Analysis of ChatGPT as a Question-Answering Tool

Pit Pichappan
Digital Information Research Labs
Chennai 600017, India
pichappan@dirf.org

M. Krishnamurthy
Documentation Research & Training Center
Indian Statistical Institute, Bangalore 5600059, India
mkrishnamurthy1304@gmail.com

P. Vijayakumar
Central University, Pondicherry 6050105
India
pvijai@gmail.com

JDIM

Journal of Digital Information Management

ABSTRACT: *ChatGPT, in recent months, has made a significant impact and exposure in the information world. Many studies have been conducted within a shorter timeframe about its efficiency, reliability, ethics, accuracy and acceptance. Besides, hundreds of opinions and perception-based analyses have also emerged. In this work, we look at the ChatGPT as a question-answering tool. We have used randomly generated prompts to solicit answers and analysed the results from a text analysis angle. The answers are compared with text analysers both manually and statistically. ChatGPT still needs more precision for linguistic effects and fails to meet comprehensive users' requirements.*

Subject Categories and Descriptors: [I.2 Artificial Intelligence]: Natural language interfaces **[I.2.4 Knowledge Representation]:** Formalisms and Methods; **[H.3.3 Information Search and Retrieval]**: Query formulation

General Terms: Artificial Intelligence Tools, ChatGPT, Natural Language Processing, Text Processing, Question-Answering

Keywords: ChatGPT, Question-answering tools, Text Analysis, AI tools, Natural Language Analysis

Received: 31 March 2023, Revised 28 April 2023, Accepted 9 May 2023

Review Metrics: Review Scale: 0/6, Review Score 4.65, Inter-reviewer Consistency: 80.8%

DOI: 10.6025/jdim/2023/21/2/50-61

1. Introduction

The infusion of AI techniques in Natural Language Processing (NLP) leads to the construction of sentences based on predictive transformers, which dictate sequence during sentence formation. Based on statistical probability, the language models form sentences with training. The significant language model is the Generative Pre-trained Transformer, and its first version GPT-1, was introduced in 2018, followed by GPT -2 and GPT- 3. GPT-3 has impacted users more than its predecessors as it holds many unparallel features. The GPT-3 uses a neural network which is termed a transformer. GPT-3 takes in a sequence of words and uses multiple layers of statistical probability to analyse the relationships between phrases and predict the next word in the sequence. GPT-3 is used for many tasks, such as language translation, summarisation, question answering, text generation, code generation etc. [1]. Dialogue generation, Text generation, Sentiment analysis, Language translation, Content generation, Classification Some applications identified include unstructured data, Information answering, Hyperpersonalization, User behaviour identification, and Improved analytics.

The attributed reason for the effectiveness of Chat GPT is the use of 175 billion parameters with ”variables in an AI system whose values are changed during training to find how the input information gets transformed into the expected output” [2,3,4].

The ChatGPT is developed as an Artificial General Intelligence tool to impart human cognitive ability. The ChatGPT ensures the patterns within the language using syntax, grammatical features, and semantics [5,6,7,8].

2. Background

As the GPT has the potential to answer any question, the academic world depends on it for knowing, learning, and writing texts, thus making the students get answers without foraging large text collections. Extensive use of GPT leads to leaving the knowledge forage activities of the students and consequently harm learning. Many researchers expressed concerns about losing creativity and end-user information foraging capabilities. The development of AI is viewed as a technological development rather than a scientific progress. Extensive use of this technical tool affects the learning process and harms creating scientific breakthroughs in future.

Whether this technological development coincides with cognition and semantic characteristics of natural language is an issue. So, we must research this issue before applying AI progress in Information handling activities. In this work, we intend to address this issue not on a broad level but specifically, the use of the specific AI tool, ChatGPT, in question answering. This work is structured as follows. We briefly review the research on the AI-generated text, followed by a brief discussion on question-answering systems. We then introduce the debate on the dataset, analysis and outcome of the exercise on the question-answering capability of the ChatGPT.

2.1. Earlier Studies and AI-generated Text

ChatGPT is viewed as an effective tool in conversation with users, seemingly naturally and intuitively [9]. The OpenAI introduced Generative Pre-Trained Transformer (GPT-)3 in 2020 as a significant AI development. GPT-3 was developed by training a large corpus with 45 terabytes of text [10]. Its data sources are drawn from Common Crawl (a non-profit organisation that crawls the web data and makes it available to open the stored information and data volume open), *WebText2, Books* and *Wikipedia*. *WebText2* is the extraction of the text of web pages using the source Reddit links from posts with a minimum of three upvotes, and *Books* is based on *Books1* & *Books2*, which are two web books corpus [11].

We summarise the crucial research progress related to GPT-3 and its predecessor, GPT-2, in higher education and research. Dehouche (2021) [12] studied whether the plagiarism identification process requires changes due to the features of GPT-3. Similarly, Fyfe (2022) [13] debated the concept and study of plagiarism, tests with GPT's earlier iteration GPT-2, and recommends that academic users manipulate writing using text editors. The efficiency issue of AI tools (such as GPT-3) is addressed by Anson & Straune (2022) [14]. It presents guidelines on how faculty can meet the challenges of their availability to users [15]. Köbis & Mossink [16] analyzed GPT-2, with users finding it difficult to identify GPT-2-created poetry reliably. Some researchers [17] are positive about the learning features of GPT-3 in the virtual learning process. Moore et al.'s (2022) [18] addressed the science courses and the user access pattern, and GPT-3 was used to assess the quality. [19] A study found that GPT-3 is an essential cognitive and semantic tool for writing as it feeds optimum writing input. Nguyen et al. 2022 [20] detect other relevant academic literature.

