

PAPER

A Proposed Framework for Human-like Language Processing of ChatGPT in Academic Writing

Mohammad

Mahyoob^{1,2}(✉),Jeehaan Algaraady³,Abdulaziz Alblwi¹¹Taibah University,
Medina, KSA²Technical Community
College, Taiz, Yemen³Taiz University, Taiz, Yemenmqassem@taibahu.edu.sa**ABSTRACT**

The study proposed a framework for analyzing and measuring the ChatGPT capabilities as a generic language model. This study aims to examine the capabilities of the emerging technological Artificial Intelligence tool (ChatGPT) in generating effective academic writing. The proposed framework consists of six principles (Relatedness, Adequacy, Limitation, Authenticity, Cognition, and Redundancy) related to Artificial Language Processing which would explore the accuracy and proficiency of this algorithm-generated writing. The researchers used ChatGPT to obtain some academic texts and paragraphs in different genres as responses to some text-based academic queries. A critical analysis of the content of these academic texts was conducted based on the proposed framework principles. The results show that despite ChatGPT's exceptional capabilities, its serious defects are evident, as many issues in academic writing are raised. The major issues include information repetition, nonfactual inferences, illogical reasoning, fake references, hallucination, and lack of pragmatic interpretation. The proposed framework would be a valuable guideline for researchers and practitioners interested in analyzing and evaluating recently emerging machine languages of AI language models.

KEYWORDS

ChatGPT, emerging technologies, OpenAI, LLMs, academic writing, natural and artificial language processing

1 INTRODUCTION

ChatGPT (Generative Pretrained Transformer) architecture is an advanced language model developed and launched by OpenAI, an AI research and deployment company, in November 2022 [1]. It is based on the Transformer architecture, a deep-learning neural network designed for NLP tasks. The model has been trained on a massive amount of text data from the internet (approximately 570 GB) and can generate human-like text responses to a wide range of prompts [2]. ChatGPT was trained on a corpus of 410 billion tokens drawn from a common, largely indiscriminate crawl of the internet and is ten times larger than any previous NLP model [3].

Mahyoob, M., Algaraady, J., Alblwi, A. (2023). A Proposed Framework for Human-like Language Processing of ChatGPT in Academic Writing. *International Journal of Emerging Technologies in Learning (iJET)*, 18(14), pp. 282–293. <https://doi.org/10.3991/ijet.v18i14.41725>

Article submitted 2023-04-25. Resubmitted 2023-06-07. Final acceptance 2023-06-07. Final version published as submitted by the authors.

© 2023 by the authors of this article. Published under CC-BY.

The tool was trained through reinforcement learning with human feedback, including from workers who screen for and label toxic text. This training enables ChatGPT to develop its generated language based on the human evaluators' inputs. As a result, the tool becomes more mature at understanding users' intents, generating human-like language, and maintaining coherence in conversation interactivity; one of the primary differentiating factors of ChatGPT from its predecessors is its multi-turn dialogue interactivity. This enables ChatGPT to perform multiple tasks within a dialogue session.

ChatGPT has several applications in various fields, including academic writing, and it can be used to provide summaries, abstracts, and conclusions. This tool can generate guidelines for academic research based on text-based academic queries and paraphrase academic texts with concise content and structure, helping researchers articulate their ideas and focus their efforts more precisely. ChatGPT produces impressively fluent writings in different genres; this tool learns from the user feedback [4] it receives. ChatGPT has gained unprecedented popularity because it shows impressive abilities in many research fields. An attractive feature of this tool is its ability to recall earlier prompts in the same conversation and allow the users to interact with the underlying large language model. Despite its exceptional capacities, its serious defects are also evident. However, ChatGPT is better than its predecessor, InstructGPT, in reducing deceitful responses, and gives more authentic responses [5]. If the researchers used ChatGPT without focusing on the content of the literature, they would be misled into ideas and details that do not exist in the same source or authors.

Moreover, this tool may output biased or nonsensical text with an authoritative tone, as [6] and [7] indicate. This algorithm's knowledge base and natural language processing capabilities are advanced and vast due to applying big data and artificial intelligence. However, text analytics specialists can identify the generated content if they have a comprehensive experience in text analysis [8] and [9]. Recently, human-machine competition in various industries has peaked, especially in language processing, which has led to the advent of many tools and their introduction to the broader public, such as spell checkers, morphological analyzers [10], [11], speech recognizers, and text classifiers [12], [13] and [14]. The latest tool that has attracted the public attention is ChatGPT. This work aims to study the accuracy of ChatGPT's humanlike-generated writings in different genres related to linguistics, online education [15], natural language processing [16] and [17], [18] translation, EFL learning [19], and Computer Science engineering. The study attempts to create a new framework for studying ChatGPT's artificial language and the responses it generates for academic inquiries. This new framework includes six principles: Relatedness, Adequacy, Authenticity, Cognition, Limitation, and Redundancy. These principles can be applied to the generated texts to test their reliability and validity.

