ChatGPT Generated Content and Similarity Index in Chemistry & Allied Sciences

Deep Kumar Kirtania

Librarian, Bankura Sammilani College

deepkrlis@gmail.com

Abstract: The main objective of this study is to verify similarity index of ChatGPT generated content in the field of chemistry and its allied subjects. To complete this study twenty sub subjects of chemistry based on controlled vocabulary tools such as Dewey Decimal Classification (DDC) system, Sears List of Subject Headings and Library of Congress Subject Headings (LCSH) have considered for sample, followed by content generation and similarity check using iThenticate, Urkund and Turnitin. The percentage of matching paragraphs is relatively low as the three plagiarism software shows 12%, 1% and 5% respectively.

Keywords: OpenAI, ChatGPT, Similarity Index, Plagiarism, Chemistry.

Introduction: In recent years, the advancements in natural language processing and machine learning have led to the development of powerful language models like ChatGPT[1]. These models, based on the GPT-3.5 architecture, are designed to understand and generate human-like text responses*2*. While these models have been extensively used for various applications, their potential in the domain of chemistry and allied sciences remains largely unexplored[3]. By leveraging the vast knowledge and data available in the field, ChatGPT has the potential to assist researchers, students, and professionals in accessing relevant information, solving problems, and facilitating scientific communication[4]. ChatGPT has the potential to revolutionize the way we access and interact with scientific knowledge in the field of chemistry and allied sciences. The generated content can encompass a wide range of topics, such as organic chemistry, inorganic chemistry, analytical chemistry, physical chemistry, biochemistry, and other related areas. Several papers[5-9] have already been published on chemistry and ChatGPT such as drug discovery, teaching learning, computational chemistry etc. From these research papers many things are known about ChatGPT and chemistry but the present work focuses only on Content and similarity index in chemistry and allied sciences

The aim of this study is to investigate the capabilities of ChatGPT in generating content related to chemistry and allied sciences and to checking the similarity index to evaluate the quality and accuracy of the generated responses.

Scope and coverage: This study covers twenty sub subjects related to Chemistry such as Agricultural chemistry, Analytical chemistry, Atmospheric chemistry, Biochemistry, Botanical chemistry, Clinical chemistry, Crystallography, Industrial chemistry, Inorganic chemistry, Microchemistry, Mineralogy, Organic chemistry, Pharmaceutical chemistry, Photochemistry,

Physical chemistry, Radiation chemistry, Space chemistry, Spectrum analysis, Textile chemistry, Theoretical chemistry.

Method used: First Chemistry and its related subjects are selected through three standard and globally renowned controlled vocabulary tools such as Dewey Decimal Classification (DDC)[10] system, Sears List of Subject Headings[11] and Library of Congress Subject Headings (LCSH)[12]. With these tools, excluding repeated or common subjects, twenty larger subjects, selected for the present work are AB and c respectively. The next step is to capture each of these terms and generate content through ChatGPT and check that content through three major plagiarism software for finding the similarity index. Finally, the similarity checking or plagiarism reports are analyzed through tables and findings and conclusions are drawn.

Limitations: Some limitations of the present work are noted which are:

I. Researchers or users have no control over the content generated by ChatGPT, so this work has been done with the answers given by ChatGPT based on the specific query.
II. While checking plagiarism or similarity index, all three software have to be relied upon and the researcher has no control over their own detection mechanism.
III. Besides, 20 subjects and 168 paragraphs have been generated for this work, but if such work is done with a larger number of samples, better results will be obtained.