3. Question-Answering system (Q & A)

Recently many web-based interactive systems use intelligent computational applications such as chatbots and work like humans. The artificial system can respond to user queries immediately with trained answer sets. Users can raise questions, and the system interacts quickly. Large Language Models (LLMs) are neural networks trained in a large volume of text which produce natural language output. The ChatGPT is developed as an Artificial General Intelligence tool to impart human cognitive ability. The ChatGPT ensures output perfection concerning the language, using basic linguistic features. [21, 22, 23, 24]. The recurrent neural network (RNN) was initially created as the state-of-the-art algorithm for generating answers to user queries, followed by the generative pre-trained transformer. (GPT) [25]. Natural Language Generation techniques have been studied mainly in conversational agents [26,27,28]. Conversational agents, or chatbots, engage users through natural language with algorithms [29].

The regenerative neural networks generate the question-answer (QA) pairs in the given contexts more efficiently than the non-deep learning methods. [30]. The AI-based tools use transformer-based deep learning models, signal the functions and tasks in a series of NLP processes, and perform question-answering reading comprehension and summarisation [31]. In a study, Zhang identified that they could outperform in NLP tasks like humans. [32]. GPT-3 recorded far-reaching performance and reflected imminent growth over its predecessor GPT-2, so it is essential to assess its effectiveness in question-answering.

4. Dataset

Randomly, we have generated 27 queries that seek descriptions and answers to questions. We ensure that the selected prompts are fed into ChatGPT. These questions are provided in the ChatGPT as Prompts, and these prompts generate answers in the ChatGPT.

The answers generated by both the GPT and Scientific Reports are now analysed using various NLP parameters to assess the effectiveness of the ChatGPT.

5. Parameters

Plagiarism/originality of the responses from ChatGPT- The Normal and AI similarity are checked by plagiarism detection tools.

Text editors' language analysis- The Grammarly text editor analyses the text.

Human Intervention- We assess the questions and answers for their content. During the text analysis, we fix the levels as below.

Domain: General
Intent: Describe
Audience: Knowledgeable

We do not analyse the comprehensiveness of the answers this process requires an additional answer system and question generation address one point of view of the issue.

6. Analysis

We have provided input into ChatGPT using randomly created factual, interpretative, evaluative and rhetorical questions. These questions are generated using the recent news published by Nature, British Medical Journal and Science. The answers extracted from the ChatGPT are subjected to a few analyses.

The answers yielded by ChatGPT are subjected to similarity measures. First, we tested the normal similarity followed by the AI similarity, which is currently added by

S. No	Prompt	Normal Similarity	AI similarity
1	Fair vocabularies	0	100
2	Data privacy laws	31	100
3	Brain-spine interface	9	27
4	Baby mice hearts	0	38
5	Quantum computers offer large gains	54	31
6	How data is used to combat misinformation	0	100
7	Health for All at the Centre of our Economies and New Economic Thinking	3	100
8	Designing, conducting, and reporting control interventions to establish a quality standard in non-pharmacological intervention research	9	68
9	Bacterial meningitis in children	10	100
10	Multispecific multiantibody	11	100
11	Cytolytic CD8+ T cells infiltrate germinal centers to limit ongoing HIV replication in spontaneous controller lymph nodes	23	100
12	Protocatechuic acid boosts continual efferocytosis in macrophages by derepressing KLF4 to transcriptionally activate MerTK	23	100
13	Respiration organizes gamma synchrony in the prefronto-thalamic network	7	100
14	Nanoscale thermal control of a single living cell enabled by diamond heater-thermometer	8	100
15	Prevalence of Brucella endocarditis	0	95

16	Global Shrinkage of Space and the Hub-and-Spoke System in the Global Trade Network	21	100
17	Deep-sea mining might rely on flawed data	0	100
18	Why chronic stress upsets the gut	10	100
19	What do you know about citation padding	0	NA
20	Neglected Tropical diseases	18	NA
21	Evaluating Knowledge graphs	0	35
22	Antidepressants induce mutation	23	NA
23	Science is less disruptive	0	NA
24	Why Himalayan Town is shrinking	6	100
25	Introducing methodological review	9	100
26	Carbon dioxide removal	44	NA
27	Science becomes less disruptive	0	100

• ***Note:*** *For a few prompts, the text editors failed to provide similarity scores*

Table 1. Similarity measures of ChatGPT answers using normal and AI tests in text editors

many text editors and plagiarism detection tools. Table 1 lists the results for the queried prompts. The standard similarity scores range from 0 to 54. Thus, checking the ChatGPT-generated text in text editors or plagiarism detection tools may not detect the data trained in the AI. AI is used for both answer provision for users and validating the originality of the texts generated.

6.1. Text Analysis

The texts generated by the AI tool are analysed for their linguistic validity using the measures such as Wordiness Instances, Rephrasing and Rearranging requirements, Text Inconsistency Occurrence, Engagement, Unclear Sentences, Determiners and Overall Text Scores. Grammatical perfection is called in language and presentation and is essential for communication. Pointless mistakes, such as improper verbs, tenses, or sentence fragments, lead English look unprofessional and ill-conceived. It is, therefore, significant for us to measure the retrieval quality in terms of language rules to ensure text consistency.

6.2. Parameters for Assessment

6.2.1. Wordiness

Limiting the word clutters, eliminating redundancy, and reducing unnecessary qualifiers, weak verbs, and round about expressions are essential in communication. The assessment of them will ensure comprehension and ensure to focus on important content.

6.2.2. Rephrasing and rearranging requirements

Many text editors suggest rephrasing to eliminate inconcision, and we tested it in the ChatGPT retrieval. Each retrieval set tests the text for the need for rephrasing, and the instances are counted.

6..2.3. Engagement

The presentation's first few lines help the reader decide whether to keep reading or leave with information consumption. We now test the retrievals for engagement issues.

6.2.4. Unclear Sentences

Clear sentences lead us to understand and interpret easily. Writing clearly and concisely helps move the text deliberately, construct carefully and use grammar correctly. We use unclear sentences as one parameter in this analysis.