Furthermore, they explore and measure the virtual text's naturalness, and, at the same time, their violation compared to human language. The authors introduced the new term NatFicial Language framework (NFLF) to describe the newly generated language by the machine (ChatGPT). NFLF is proposed for analyzing the newly generated language, which is produced artificially by the tool using natural human language data, as the tool was trained on big data through reinforcement learning with human feedback. The term 'NatFicial' is a blend of the two words "natural" and "artificial", which were merged to produce a new name for the state-of-the-art, newly created language that is neither entirely natural nor wholly artificial. In other words, the term 'NatFicial' is clipped and overlapped in the context of a natural and artificial language. The rest of the paper is organized as follows: section two highlights the research questions and the significance of the study. Section three describes the proposed framework and the methodology. Section four introduces the

analysis and discusses the results. Section five concludes this analysis and indicates the limitations and future implications.

2 RESEARCH QUESTIONS

In this study, we aimed to explore the accuracy and authenticity of ChatGPT3's academic writing and its similarity to natural human texts. To reach this, we formulated three main research questions. Firstly, to what extent the tool's academic writing is accurate and authentic. Secondly, what is the degree of similarity between the generated virtual texts and natural human texts? Finally, what are the differences between the tool's generated texts and natural human language in the context of academics? Through our analysis, we provide a wide-ranging understanding of the capabilities and limitations of ChatGPT3 in producing academic writing.

3 MATERIALS AND METHODS

Due to the latest advancements in large language models (LLMs), a new generic language model (ChatGPT) emerged that can produce human-like language professionally for the first time. The authors did not find an appropriate framework to explore and validate the naturalness of the generated texts. They introduced their new framework as an outline for conducting a content analysis on the tool's academically generated texts and thus tested the language in terms of fluidity, naturalness, and suitability for academic writing. They performed an experimental study on AI models that produced text from various genres to understand their relatedness, structure, style, reference, and well-formedness. There are many theories in natural texts and discourse analysis, such as Critical Discourse Analysis, Conversational Analysis, Social Constructionist Discourse Analysis, and Social Media Text Analytics Framework. All these approaches can be used to analyze natural human language. Applying the existing natural language perspectives to artificial language processing will be meaningless as Artificial Intelligence (AI) is being trained to produce human-like language vastly. Therefore, we introduce new theories and approaches to be utilized with recent technological issues, mainly language cognition and productivity. These approaches must be applied in processing human-like language produced by machines. After two months of searching and analyzing ChatGPT's language and texts, the researchers gained a deeper understanding of the generated texts and how it reflects the information. This understanding motivates the authors to propose a new framework for analyzing academic human-like language provided by the ChatGPT tool. This framework examines the naturalness of the artificial language based on five principles (Relatedness, Adequacy, Authenticity, Cognition, and Redundancy). The authors applied these methods to cover most machine language issues related to academic writing. These approaches are applied to most academic genres to understand the tool's potential performance. The findings of applying NFLF to the generated academic text analytics proposed that the higher the principles achieved, the more effective and natural the language is. The concept of these principles is specified as follows:

Relatedness: Focus on the relationship between the components of the artificial language and the text-based academic queries given to the tool.

Adequacy: Focus on the generated information's efficiency, sufficiency, and thoroughness.

Limitation: Concentrate on the capacity of the provided information about the topic.

Authenticity: Investigate the factuality of the provided information, such as references, names, ideas, definitions, etc.

Cognition: Evaluate human-like language use, including perception, comprehension, and production.

Redundancy: Find out the repetition of the generated texts' ideas, information, and references.

Figure 1 represents the principles' framework and the academic text processing.

