iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62586

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

BioGPT: The ChatGPT of Life Sciences

Shelar Aniket[1], Kowe Ankit[2], Prof Hiranwale S. B[3]
Department of Computer Engineering, HSBPVT's GOI FOE, Kashti, Maharashtra, India

Abstract: ***In the medical field, pictures of the body are really important. But answering questions about these images to help diagnose problems is tricky. We made a new model called CGMVQA to help with this. It's good at sorting out different types of questions and finding answers. We use special techniques to understand written questions and make images clearer. We also made the model work faster by changing some settings. Our model is really good! It's the best one yet at answering medical image questions, like figuring out what's in a picture or matching words to images. This means doctors can use it to help with diagnoses. BioGPT is another helpful tool. It's great for scientists and teachers in biology. It gives quick and accurate answers to complex biology questions. As technology gets better, BioGPT will help us learn more about life and speed up biological research.***

Keywords: BioGPT, ChatGPT, life sciences, molecular biology, genomic analysis, genetic engineering, bioinformatics, research tool, data-driven insights, comprehensive, accuracy, quick responses, advancements, biological sciences.

I. INTRODUCTION

In recent advancements within the realm of artificial intelligence (AI), Microsoft garnered attention with the launch of ChatGPT, an innovative chatbot developed by OpenAI, in November of the preceding year. However, a lesser-known yet equally significant unveiling by Microsoft occurred in January of the subsequent year with the introduction of BioGPT. Unlike ChatGPT, BioGPT is tailored specifically for the biomedical domain, offering a unique AI tool designed to evaluate biomedical research and provide insights into complex biomedical queries. Leveraging generative language models trained on millions of published biomedical research articles, BioGPT possesses the capability to extract relevant information, generate text, and offer answers to biomedical questions. This introduction of BioGPT signifies a pivotal step in empowering researchers with a powerful AI-driven resource for gaining fresh perspectives and enhancing biomedical research endeavors. In this paper, we delve into the development and capabilities of BioGPT, a generative pretrained Transformer language model, highlighting its potential in revolutionizing the landscape of biomedical text generation and analysis.