6.2.5. Determiners

A determiner is a word that modifies, describes, or introduces a noun with possessives and demonstratives. Thus, each prompt and its retrieval content are tested using the above-described parameter and the results are presented in Table 2. Table 2 presents the results of the various text

S.No	Prompt	Overall Text Score	Wordiness Instances	Rephrasing and rearranging	Text inconsistency occurrence	Engagement	Unclear sentences	Determiners
1.	Fair Vocabularies	82%	5	1	0	1	1	0
2.	Data Privacy	94%	3	4	0	1	3	0
3.	Brian-spine interface	95%/	3	2	0	0	1	0
4.	Babymice heart	96%/	1	0	0	1	2	2
5.	Quantum computers	95%/	2	1	0	3	0	1
6.	How to combat data	83%	5	2	0	1	2	0
7.	Health for all	90%	7	1	0	1	4	0
8.	Designing control inventions	86%	6	2	0	5	1	0
9.	Bacterial meningitis	92%	1	0	0	2	1	1
10.	Multispecific multiantibody	73%	4	4	0	2	4	0
11.	Cytolytic CD8+ T cells infi	66%	7	1	0	1	1	0
12.	Protocatechuic acid boosts continual efferocytosis	84%	1	2	0	1	1	3
13.	Respiration organizes	77%	2	2	0	3	2	0
14.	Nanoscale thermal control	84%	5	2	0	1	2	1
15.	Prevalence of Brucella endocarditis	97%	1	0	0	2	1	0
16.	Global Shrinkage of Space	88%	4	0	0	4	2	0
17.	Deep-sea mining	88%	1	1	0	3	3	0
18	Why chronic stress upsets the gut	86%	1	1	0	3	3	1
19	Citation padding	96%	1	0	0	2	1	0
20.	Neglected tropical diseases 2023	86%	2	1	0	0	0	0
21.	Evaluating Knowledge Gaps in Sea-Level Rise	79%	2	1	0	3	6	3
22.	Antidepressants induce mutation	98%	2	1	0	1	1	0
23.	Science is less disruptive	87%	2	0	0	1	1	0
24.	Why Himalayan town is sinking	82%	6	1	0	3	2	0
25.	Introducing methodological review	83%	5	1	0	4	2	0
26.	Carbon dioxide removal 2022	87%	3	1	0	1	3	1
27.	Science becomes less disruptive	85%	2	2	0	5	3	0

Table 2. Text Analysis Results

analysis made from the answers of the ChatGPT. The Overall Text Scores range from 77 to 96%, leaving different values. The overall text scores are arrived in text editors based on all the text examined in their platform. Thus, it is a relative one based on the examined scores online by many users. We use the text editor platform to arrive at the Overall Text Scores for the ChatGPT prompts.

The wordiness instances are significant, considering the small text size in the answers, while the rephrasing suggestions are moderate. The text editor has not reported any text consistency for all our prompts.

The other three parameters, Engagement, Unclear Sentences and Determiners, are moderate in the analysed text. We then tested the relation between various parameters used for text testing. The results are produced below in the figures.

6.3. Relationship measure among variables in Text analysis

In the analysis below, we correlated the combination of two variables. The text's Overall Text Score and Wordiness match; Figure 1 below shows the relations. A high-level correlation is observed, which indicates the text's

