired summarization outcomes. While BERT Score proved to be a valuable evaluation metric, the study emphasized the need for considering additional factors such as coherence, fluency, and relevance in future research. Overall, this research contributes valuable insights to the evolving landscape of NLP and automated text summarization, emphasizing the significance of prompt engineering in influencing the quality of generated summaries.

VI. REFERENCES

[1] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," arXiv.org. Accessed: Oct. 10, 2023. [Online]. Available: https://arxiv.org/abs/1904.09675v3

[2] Z. Luo, Q. Xie, and S. Ananiadou, "ChatGPT as a Factual Inconsistency Evaluator for Text Summarization." arXiv, Apr. 13, 2023. Accessed: Oct. 10, 2023. [Online]. Available: http://arxiv.org/abs/2303.15621

[3] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," Internet of Things and Cyber-Physical Systems, vol. 3, pp. 121–154, Jan. 2023, doi: 10.1016/j.iotcps.2023.04.003.

[4] X. Yang, Y. Li, X. Zhang, H. Chen, and W. Cheng, "Exploring the Limits of ChatGPT for Query or Aspect-based Text Summarization." arXiv, Feb. 15, 2023. Accessed: Oct. 10, 2023. [Online]. Available: http://arxiv.org/abs/2302.08081

[5] E. M. Chung, S. C. Zhang, A. T. Nguyen, K. M. Atkins, and M. Kamrava, "Feasibility and Acceptability of ChatGPT Generated Radiology Report Summaries for Cancer Patients," International Journal of Radiation Oncology, Biology, Physics, vol. 117, no. 2, p. e463, Oct. 2023, doi: 10.1016/j.ijrobp.2023.06.1662.

[6] T. B. Brown et al., "Language Models are Few-Shot Learners." arXiv, Jul. 22, 2020. doi: 10.48550/arXiv.2005.14165.

[7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners".