Molecular Therapy
Nucleic Acids
Commentary

Can artificial intelligence-strengthened ChatGPT or other large language models transform nucleic acid research?

Srijan Chatterjee,[1,4] Manojit Bhattacharya,[2] Sang-Soo Lee,[1] and Chiranjib Chakraborty[3,4]

https://doi.org/10.1016/j.omtn.2023.06.019

The impressive accomplishments of large language models (LLMs), particularly the "Chat Generative Pre-Trained Transformer" created by OpenAI, usually referred to as "ChatGPT," have been the subject of significant media coverage in recent times.[1] These LLMs are artificial intelligence (AI) programs that create text closely resembling human language using sophisticated recurrent neural networks trained on large datasets.[2] Within a brief period, the launch of ChatGPT has significantly impacted the academic community. This technology will significantly alter how researchers conduct their work (https://openai.com/blog/chatgpt/). LLMs mostly use deep learning to convincingly mimic human language. These models are now used in content marketing, customer support, and many corporate contexts, and their usage is on the rise.[3] AI is already used in healthcare, and it has the power to completely change how patients are cared for and how administrative tasks are carried out in hospitals and pharmaceutical firms. The potential of AI in healthcare was covered by Davenport and Kalakota,[4] who found that healthcare providers and life science businesses already use various forms of AI. These AI applications can be broadly divided into administrative chores, patient engagement and adherence, and diagnosis and therapy recommendations.[4] Patel and Lam demonstrated how ChatGPT could generate a patient discharge statement based on a short prompt, demonstrating the technology's potential to automate and speed up hospital discharges. This automation offers the advantage of preserving the essential degree of detail while freeing up doctors' time for patient care and professional development.[5] Another study looked at ChatGPT's usefulness in simplifying radiology reports, and the results were determined to be accurate, complete, and unlikely to pose substantial dangers to patients.[6] Thus, ChatGPT has the potential to significantly change medical research in several ways. Before incorporating ChatGPT deeply into clinical research and medical practice, in-depth conversations must be conducted to improve its originality, accuracy, and academic integrity.[7] Notably, researchers are trying to evaluate the application of these LLMs in various fields. However, few studies have examined the application of LLMs to nucleic acid research (Figure 1).

In a recent editorial, Page et al. demonstrated that AI could transform microbial genomics research by improving data processing and speeding up procedures. AI can aid in identifying several essential components of genomic research, such as regulatory elements and the functions of several genes. Furthermore, AI can predict microbial behavior, disclose new gene clusters, and propose ideas for experimental verification, resulting in a significant acceleration in discoveries and a better understanding of bacteria and their interactions with the environment.[8] At the same time, researchers have tried to assess the level of understanding of ChatGPT in nucleic acid research using the GeneTuring test.[9] The GeneTuring test, devised by Hou and Ji, can help evaluate the fitness of GPT models for genetic research. The researchers wanted to see how well GPT models perceive and generate genomics-related information. The models were tested on various genomic tasks, including gene prediction, variation analysis, and DNA sequence generation. Armed with rigorous metrics and benchmarks, the researchers demonstrated the ability of GPT models to reliably anticipate genetic information and generate relevant insights in genomics. The findings established the superiority of AI models in genomics and showed promise for future breakthroughs.[9]

Scientists have also tried to compare the ability of LLMs and humans to answer genetic questions. Duong and Solomon compared the performance of humans and ChatGPT in answering 85 human genetics-related multiple-choice questions[10]; ChatGPT showed 68.2% accuracy. These models can have an immediate impact by quickly and accurately answering many genetics-related questions. These strategies can help medical practitioners diagnose and treat genetic illnesses and provide readily available information about conditions to patients and their families.