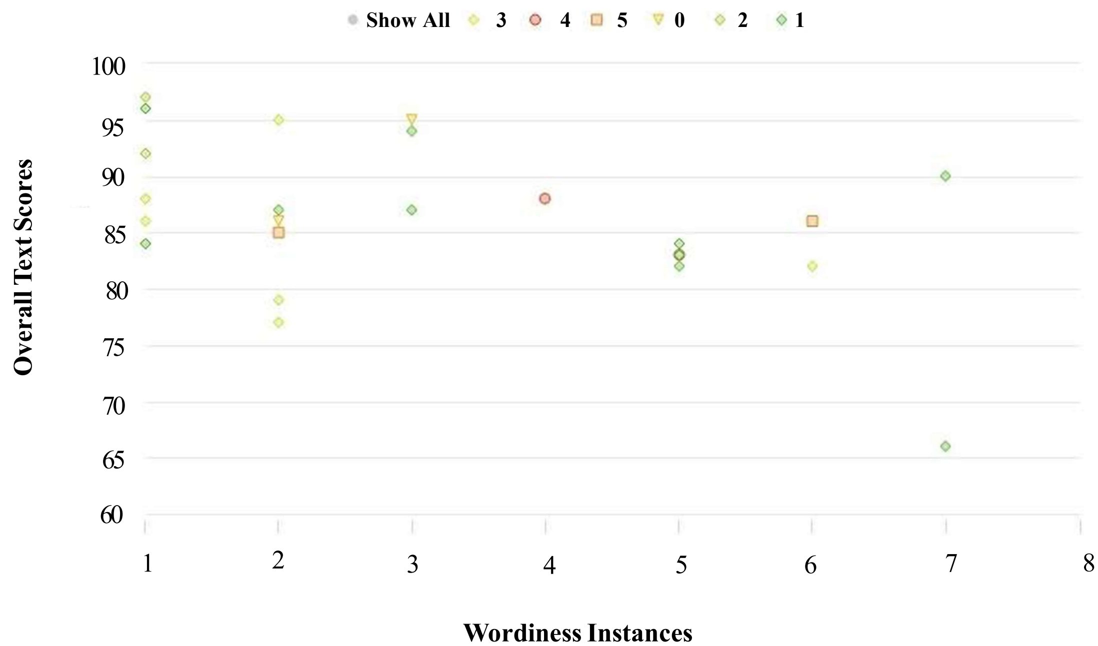

Figure 1. Relationship between Overall Text Scores and Wordiness

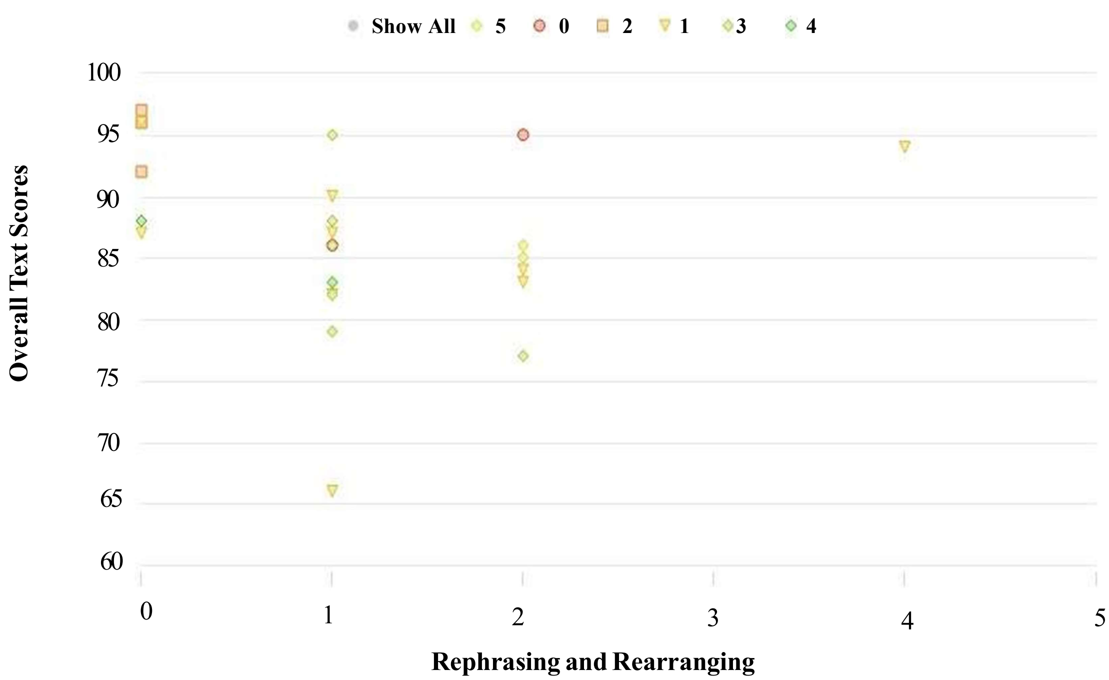

Figure 2. Relationship between Overall Text Scores and rephrasing

overall score is high when wordiness is less. The relation between Overall text and rephrasing requirement is moderate. Otherwise, the rephrasing requirement only impacts overall text scores a little.

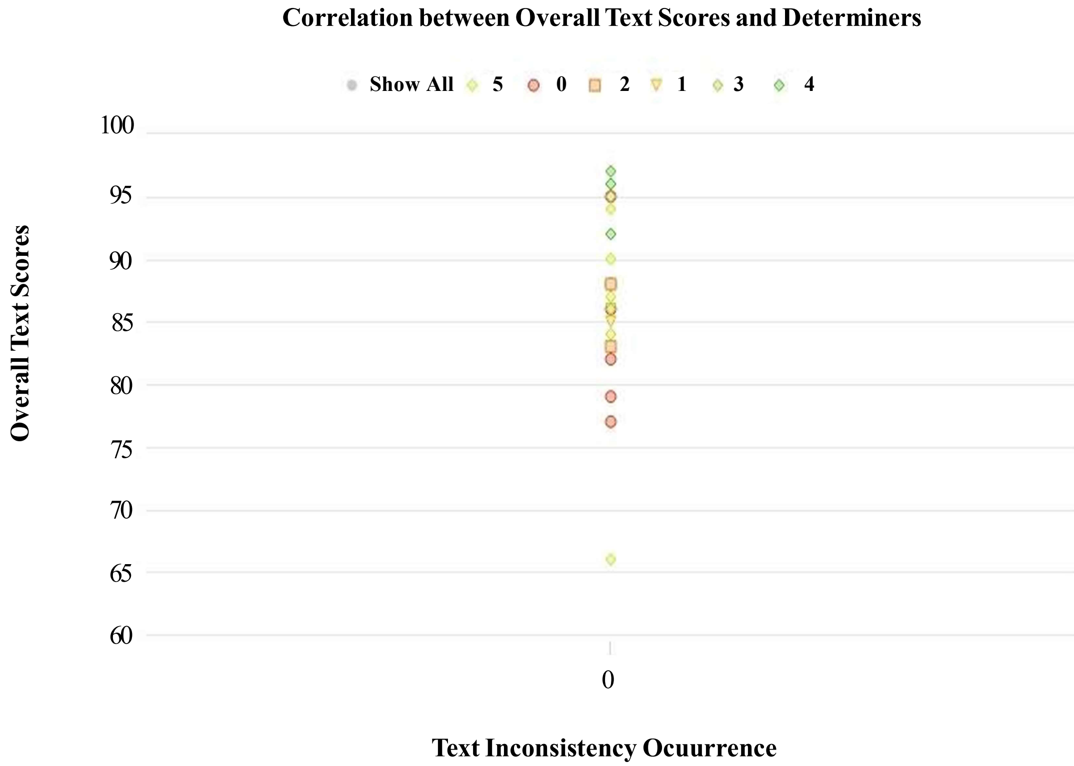

Figure 3. Relationship between Overall Text Scores and Text Inconsistency

Figure 3 has no result as the text editor does not report text inconsistency during analysis. Thus, no inference is generated in this relationship.