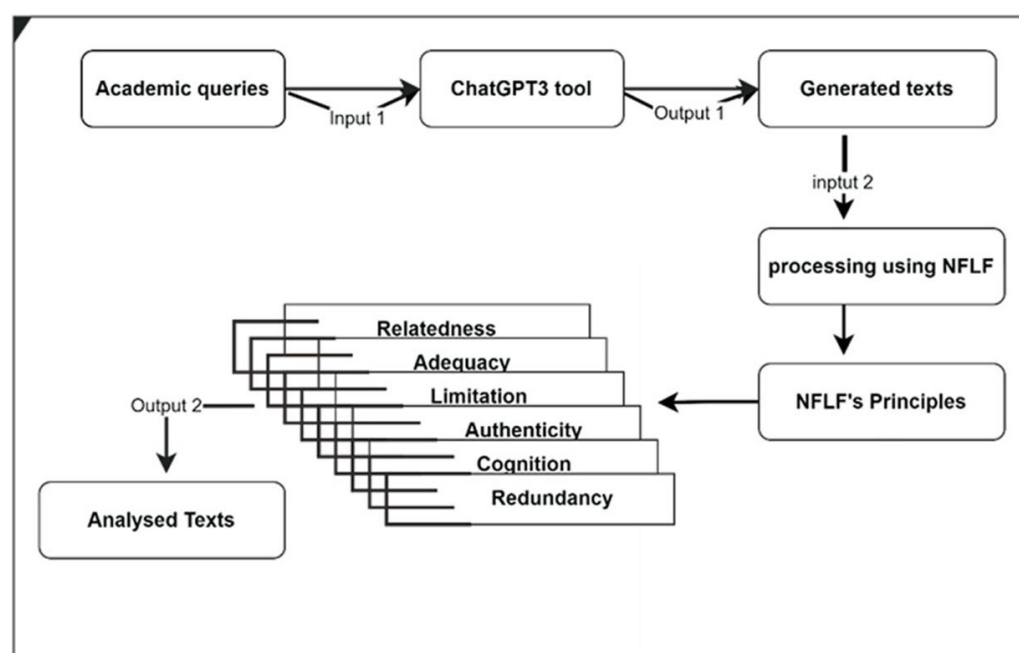


Fig. 1. The model of the study

The generated texts used in this study were collected from the ChatGPT platform by providing different academic questions about various genres related to linguistics, online education, natural language processing, translation, E-Learning, and Computer Science engineering. The research focuses on the virtual text's content to understand the language following a deep analysis using the above framework principles.

4 FINDINGS AND DISCUSSION

The chatbot AI model uses big data, vastly trained corpora, and artificial intelligence to provide texts, translation, and codes. This analysis was conducted to understand how the generated texts are related to the searched queries and to investigate the accurate and inaccurate inferences of the virtual language compared to the natural human language. In this investigation, the authors applied the five principles of their proposed framework (NFLF) to perform content analysis for the ChatGPT-produced language.

4.1 Adequacy, limitation, and cognition principle

When analyzing the generated academic texts, it is noticed that the tool produces a robust language, but the limitations could potentially be proven in many academic

contexts. The researchers first studied many examples in different genres to identify the type of language the tool produces. The following two emails were written by a student and discussed in [20]. A student was asked to write an email to his professor in English and the same email in Japanese and translate the Japanese version into English. The tool was asked to analyze these emails and provide the difference between the two writing without providing the background of the emails, as in Figure 2 below. ChatGPT explored the formality of the language and the writing style without providing any social or cultural differences in the two texts, as shown in Figure 3 below:

OS Analyze the two emails which were written by a student to his professor, what is the difference?
 Dear Jim
 Hello, I am currently working on my graduation thesis, and would like to know if it is good or not. Would you mind reading one of the chapters for me? I would really appreciate it.
 Thanks
 Tetsuya Fujimoto
 Greetings, Professor Nakamura
 Early spring, in this sizzling day, how are you spending your day? This time, I would like you to do me a favor, and this I why I take up my pen (In Japanese this means 'to write' in a formal way).
 I am now writing my graduation thesis, and even though I am afraid to ask, would you mind seeing my work . . . of course, as long as it does not bother you. If it is not inconvenient for you, could you please consider it?
 I beg you again
 Sincerely
 Tetsuya Fujimoto

Fig. 2. The question posed to the ChatGPT search engine

The first email is a straightforward request for help from Jim. Tetsuya introduces himself and the reason for the email - he is working on his graduation thesis and would like feedback on one of the chapters. Tetsuya uses a polite tone and ends the email with a simple "thanks," expressing gratitude for any help Jim can provide. The email is brief and to the point, indicating that Tetsuya values Jim's time and does not want to waste it with unnecessary information.

The second email is more formal and uses a more polite tone than the first. Tetsuya addresses Professor Nakamura directly, using a formal greeting and expressing interest in how the professor is spending his day. Tetsuya introduces his request for help - to read his graduation thesis - and uses formal language to ask if it is possible. He expresses his fear in asking and uses a very polite tone, even begging the professor for assistance. The email is longer than the first, using more words to convey the same request. It is clear that Tetsuya values the professor's time and expertise and wants to be as polite as possible in making his request.