Results:

I. iThenticate indicates a similarity index of 12%, suggesting that approximately 12% of the content generated by ChatGPT has similarities with existing sources. Out of the 168 paragraphs checked, 75 were found to have matches, while 93 paragraphs did not have any matching content.
II. Urkund reveals a lower similarity index of 1%, indicating minimal similarities between the ChatGPT generated content and existing sources. Only 23 out of the 168 paragraphs were found to have matches, while the majority of the paragraphs (145) did not show any matching content.
III. Turnitin reports a similarity index of 5%, indicating moderate similarities between the ChatGPT generated content and available sources. Out of the 168 paragraphs, 37 were identified as matching content, while 131 paragraphs did not exhibit any similarities.
IV. Subject-wise analysis (Vide. Table 2) provides insights into the similarity between the ChatGPT generated content and existing sources across different areas of chemistry.

The results obtained from these plagiarism checker software indicate varying degrees of similarity between the ChatGPT generated content and existing sources. iThenticate and Turnitin demonstrate higher similarity indices compared to Urkund. It is important to note that these percentages only provide a quantitative measure of similarity and further analysis is required to determine the nature and context of the matches.

Table 1: Similarity Index of ChatGPT Generated Content using various software

Software Name	Similarity Index	Total number of Paragraph	Similarity Matching paragraph	Not matching paragraph
iThenticate	12%	168	75	93
Urkund	1%	168	23	145
Turnitin	5%	168	37	131

Table 2: Subject wise distribution of Similarity Index of ChatGPT Generated Content

Subject	Total Number of Paragraph	Similarity Matching paragraph					
		iThenticate		Urkund		Turnitin	
		Matching	Not Matching	Matching	Not Matching	Matching	Not Matching
Agricultural chemistry	8	1	7	0	8	1	7
Analytical chemistry	8	5	4	0	8	2	6
Atmospheric chemistry	7	5	2	2	5	2	5
Biochemistry	8	1	7	3	5	2	6
Botanical chemistry	9	2	7	1	8	2	7
Clinical chemistry	8	1	7	3	5	0	8
Crystallography	8	6	2	4	4	2	6
Industrial chemistry	9	2	7	1	8	1	8
Inorganic chemistry	8	6	2	0	8	2	6
Microchemistry	9	2	7	0	9	0	9
Mineralogy	9	3	6	0	9	1	8
Organic chemistry	8	7	1	1	7	2	6
Pharmaceutical chemistry	8	4	4	1	7	1	7
Photochemistry	10	3	7	1	9	2	8
Physical chemistry	9	5	4	2	7	3	6
Radiation chemistry	8	5	3	1	7	3	5
Space chemistry	7	3	4	0	7	2	5
Spectrum analysis	11	7	4	3	8	4	7
Textile chemistry	8	3	5	0	8	3	5
Theoretical chemistry	8	4	4	0	8	2	6
Total	**168**	**75**	**93**	**23**	**145**	**37**	**131**

Major Findings:

I. The similarity index varies across different subjects in chemistry. Some subjects show a higher similarity index, indicating more matching content, while others exhibit lower similarity, indicating less similarity with existing sources.
II. Among the subjects analyzed, organic chemistry shows the highest similarity index across all three plagiarism checker software, with iThenticate reporting 7 matching paragraphs, Urkund reporting 1 matching paragraph, and Turnitin reporting 2 matching paragraphs.
III. Microchemistry, mineralogy, and theoretical chemistry show the lowest similarity index across all three-plagiarism checker software, with either no matching paragraphs or a minimal number of matching paragraphs.
IV. There is variation in the results obtained from different plagiarism checker software. For example, iThenticate often reports a higher number of matching paragraphs compared to Urkund and Turnitin for many subjects.
V. Some subjects exhibit inconsistencies in matching paragraphs across different plagiarism detectingtool. For instance, inorganic chemistry shows a discrepancy in results, with iThenticate and Turnitin reporting 6 matching paragraphs while Urkund does not find any matching paragraphs.
VI. Overall, the percentage of matching paragraphs is relatively low for all subjects, indicating that the ChatGPT generated content generally does not have extensive similarities with existing sources.