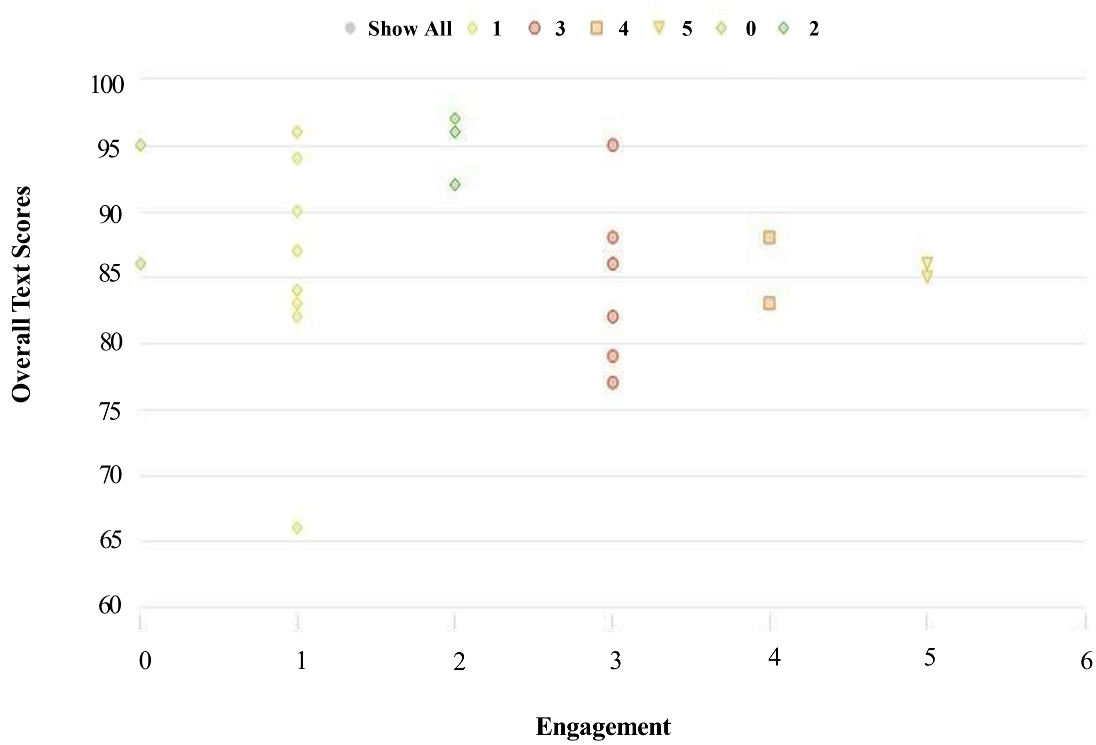

Figure 4. Relationship between Overall Text Scores and Engagement

The impact of engagement on the Overall text score is moderate, as the overall relation shows positive results.

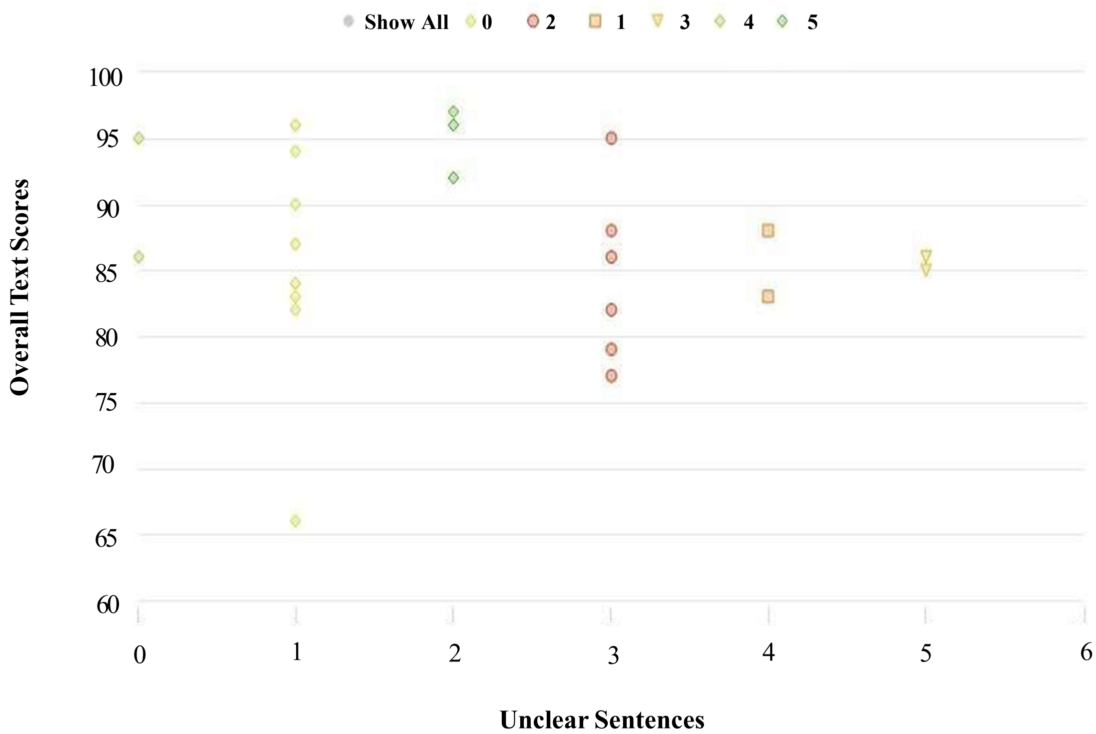

Figure 5. Relationship between Overall Text Scores and Unclear Sentences

The plotting of the points is scattered as we do not find the plots over the median line in the graph. It confirms the less impact of the unclear sentence occurrence over the Overall Text Scores.