Overall, both emails are polite and respectful, but the second email is more formal and uses a more elaborate tone. Tetsuya appears to be more cautious and careful in the second email, perhaps indicating that he has a higher level of respect for Professor Nakamura.

Fig. 3. Sample one of ChatGPT generated responses

It is noticed that the limitation principle is achieved in the style of the written language. However, when the researchers provided the tool with the email background, as shown in Figure 4 below, it generated a different discussion with cultural

and social backgrounds. GPT could partially understand cross-cultural pragmatics. As noticed in Figure 5 below, the tool interpreted the greetings of the second email as a poetic expression, not as a cultural or social behavior in the Japanese language. The tool analysis does not entirely fulfill this context's adequacy and cognition principles. However, the tool could interpret the style and structure of the two samples. These infringements prove that ChatGPT cannot handle commonsense reasoning tasks, as stated by [21] and [22].

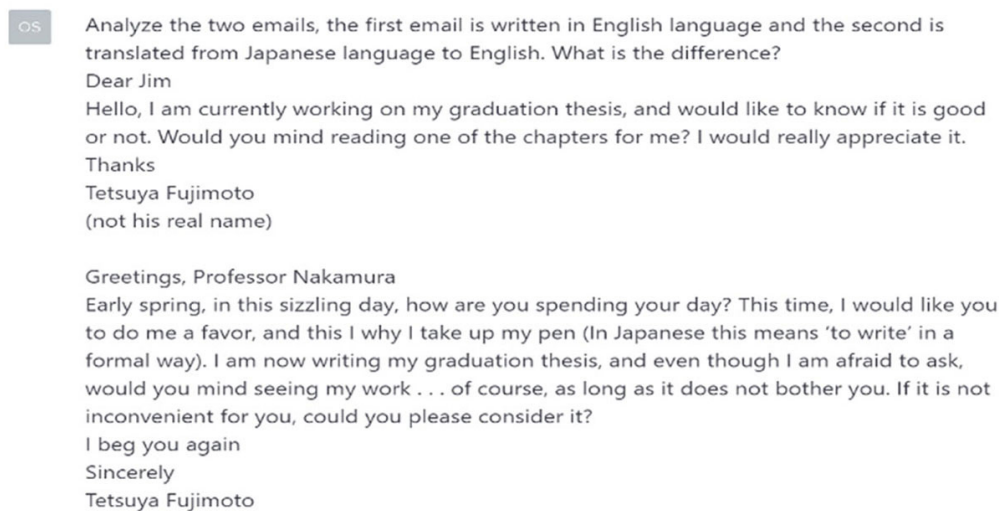


Fig. 4. The question posed to the ChatGPT search engine refers to text analysis

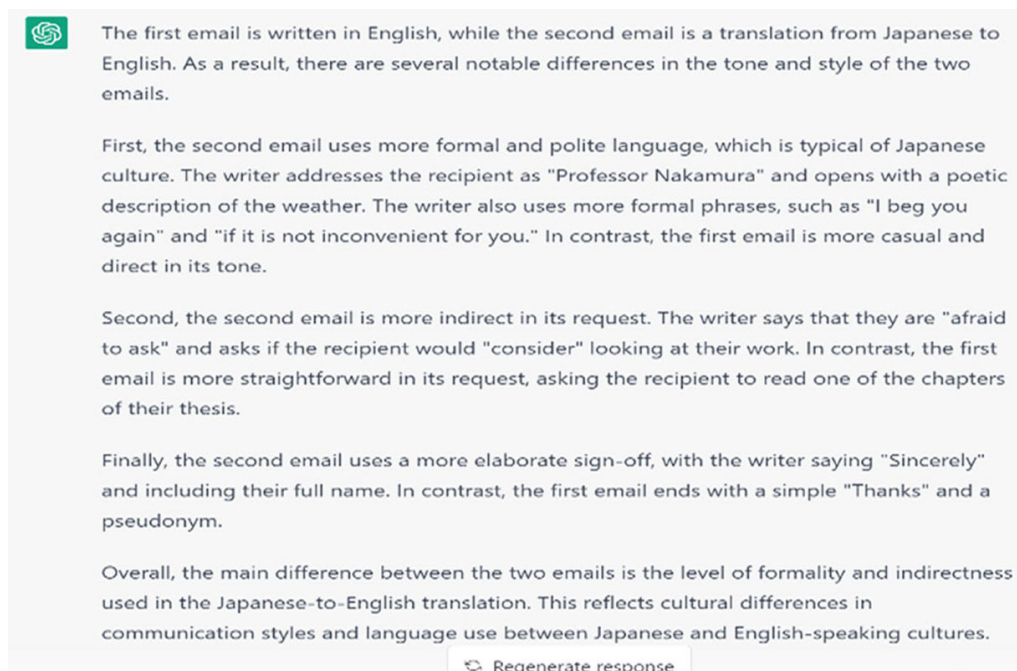


Fig. 5. Sample two of ChatGPT generated response

The tool tends to create a response based on users' instructions rather than engagement in genuine interaction, as in the response generated in Figures 3 and 5. The tool makes assumptions about what the user wants to hear rather than seeking clarifications.

4.2 Redundancy, authenticity, and relatedness principles

In the GPT3's generated texts concerning "the impact of COVID-19 on online learning post the pandemic," it provided this response:

"One of the major impacts of the pandemic on EFL pedagogy has been the increased use of technology in the classroom. With in-person instruction no longer possible, many educators have turned to virtual classrooms, video conferencing, and online tools to deliver lessons and engage with students (Chen & Liang, 2020; Liu & Liang, 2020). This has led to an increased focus on digital literacy and the use of technology in language teaching, with many EFL educators, now incorporating online resources, such as online dictionaries, grammar checkers, and language learning apps, into their instruction (Chen, Li, & Liang, 2020). Another impact of the pandemic on EFL pedagogy has been the increased use of gamification and interactive activities in online instruction (Chen & Liang, 2020; Li, 2020). With the lack of physical presence, teachers must find new ways to engage with students and keep them interested. Therefore, teachers have turned to gamification strategies, such as digital scavenger hunts and interactive quizzes, to make online learning more engaging and interactive for students." ChatGPT, 2023.

The text simultaneously infringes the redundancy and authenticity principles, i.e., the redundancy is flouted in case of citations in the provided texts, where the same author, e.g., (Chen & Liang, 2020), is repeated several times in obtained paragraphs. The authenticity principle is infringed where the citation is inevitably incorrect or misattributed. ChatGPT shows the tendency to fabricate total citations. In the above context, neither the authors nor the studies were found in the mentioned journals. It seems that the chat tool manufactured the citation where the journals, volumes, and issues exist, but the cited author and papers do not exist; more examples are shown in Appendix A & B.