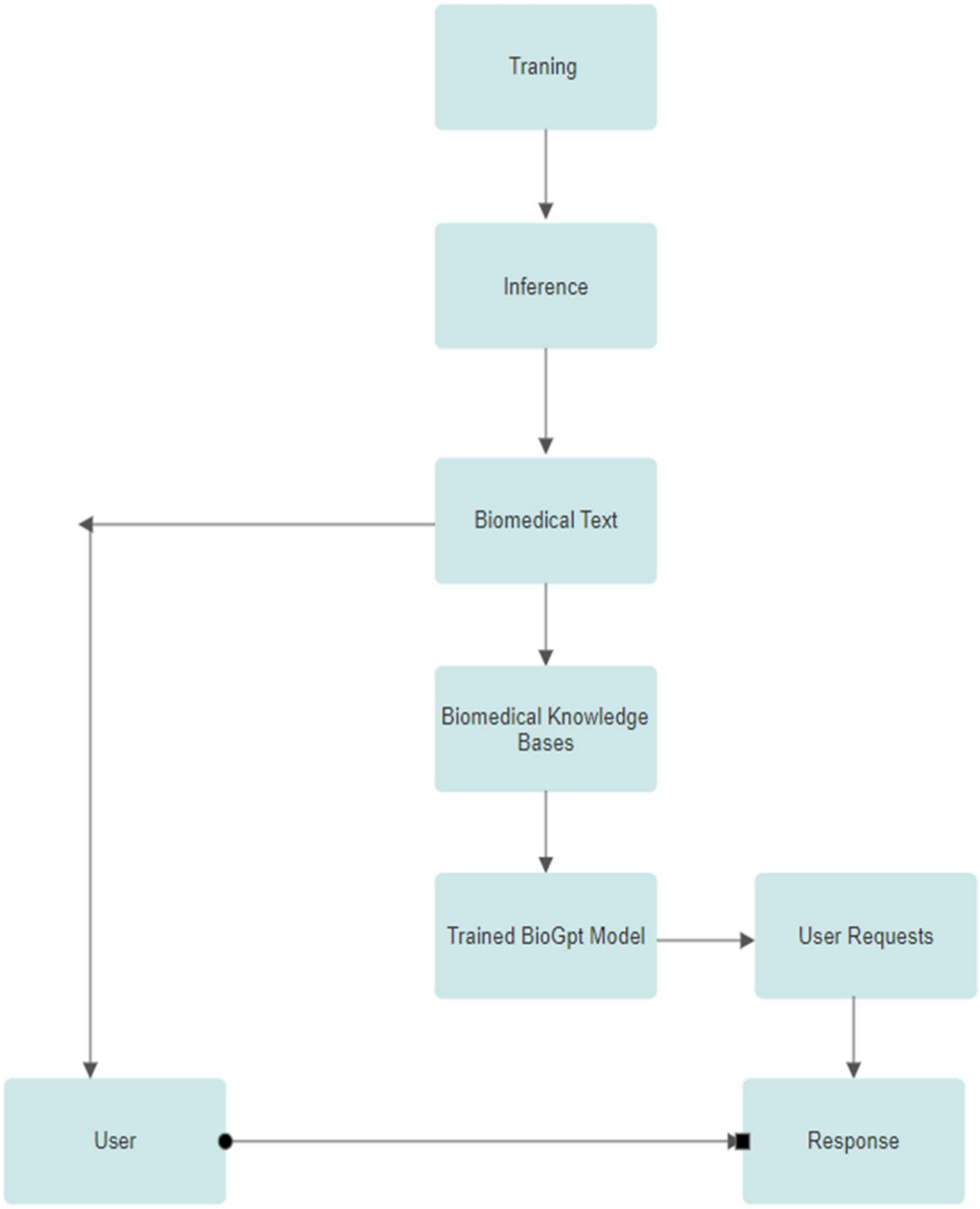
II. PROJECT SCOPE

The project aims to develop a large language model that can generate and understand biomedical text in a comprehensive and informative way.

This includes the following goals:

1) To train a language model that can generate text in a variety of biomedical formats, such as scientific papers, clinical trial protocols, and drug discovery reports.
2) To train a language model that can answer questions about biomedical topics in a comprehensive and informative way.
3) To train a language model that can extract relationships between entities in biomedical text, such as gene-disease interactions and drug-target interactions.
4) To train a language model that can classify biomedical documents into different categories, such as disease types and drug classes.
5) To develop a comprehensive documentation for BioGPT that includes instructions on how to use BioGPT and how to interpret its results.
6) To develop a set of benchmark datasets for BioGPT that can be used to evaluate its performance on a variety of biomedical tasks.
7) To release BioGPT to the public so that it can be used by researchers and clinicians to advance biomedical research and to improve the lives of patients.

Fig.1: Data Flow



III. MATHEMATICAL MODEL

A mathematical model for BioGPT could involve various mathematical techniques and algorithms to represent the underlying processes involved in natural language processing (NLP) and machine learning. Since BioGPT is designed for bioinformatics applications, the mathematical model would need to incorporate domain-specific knowledge and techniques tailored to analyzing biological data.

1) *Statistical Models:* Statistical models are commonly used in NLP for tasks such as language modeling, text classification, and sequence labeling. Techniques like n-gram models, hidden Markov models (HMMs), and conditional random fields (CRFs) could be employed to capture the statistical properties of biological text data
2) *Deep Learning Models:* Deep learning models, particularly recurrent neural networks (RNNs) and transformer- based architectures like the one used in GPT (Generative Pre-trained Transformer), are effective for capturing complex patterns and dependencies in sequential data such as text.