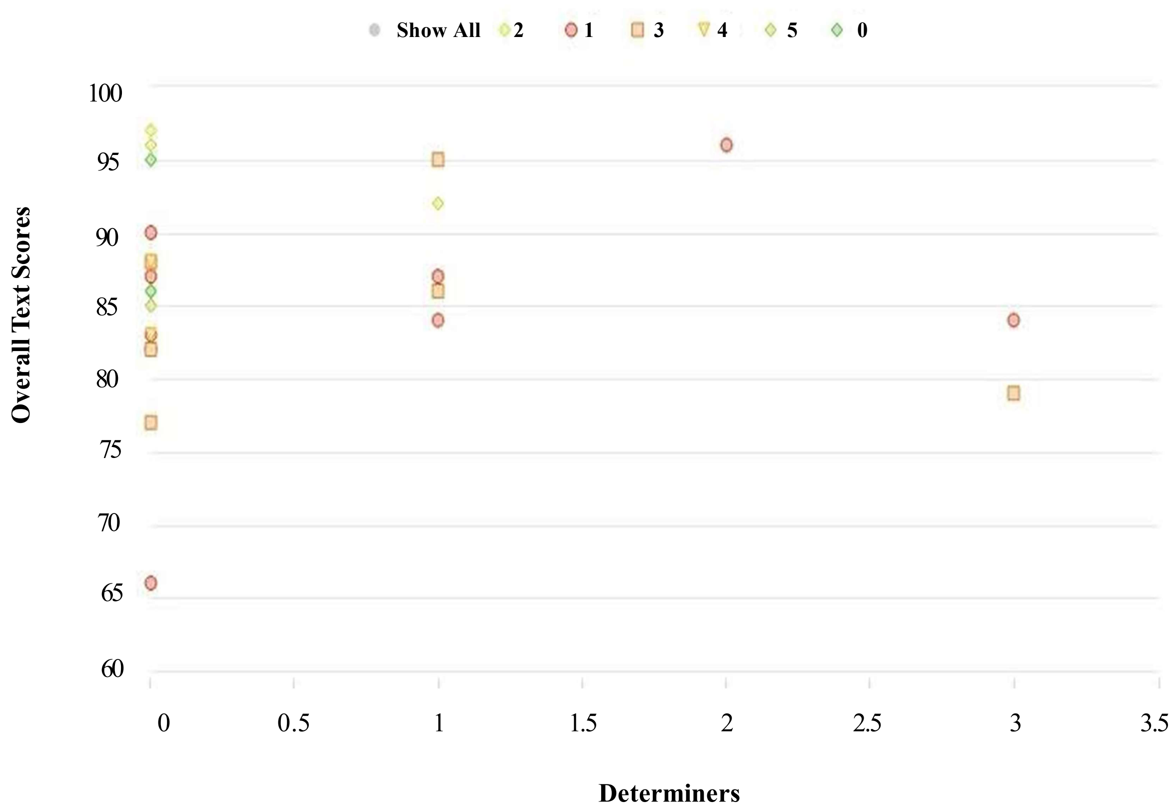

Figure 6. Relationship between Overall Text Scores and Determiners

The determines do not impact the Overall Text Scores, and the plotted points lie above the possible median line.

6.3.1 Correlation among multiple variables

The deployment of a few selected variables and their data is subjected to a multi-correlation analysis, and the result is presented in Figure 7.

The x-axis is used to represent seven variables, and the Y-axis is used to describe the overall text scores. Figure 7 illustrates the overall general relationship results, showing varying relations among the variables used in the text analysis.

Writing and presenting the text is difficult for many people, and correcting the language and presentation is tough. Users, particularly non-native users, need help transforming wordy sentences into more concise ones and ensuring error-free sentences. AI tools are expected to infuse linguistic constructs into their platform. However, the above analyses found that AI tools must provide error-free content in answering questions. AI-based tools should incorporate more features to bring successful language models.

For a few prompts, the ChatGPT replies that data was trained till 2021 only, and the updated content is not generated. During the manual analysis of the answers, we also found that the content is more elementary, and the critical part is omitted. The efficiency of ChatGPT is behind human expertise, particularly in perception and comprehension.

The success of ChatGPT depends on the training dataset. The dataset has yet to incorporate many newer research reports and relies on open content on the web.

Query refinement is required in many places as ChatGPT intelligence needs to understand user perception. A query on '*Science is less disruptive*' leads to a denial response from ChatGPT. This phrase is drawn from the Nature report where ChatGPT fails to recognise and respond to it negatively. When an advanced query, '*Science becomes less disruptive,'* it can sense and respond. In the question-answering systems, the user queries are modified with human intervention, which is inattentive in AI-based systems.

The texts generated from ChatGPT fail to ensure consistency in the answers to the questions raised, confirming that the training needs more precision and the call for more variables to make the AI conversations more oriented to language and language models impact the user community.

7. Issues in Chat GPT

ChatGPT is found to have fabricated the text with several factual errors, misrepresentations and erroneous data. These errors are reported mainly due to the absence of relevant and significant content in ChatGPT's training set, a failure to filter the correct data or difficulty distinguishing between credible and less-credible sources [33]. Similar biases, such as availability, selection and confirm