Most developed language models produced nonfactual information [23], [24], [25] and [26] similar to other LLMs [27], [28], [29] and [30], the ChatGPT tool hallucinated some facts. For example, in Figure 6 below, when the researchers asked the tool about the flag of Yemen, a part of the generated response was correct, while another part of the information was revealed to be incorrect when verifying with the source. This is a common issue with LLMs that they do not have access to external knowledge bases that are consistent with [31]. The tool provided a nonfactual response, "On the hoist of the red band, there is a vertical stripe of red that features a white-five-pointed star," while the fact is that the Yemeni flag consists of three equal horizontal bands of red, white, and black only. It hallucinates with human-like fluency and persuasiveness ideas without truth, which is consistent with the study conducted by [23]. However, the relatedness principle is achieved by providing partial information relevant to the same searched topic.



Certainly, here is the flag of Yemen:

ve The flag of Yemen consists of three equal horizontal bands of red, white, and black. On the hoist side of the red band, there is a vertical stripe of red that features a white, five-pointed star. The red band represents the bloodshed and unity of the country's struggle for independence. The white band represents a bright future for Yemen, and the black band represents the dark past and present of Yemen's internal strife. The star on the red stripe symbolizes the unity of the country and its hope for a bright future.

Fig. 6. A sample of nonfactual response information

More than one of our framework's principles often overlaps in the same generated text. However, the tool exploits the principles of relatedness, adequacy, authenticity, and redundancy to provide the answer to the inquiries as shown in the sample of Appendix 1.

As shown in Figure 7, the authors asked the ChatGPT tool to discuss the sentence, "I am feeling very cold, I told my friend who is sitting next to the window," The answer to the query was simply an explanation of the situation and the tool could not provide a robust understanding of the context pragmatically. The actual perception is not discussed by the tool that the speaker wants his or her friend to close the window. Here, the tool exploits the cognition principle without comprehending that the speaker indirectly requests the friend to close the window because of the cold weather. The answer to the implicature of the following sentence by GPT indicates that the principles of relatedness and cognition are infringed because the tool could not explore the context and conceptual relatedness, which aligns with [30]. It is also noted that the ChatGPT could not identify the pronoun's antecedent as singular or plural and feminine or masculine. However, it is singular in the context, where the tool interprets "their" as an antecedent of "my friend," whereas the human interpretation would be "he/ she" to refer to my friend.