Furthermore, ChatGPT's ability to understand and reply to simple language questions can improve access to genetic information for people without prior knowledge of the subject. As genetic research advances, natural language processing models such as ChatGPT will become more important in research and medical settings.[10] Similarly, LLMs can be used in bioinformatics teaching, including demonstrations of phylogenetic analysis. In one such study, Shue et al. examined ChatGPT's assistance to students in phylogenetic studies. The researchers tasked the Chatbot with creating R code for developing a phylogenetic tree comprising

[1]Institute for Skeletal Aging & Orthopaedic Surgery, Hallym University-Chuncheon Sacred Heart Hospital, Chuncheon-si, Gangwon-do 24252, Republic of Korea; [2]Department of Zoology, Fakir Mohan University, Vyasa Vihar, Balasore, Odisha 756020, India; [3]Department of Biotechnology, School of Life Science and Biotechnology, Adamas University, Kolkata, West Bengal 700126, India
[4]These authors contributed equally

Correspondence: Prof. Chiranjib Chakraborty, MSc, PhD, Department of Biotechnology, School of Life Science and Biotechnology, Adamas University, Kolkata, West Bengal 700126, India.
E-mail: drchiranjib@yahoo.com

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

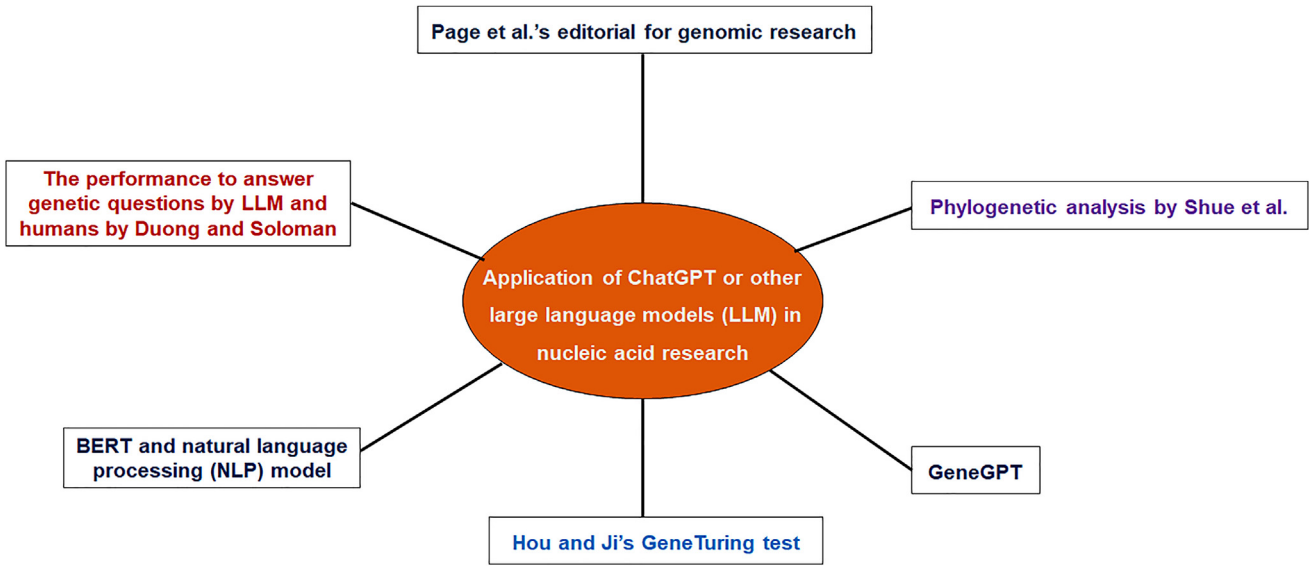

Figure 1. Recent applications of ChatGPT or other large language model (LLM) in nucleic acid research

nine different species. The work began with aligning protein-coding sequences from the TP53 tumor suppressor gene. Initially, the Chatbot was given instructions explaining the major processes required to produce an unrooted tree. The Chatbot successfully created functioning code capable of creating a reasonably accurate unrooted phylogenetic tree after two rounds of changes and incorporating feedback from humans regarding error messages observed during code execution. At the same time, the Chatbot was instructed to design a rooted phylogenetic tree with designated species; however, the Chatbot failed to provide a valid result.[11]