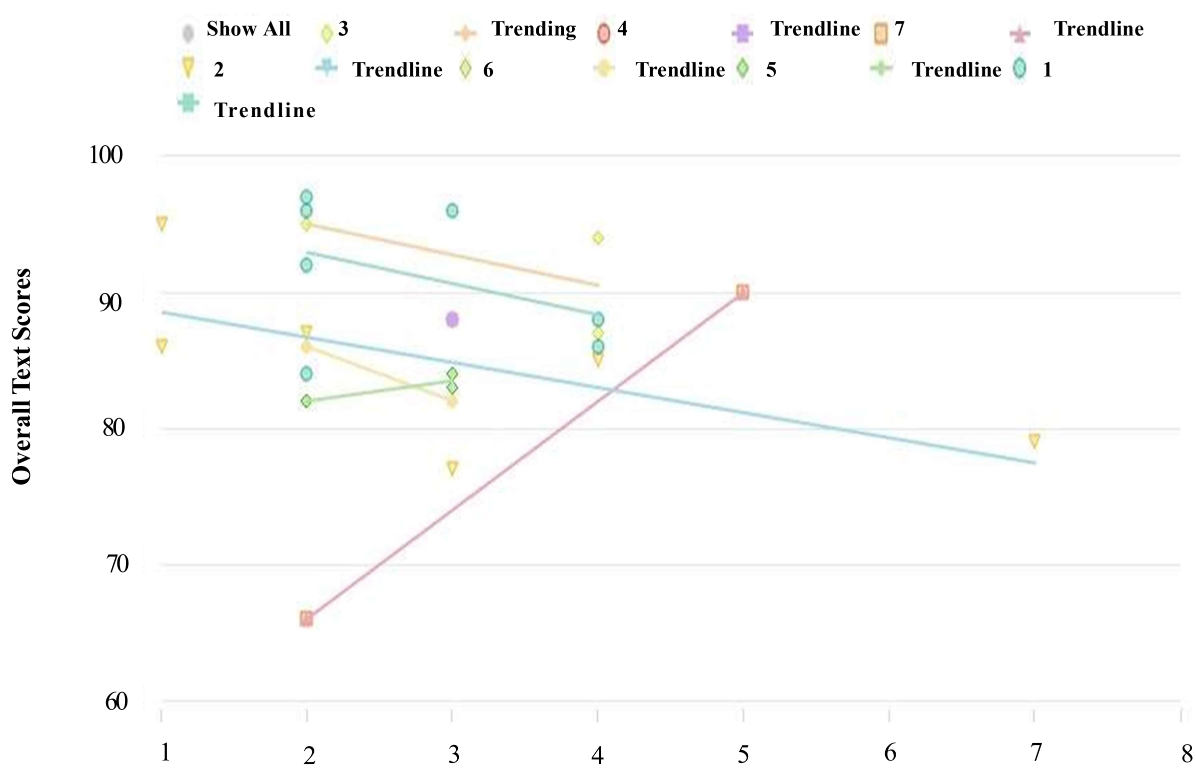

Figure 7. Relationship between all the variables used for text analysis

Learntion discrepancies, are reported and frequently amplified in conversational AI [34].

The ChatGPT-generated summaries are subjected to the plagiarism checker tools: The median similarity was 0%, indicating no plagiarism was identified. The AI-output similarity was determined as 66% of the produced abstracts. But the expert scanning could have done better: they correctly identified only two-thirds of the produced abstracts and 86% of the genuine abstracts. They wrongly detected nearly one-third of the produced abstracts as being accurate and 14% of the natural summaries as being developed. [35]

Of late, the scientific information world faces a challenge induced by AI technology's contravention of its scientific values, practices and norms. Science can move to utilise the developments and, at the same time, manage the risks. The scientific world has the potential to make use of AI in a meaningful and exceptional way without affecting the fundamental function of science and promoting curiosity, imagination and discovery [36].

8. Summary

We measured the similarity, comparison of results with text editors and a few more parameters. It may be helpful to write primary content, whereas, in research, this AI tool cannot perform well. Understanding the contextual relations between concepts is possible only by human intervention. It cannot make sensible interference from logical relations between scientific concepts. Based on what is already produced in the literature is combined by machines.

Content is primarily drawn from web-based sources; it is paraphrasing as the content is not original and pre-trained. It affects creativity even if drawn from many sources; the sources must be acknowledged. If the chat GPT system is honest, the references where the information was extracted should be acknowledged.

9. Conclusion

The ChatGPT uses reinforcement learning, deep learning, supervised learning, and a large corpus dataset to answer user questions with predictive text effectively. [37] Unlike other QA systems, the primary dataset used to respond to questions needs to be clarified as, except Wikipedia, all others are crawled from many sources. The comprehensiveness of them is debatable. One issue is the provision of current information by the ChatGPT. The last trained dataset date back to 2021. Text generation needs a so-called quality-diversity trade-off problem [38]. ChatGPT, like an intelligent plagiarist, denies knowing the original text used to generate answers. It may have a long-term negative impact on human synthesis in learning and understanding. If AI succeeds ultimately, the machines may take over the role of humans, and humans work like machines with limited logic and creativity.

The inevitable use of LL Models in different domains and the academic world is open, and users largely accept to practice it. The experts' studies and opinions should be considered while using it. It is essential to follow the ethics, guidelines and regulations, and the engagement of all stakeholders is required to ensure the responsible use of ChatGPT powers. [39].

Funding: This work is not supported by any grants.

Conflict of Interest: None

References

[1] https://fireflies.ai/blog/generative-ai-or-gpt-3-apps

[2] Number of parameters in notable Artificial Intelligence Systems, Our World in Data. [Online]. Available: https://ourworldindata.org/grapher/artificial-intelligence-parametercount. [Accessed: 22-Mar-2023].

[3] Hughes, A. (2023). CHATGPT: Everything you need to know about OpenAI's GPT-4 tool, ChatGPT: Everything you need to know about OpenAI's GPT-4 upgrade — BBC Science, Focus Magazine, 16-Mar-2023. [Online]. Available: https://www.sciencefocus.com/futuretechnology/gpt-3 [Accessed: 22-Mar-2023].

[4] https://towardsdatascience.com/gpt-3-a-complete-overview-190232eb25fd

[5] Yu, Caiyang., Liu, Xianggen., Tang, Chenwei., Feng, Wentao., Jiancheng GPT-NAS: Neural Architecture Search with the Generative Pre-Trained Model. Lv.arXiv:2305.05351v1 [cs.CV] 9 May 2023]

[6] Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., Awadalla, H. H. (2023). How good are gpt models at machine translation? a comprehensive evaluation, arXiv preprint arXiv:2302.09210, 2023

[7] Mathew L. Bindu, V. (2020). A review of natural language processing techniques for sentiment analysis using pre-trained models, In: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). *IEEE*, 2020, p. 340–345.

[8] Bongini, P., Becattini, F., Del Bimbo, A. (2022). Is gpt-3 all you need for visual question answering in cultural heritage? *In:* Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, *Proceedings, Part I*. Springer, 2023, p. 268–281.

[9] Limna, Pongsakorn., Kraiwanit, Tanpat., Jangjarat, Kris., KlayklungPrapasiri, Chocksathaporn, Piyawatjana (2023). The use of ChatGPT in the digital era: Perspectives on chatbot implementation, *Journal of Applied Learn-*

-*ing & Teaching*. Vol.6 (1) 1-22.

[10] Cooper, K. (2021, November 1). OpenAI GPT-3: Everything you need to know. Springboard, https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai

[11] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.