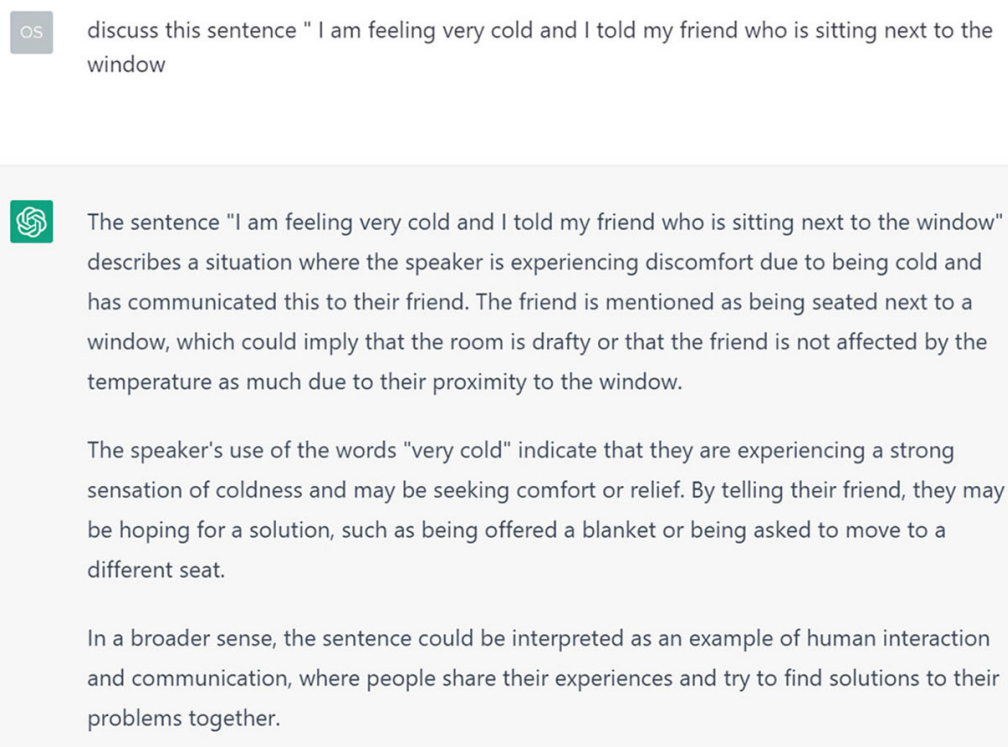


Fig. 7. Response for a simple sentence analysis

In Figure 8, the researchers added the feminine pronoun to the same sentence in the previous sample. It could interpret using the correct antecedent. However, the implicature of the sentence could not be understood/extracted by the machine, whereas human beings can easily predict the speaker's intention. Thus, the tool did not achieve the cognition principle in the two previous samples. This proves that LLMs models are intuitively text-bound [33].

os discuss this sentence "She is feeling very cold and she told her friend who is sitting next to the window"



The sentence describes a situation where a person, referred to as "She," is feeling very cold. She has communicated this to her friend, who is sitting next to the window. The sentence does not specify the location or the context, but it implies that the friend is close enough to the speaker to be able to hear the statement.

It could be a simple observation, or it might be a request for help or a suggestion to move away from the window. Without more information, it is impossible to say for sure. However, the sentence conveys the idea that the speaker is experiencing a discomfort due to the cold temperature and has conveyed this to her friend.

Fig. 8. Sample with gender specification

The results revealed that, although the AI model is an advanced language generator, it still shows deficiencies in many contexts, especially in logical reasoning, rhetoric, and pragmatics. So when we use ChatGPT technology, it is necessary to proceed with caution. Furthermore, academic writers must know that relying entirely on the tool's generated writings inevitably means they will be provided with fabricated writings supported by synthetic data and misattributed or erroneous citations.

The multi-model principles used in this framework are based on a thorough analysis of the language patterns used in academic texts. By examining the academic language's structure, vocabulary, semantics, and other factors, we can better understand how ChatGPT produces language partially suitable for academic purposes. However, apparent defects are evident. This framework can be used to analyze and evaluate the quality of academic language produced by ChatGPT and can also be used to guide the development of new language models that are better suited for academic writing.