Scientists are also trying to create LLMs specific for the nucleic acid field. For example, Jin and colleagues' novel GeneGPT model trains LLMs to use NCBI web APIs. GeneGPT incorporates specialized techniques such as gene mention identification, entity linkage, and database integration to improve comprehension and create biomedical information. The usefulness of GeneGPT has been evaluated through several tests, such as creating protein-protein interaction networks and resolving biomedical concerns. Interestingly, GeneGPT outperformed traditional language models in biological applications by providing more precise and contextually relevant information.[12] Transformer-based LLMs have also found success in interpreting lengthy DNA sequences in recent years. DNABERT, a pre-trained bidirectional encoder representation and natural language processing model, can understand global and transferable genomic DNA sequences, which it accomplishes by considering both the upstream and downstream nucleotide environments.[13] For example, DeepMind's Enformer transformer model uses self-attention methods to incorporate more detailed DNA context. As a result, it obtains higher precision when forecasting gene expression using DNA sequences.[14] However, additional research is required to determine the effects of various *cis*-acting DNA components' *trans*-acting factors and predict where enzyme molecules will bind.[15]

LLMs such as ChatGPT can revolutionize nucleic acid research. The ability of LLMs to analyze and output massive amounts of text-based information has already proven helpful in various sectors, including nucleic acid research. ChatGPT can potentially transform nucleic acid research, but its implementation requires prudence and responsibility. When properly safeguarded, LLMs can become vital instruments, allowing researchers to make substantial advances in understanding the complex world of nucleic acids and their function in biology and health. They can be virtual assistants, answering questions, reviewing literature, and summarizing study findings. Using these tools can significantly speed up the research process and boost the likelihood of success in specific nucleic acid-based experiments and publishing the results. This ease of access to information can stimulate interdisciplinary partnerships, allowing scientists from various fields to apply nucleic acid research more efficiently. However, further research is needed to fully understand the capabilities of ChatGPT or other LLMs to alter the field of nucleic acid research while keeping ethical concerns about data privacy, bias, and the responsible usage of LLMs in mind.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Looi, M.K. (2023). Sixty seconds on ChatGPT. BMJ *380*, 205.
2. Alberts, I.L., Mercolli, L., Pyka, T., Prenosil, G., Shi, K., Rominger, A., and Afshar-Oromieh, A. (2023). Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? Eur. J. Nucl. Med. Mol. Imaging *50*, 1549–1552.

3. Editorials. (2023). Will ChatGPT transform healthcare? Nature medicine *29*, 505–506.

4. Davenport, T., and Kalakota, R. (2019). The potential for artificial intelligence in healthcare. Future Healthc. J. *6*, 94–98.

5. Patel, S.B., and Lam, K. (2023). ChatGPT: the future of discharge summaries? Lancet. Digit. Health *5*, e107–e108.

6. The Lancet Digital, H. (2023). ChatGPT: friend or foe? The Lancet Digital health *5*, e102.

7. Ruksakulpiwat, S., Kumar, A., and Ajibade, A. (2023). Using ChatGPT in Medical Research: Current Status and Future Directions. J. Multidiscip. Healthc. *16*, 1513–1520.

8. Page, A.J., Tumelty, N.M., and Sheppard, S.K. (2023). Navigating the AI frontier: ethical considerations and best practices in microbial genomics research. Microb. Genom. 9. https://doi.org/10.1099/mgen.0.001049.

9. Hou, W., and Ji, Z. (2023). GeneTuring tests GPT models in genomics. Preprint at bioRxiv. https://doi.org/10.1101/2023.03.11.532238.

10. Duong, D., and Solomon, B.D. (2023). Analysis of large-language model versus human performance for genetics questions. Eur. J. Hum. Genet. https://doi.org/10.1038/s41431-023-01396-8.

11. Shue, E., Liu, L., Li, B., Feng, Z., Li, X., and Hu, G. (2023). Empowering Beginners in Bioinformatics with ChatGPT. Preprint at bioRxiv. https://doi.org/10.1101/2023.03.07.531414.

12. Jin, Q., Yang, Y., Chen, Q., and Lu, Z. (2023). GeneGPT: Augmenting Large Language Models with Domain Tools for Improved Access to Biomedical Information. Preprint at arXiv. https://doi.org/10.48550/arXiv.2304.09667.

13. Ji, Y., Zhou, Z., Liu, H., and Davuluri, R.V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. Bioinformatics *37*, 2112–2120.

14. Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D.R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods *18*, 1196–1203.

15. Wang, D.-Q., Feng, L.-Y., Ye, J.-G., Zou, J.-G., and Zheng, Y.-F. (2023). Accelerating the integration of ChatGPT and other large-scale AI models into biomedical research and healthcare. MedComm – Future Medicine *2*, e43.