[12] Dehouche, N. (2021). Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). *Ethics in Science and Environmental Politics*, 21, 17-23

[13] Fyfe, P. (2022). How to cheat on your final paper: *Assigning AI for student writing. AI & Society*, 1-11.

[14] Anson, C. M., Straume, I. (2022). Amazement and trepidation: Implications of AI-based natural language production for the teaching of writing. *Journal of Academic Writing*, 12 (1) 1-9.

[15] Anson, C. M. (2022). *AI-based text generation and the social construction of "fraudulent authorship": A revisitation. Composition Studies,* 50, 37-46.

[16] Köbis, N., Mossink, L. D. (2021). *Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. Computers in Human Behavior,* 114, 106553

[17] Tate, T., Doroudi, S., Ritchie, D., Xu, Y. (2023). Educational research and AI-generated writing: Confronting the coming tsunami. Pre-print. https://edarxiv.org/4mec3

[18] Moore, S., Nguyen, H. A., Bier, N., Domadia, T., Stamper, J. (2022). Assessing the quality of student-generated short answer questions using GPT-3. In: *European Conference on Technology Enhanced Learning* (p. 243-257). Springer.

[19] Elkins, K., Chun, J. (2020). Can GPT-3 pass a writer's Turing test? *Journal of Cultural Analytics*, 5 (2) 17212.

[20] Nguyen, H. A., Bhat, S., Moore, S., Bier, N., Stamper, J. (2022). Towards generalized methods for automatic question generation in educational domains. *In: European Conference on Technology Enhanced Learning* p. 272-284. Springer.

[21] Yu, Caiyang., Liu, Xianggen., Tang, Chenwei., Feng, Wentao., Lv, Jiancheng.(GPT-NAS: Neural Architecture Search with the Generative Pre-Trained Model.arXiv:2305.05351v1 [cs.CV] 9 May 2023

[22] Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M. Awadalla, H. H (2023). How good are GPT models at machine translation? a comprehensive evaluation, arXiv preprint arXiv:2302.09210

[23] Mathew, L., Bindu, V (2020). A review of natural language processing techniques for sentiment analysis using pre-trained models, In: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC). *IEEE*, 2020, p. 340–345.

[24] Bongini, P., Becattini, F. Del Bimbo, A (2022). Is gpt-3 all you need for visual question answering in cultural heritage? In: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I. Springer, 2023, p. 268–281.

[25] Li, Chenglu., Xing, Wanli. (2021). Natural Language Generation Using Deep Learning to Support MOOC Learners. *International Journal of Artificial Intelligence in Education* v. 31, p.186–214.

[26] Caballé, S., Conesa, J. (2019). Conversational Agents in Support for Collaborative Learning in MOOCs: An Analytical Review. In F. Xhafa, L. Barolli, & M. Greguš (Eds.), Advances in Intelligent Networking and Collaborative Systems. INCoS 2018. *Lecture notes on data engineering and communications technologies* (Vol. 23). Cham: Springer.

[27] Kumar, R., Rose, C. P. (2010). Architecture for building conversational agents that support collaborative learning. *IEEE Transactions on Learning Technologies*, 4 (1) 21–34. https://doi.org/10.1109/TLT.2010.41

[28] Mittal, A., Vigentini, L., Djatmiko, M., Prusty, G., Sharma, Y., King, M. E. (2018). MOOC-O-Bot: Using Cognitive Technologies to Extend Knowledge Support in MOOCs. *Proceedings of the 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 69–76. https://doi.org/10.1109/tale.2018.8615453.]

[29] Shawar and Atwell Abu Shawar, B., Atwell, E. (2007). *Chatbots: Are They Really Useful? LDV-Forum Band*, 22, 29-49.

[30] Indurthi, S. R., Raghu, D., Khapra, M. M., Joshi, S. (2017). Generating natural language question-answer pairs from a knowledge graph using a RNN-based question generation model. In: M. Lapata, P. Blunsom, A. Koller (Eds), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers (pp. 376-385). *Association for Computational Linguistics*. https://doi.org/10.18653/v1/e17-1036.

[31] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1 (8) 9. Retrieved 4 Jan 2020, from http://www.persagen.com/files/misc/

radford2019language.pdf

[32] Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Dolan, B. (2020). DialoGPT: Large-scale generative pre-training for conversational response generation. *In*: A. Celikyilmaz, T. Wen (Eds), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (p. 270–278). *Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-demos.30

[33] Eva A. M., van Dis, Bollen, Johan., Rooij, Robert van., Zuidema, Willem., Bockting, Claudi L. (2023). ChatGPT: five priorities for research, *Nature* V. 614. 224-2266. 9 February 2023.

[34] Rich, A. S., Gureckis, T. M.(2019). *Nature Mach. Intell*. 1, 174–180. 2019.

[35] Else, Holly., Gao, C. A. et al (2022). ABSTRACTS WRITTEN BY CHATGPT FOOL SCIENTISTS by. *Nature* V. 613, 19 January 2023 p. 423. Preprint at bioRxiv https://doi.org/10.1101/2022.12.23.521610.

[36] ChatGPT: five priorities for research by Eva A. M. van Dis, Johan Bollen, Robert van Rooij, Willem Zuidema & Claudi L. Bockting- 224-2266, *Nature*, Vol 614 . 9 February 2023.

[37] JOUR Kenney, Noah A Brief Analysis of the Architecture, Limitations, and Impacts of ChatGPT DO - 10.5281/zenodo.7762245 2023/03/23

[38] Zhu, Yaoming., Lu, Sidi., Zheng, Lei., Guo, Jiaxian., Zhang, Weinan., Wang, Jun ., Yu, Yong. (2018). Texygen: A Benchmarking Platform for Text Generation Models. *ACM SIGIR*'18, July 8-12, 2018, p. 1097-1100. Ann Arbor, MI, USA

[39] Sallam, Malik. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns, *Healthcare*, 11 (6) 887; https://doi.org/10.3390/healthcare11060887