5 CONCLUSION

This paper provides a framework and multi-model principles for understanding the characteristics of academic language produced by ChatGPT, and these principles can be combined to produce a more complete and nuanced analysis. The findings indicate that ChatGPT failed to produce well-formed academic texts and articles with authentic and reliable information. The tool can generate human-like language with grammatically correct sentences and well-formed structure due to the large, trained corpus. However, professional academic analysis and writing are still inadequate, mainly if there are issues related to pragmatics and social norms and values. The tool can be an assistive medium to guide and provide information about the various genres. However, researchers should not depend on the tool to write articles; it can be identified easily as (natficial) a fabricated article and nonhuman language. This study recommended that, when using ChatGPT in academics, it is necessary to proceed cautiously.

6 REFERENCES

- [1] Roose, K. (2023). The Brilliance and Weirdness of ChatGPT. *NYTimes*. December 5, 2022.
- [2] Shen, Y., Heacock, L., Elias, J., Hentel, K. D., Reig, B., Shih, G., & Moy, L. (2023). ChatGPT and Other Large Language Models are Double-edged Swords. *Radiology*, 230163. <https://doi.org/10.1148/radiol.230163>
- [3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... & Amodei, D. (2020). Language models are few-shot learners. *ArXiv*.
- [4] Tate, T., Doroudi, S., Ritchie, D., & Xu, Y. (2023). Educational Research and AI-Generated Writing: Confronting the Coming Tsunami. <https://doi.org/10.35542/osf.io/4mec3>
- [5] Chawla, Raveen (December 26, 2022). What is ChatGPT? History, Features, Uses, Benefits, Drawbacks 2023. Archived from the original on January 7, 2023. Retrieved December 27, 2022.
- [6] Bowman, Emma (December 19, 2022). A New AI Chatbot Might Do Your Homework for You. But it's Still Not an A+ Student. *NPR*. Archived from the original on January 20, 2023. Retrieved December 19, 2022.
- [7] OpenAI. ChatGPT: Optimizing Language Models for Dialogue. <https://openai.com/blog/chatgpt/>. 2022.
- [8] Mahyoob, M. Semi-Automatic Annotation of Arabic Corpus: A Morpho-syntactic Study, Ph.D. dissertation, Aligarh Muslim University, Aligarh, India, 2015.
- [9] Ahmed, B., Ali, G., Hussain, A., Baseer, A., & Ahmed, J. (2021). Analysis of Text Feature Extractors using Deep Learning on Fake News. *Eng. Technol. Appl. Sci. Res.*, vol. 11, no. 2, pp. 7001–7005. <https://doi.org/10.48084/etasr.4069>
- [10] Al-Garaady, J., & Mahyoob, M. (2022). Public Sentiment Analysis in Social Media on the SARS-CoV-2 Vaccination Using VADER Lexicon Polarity. *Humanities and Educational Sciences Journal*, no. 22, pp. 591–609. <https://doi.org/10.55074/hesj.v0i22.476>
- [11] Mahyoob, M., & Al-Garaady, J. (2018). Towards Developing a Morphological Analyzer for Arabic Noun Forms. *International Journal of Linguistics and Computational Applications (IJLCA)* ISSN, vol. 5, no. 3, pp. 2394–6385. <https://doi.org/10.30726/ijlca/v5.i3.2018.52012>
- [12] Mahyoob, M. (2020). Developing a Simplified Morphological Analyzer for Arabic Pronominal System. Available at SSRN 3599719. <https://doi.org/10.2139/ssrn.3599719>
- [13] Al-Garaady, Jeehaan, An Analysis of Yemenis' Responses and Sentiments on Social Media Towards the Emergence of the COVID-19 Pandemic. (2022). *Humanities and Educational Sciences Journal*, no. 27, 589–607. Available at SSRN: <https://ssrn.com/abstract=4318853>; <https://doi.org/10.55074/hesj.v0i27.621>
- [14] Zneit, R. S. A. Online Handwriting Signature Verification Based on Using Extreme Points Extraction. *Eng. Technol. Appl. Sci. Res.*, vol. 6, no. 4, pp. 1084–1088, 2016. <https://doi.org/10.48084/etasr.694>
- [15] Palasundram, K., Mohd Sharef, N., Nasharuddin, N. A., Kasmiran, K. A., & Azman, A. (2019). Sequence to Sequence Model Performance for Education Chatbot. *International Journal of Emerging Technologies in Learning (ijET)*, vol. 14, no. 24, pp. 56–68. <https://doi.org/10.3991/ijet.v14i24.12187>
- [16] Benjelloun Touimi, Y., Hadioui, A., El Faddouli, N., & Bennani, S. (2020). Intelligent Chatbot-LDA Recommender System. *International Journal of Emerging Technologies in Learning (ijET)*, vol. 15, no. 20, pp. 4–20. <https://doi.org/10.3991/ijet.v15i20.15657>
- [17] Alblwi, A., Mahyoob, M., Algaraady, J., & Mustafa, K. S. (2023). A Deterministic Finite-State Morphological Analyzer for Urdu Nominal System. *Eng. Technol. Appl. Sci. Res.*, vol. 13, no. 3, pp. 11026–11031. <https://doi.org/10.48084/etasr.5823>