Conclusion: In conclusion, the analysis of the similarity index of ChatGPT generated content, subject-wise, using various plagiarism checker software reveals several important findings. The ChatGPT generated content demonstrates a relatively low level of similarity with existing sources across different subjects in chemistry and majority of the paragraphs do not exhibit significant matches.The similarity index varies among different subjects. There are inconsistencies in the results obtained from different plagiarism checker software. While the similarity index provides a quantitative measure of similarity, it is crucial to conduct manual examination and contextual analysis of the matching paragraphs to determine the appropriateness, originality, and accuracy of the ChatGPT generated content.Plagiarism checker software should be used as complementary tools alongside manual assessment to make informed judgments about the quality and originality of the content generated by ChatGPT.Based on these findings, it can be concluded that ChatGPT, when trained on a comprehensive dataset of chemistry and allied sciences, has the potential to generate content with limited similarity to existing sources. Future research and improvements in NLP models can further enhance the reliability and credibility of the generated content, expanding the possibilities for scientific communication and knowledge dissemination in the field of chemistry and allied sciences.

References:

(1) Castro Nascimento, C. M.; Pimentel, A. S. Do Large Language Models Understand Chemistry? A Conversation with ChatGPT. *J. of Chem. Inf. and Model.* **2023**, *63*(6), 1649-1655.

(2) Kirtania, D. K.,; Patra, S. K.OpenAI ChatGPT Generated Content and Similarity Index: A study of selected terms from the Library & Information Science. *Annals of Lib. & Inf. Studies*. **2023,** *70*(2), 99-101.

(3) Pimentel, A.; Wagener, A.; da Silveira, E. F.; Picciani, P.; Salles, B.; Follmer, C.; Oliveira Jr, O. N. Challenging ChatGPT with Chemistry-Related Subjects. **2023**, https://chemrxiv.org/engage/api-gateway/chemrxiv/assets/orp/resource/item/646d22f3b3dd6a65309f5119/original/challenging-chat-gpt-with-chemistry-related-subjects.pdf

(4) Emenike, M. E.; Emenike, B. U. (2023). Was This Title Generated by ChatGPT? Considerations for Artificial Intelligence Text-Generation Software Programs for Chemists and Chemistry Educators. *J. of Chem. Ed.* **2023**, *100*(4), 1413-1418.

(5) Humphry, T.; Fuller, A. L. Potential ChatGPT Use in Undergraduate Chemistry Laboratories. *J. of Chem. Ed.* **2023**, *100*(4), 1434-1436.

(6) Lolinco, A.; Holme, T. Developing a curated chatbot as an exploratory communication tool for chemistry learning. **2023**, https://chemrxiv.org/engage/api-gateway/chemrxiv/assets/orp/resource/item/64934176a2c387fa9ab585e8/original/developing-a-curated-chatbot-as-an-exploratory-communication-tool-for-chemistry-learning.pdf

(7) Mahjour, B.; Hoffstadt, J.; Cernak, T. Designing Chemical Reaction Arrays using phactor and ChatGPT.**2023,**https://chemrxiv.org/engage/api-gateway/chemrxiv/assets/orp/resource/item/63fd0c5d937392db3d25b507/original/designing-chemical-reaction-arrays-using-phactor-and-chat-gpt.pdf

(8) Santos, R. P. D. Enhancing Chemistry Learning with ChatGPT and Bing Chat as Agents to Think With: A Comparative Case Study.**2023**, *arXiv preprint arXiv:2305.11890*.

(9) Sharma, G.; Thakur, A. ChatGPT in drug discovery.**2023**, https://chemrxiv.org/engage/api-gateway/chemrxiv/assets/orp/resource/item/63d56c13ae221ab9b240932f/original/chat-gpt-in-drug-discovery.pdf

(10) Dewey, M. (2011). *Dewey Decimal Classification and Relative Index* (Vol. II). OCLC.

(11) Sears List of Subject Headings. https://searslistofsubjectheadings.com/page/frontmatter

(12) Library of Congress Subject Heading. https://www.loc.gov/aba/cataloging/classification/lcco/lcco_q.pdf