- [18] Kharis, M., Schön, S., Hidayat, E., Ardiansyah, R., & Ebner, M. (2022). Mobile Gramabot: Development of a Chatbot App for Interactive German Grammar Learning. *International Journal of Emerging Technologies in Learning (IJET)*, vol. 17, no. 14, pp. 52–63. <https://doi.org/10.3991/ijet.v17i14.31323>
- [19] Al-Garaady, J., & Mahyoob, M. (2021). Social Network Communication: Emojis and EFL Learners' Writing Issues. *TESOL International*, vol. 16, no. (3.1). <https://doi.org/10.31235/osf.io/nbu8y>
- [20] Paltridge, B. (2021). *Discourse analysis: An introduction*. Bloomsbury Publishing.
- [21] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? Comparison corpus, evaluation, and detection. arXiv preprint arXiv:2301.07597.
- [22] Ernest Davis. 2023. Mathematics, word problems, common sense, and artificial intelligence. arXiv preprint arXiv:2301.09723.
- [23] Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. In *Advances in Neural Information Processing Systems*.
- [24] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022b. Survey of hallucination in natural language generation. *ACM Comput. Surv.* Just Accepted.
- [25] Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. 2022c. Rho (ρ): Reducing hallucination in open-domain dialogues with knowledge grounding. arXiv preprint arXiv:2212.01588.
- [26] Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Read before generate! faithful long form question answering with machine reading. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 744–756. <https://doi.org/10.18653/v1/2022.findings-acl.61>
- [27] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Cross-lingual generalization through multitask finetuning. arXiv preprint arXiv:2211.01786.
- [28] Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2022b. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. ArXiv, abs/2210.07688.
- [29] Rohrbach A, Hendricks LA, Burns K, Darrell T, Saenko K. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing 2018* (pp. 4035–4045). <https://doi.org/10.18653/v1/D18-1437>
- [30] Xiao Y, Wang WY. On Hallucination and Predictive Uncertainty in Conditional Language Generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume 2021 Jan.* <https://doi.org/10.18653/v1/2021.eacl-main.236>
- [31] Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., ... & Fung, P. (2023). A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *arXiv preprint arXiv:2302.04023*.
- [32] Yiqiu Shen, Laura Heacock, Jonathan Elias, Keith D. Hentel, Beatriu Reig, George Shih, & Linda Moy. (2023). Chatgpt and other large language models are double-edged swords. <https://doi.org/10.1148/radiol.230163>
- [33] Phillips, T., Saleh, A., Glazewski, K. D., Hmelo-Silver, C. E., Mott, B., & Lester, J. C. (2022). Exploring the use of GPT-3 as a tool for evaluating text-based collaborative discourse. *Companion Proceedings of the 12th*, 54.

7 AUTHORS

Mohammad Mahyoob is an Associate Professor of Computational linguistics at the Department of Languages & Translation, Al-Ula Campus, Taibah University, Madina, KSA. He is also an Associate Professor of Computational linguistics at Technical Community College -Taiz, Yemen. He holds his PhD in Computational Linguistics. His research interests include Computational linguistics, Natural Language Processing, Digital Media Studies, Linguistics, ELE and e-learning. ORCID: <https://orcid.org/0000-0002-6664-1017>

Jeehaan Algaraady is an Assistant Professor at the Centre of Languages of Taiz University, Yemen. She received her PhD in Computational Linguistics. Her research interests include Computational Linguistics, Data Mining, Theoretical Linguistics, ELT, and e-learning. ORCID: <https://orcid.org/0000-0003-3901-4648>

Abdulaziz Ablwi is an Assistant Professor at the Applied College of Taibah University, Al-Ula Campus, Madina, KSA. He received his PhD in Information Technology from Bournemouth University, UK. His research interests include Digital addiction, Requirements Engineering, Persuasive Technology, Information technology, and Machine Learning. ORCID: <https://orcid.org/0000-0002-5664-5145>