3) *Word Embeddings:* Word embeddings are dense vector representations of words that capture semantic relationships between words based on their co-occurrence patterns in a corpus of text. Techniques like Word2Vec, GloVe, and BERT (Bidirectional Encoder Representations from Transformers) could be used to generate word embeddings for biological terms and documents, enabling BioGPT to understand the context and meaning of biological text data.
4) *Domain-Specific Knowledge Representation:* Incorporating domain-specific knowledge into the mathematical model is essential for understanding and processing biological text effectively. This could involve using ontologies, semantic networks, or structured databases to represent biological concepts, relationships, and annotations in a mathematical form that can be utilized by BioGPT.
5) *Feature Engineering:* Feature engineering involves transforming raw input data into a suitable representation for machine learning algorithms. For BioGPT, feature engineering could involve extracting features from biological text data, such as word frequencies, n-grams, syntactic patterns, or domain-specific features like gene names, protein sequences, and biological annotations.
6) *Evaluation Metrics:* Mathematical models for BioGPT would also need to incorporate evaluation metrics to assess their performance on specific bioinformatics tasks, such as sequence classification, named entity recognition, or document classification.

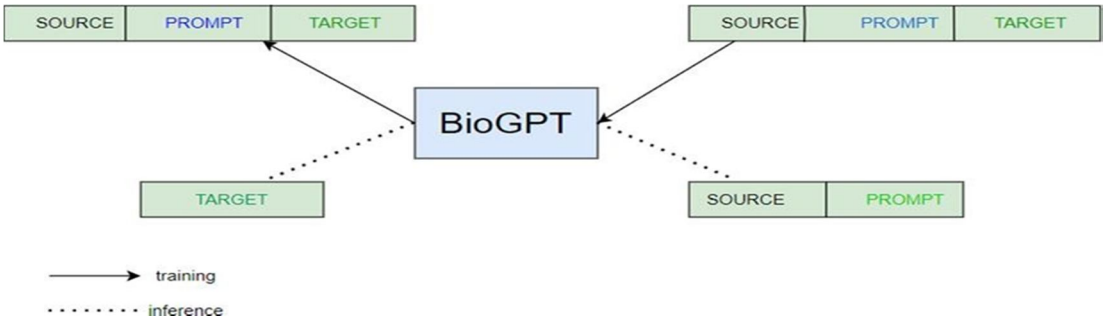


Fig: System Architecture.

The BioGPT system architecture consists of the following components:

a) *Data Layer:* The data layer stores the data that is used to train and deploy BioGPT. This data includes a large dataset of unlabeled biomedical text and a labeled dataset for fine-tuning BioGPT on specific biomedical tasks.
b) *Training Layer:* The training layer is responsible for training the BioGPT model. The training layer uses the data from the data layer to train the BioGPT model to perform a variety of tasks, such as generating text, answering questions, and extracting relationships between entities in biomedical text.
c) *Inference Layer:* The inference layer is responsible for deploying the BioGPT model and making it available to users. The inference layer receives requests from users and uses the BioGPT model to generate responses.
d) *API Layer:* The API layer provides a way for users to interact with the BioGPT system. The API layer exposes a set of endpoints that users can call to generate text, answer questions, andextract relationships between entities in biomedical text.

The BioGPT system architecture is designed to be scalable and reliable. The data layer is distributed across multiple servers to ensure that it can handle a large volume of data. The training layer is also distributed across multiple servers to speed up the training process. The inference layer is stateless, which means that it can be scaled horizontally to handle a large number of concurrent requests.

IV. SOFTWARE AND HARDWARE REQUIREMENT

A. *Software Requirements*

1) *Cloud Computing Platforms*

a) *Amazon Web Services (AWS):* AWS offers a comprehensive set of cloud computing services, including compute, storage, database, and machine learning services. BioGPT can leverage AWS Lambda for serverless computing, Amazon S3 for scalable storage, andAmazon SageMaker for machine learning model training and deployment.
b) *Microsoft Azure:* Azure provides a wide range of cloud services, including virtual machines, databases, AI services, and DevOps tools. BioGPT can utilize Azure Functions for serverless computing, Azure Blob Storage for data storage, and Azure Machine Learning for model training and deployment.

2) Containerization Platforms

a) Docker: Docker enables packaging BioGPT and its dependencies into lightweight, portable containers that can run consistently across different environments. BioGPT containers can be deployed on-premises or in cloud environments, providing flexibility and portability.

3) Serverless Computing Platforms

a) AWS Lambda: AWS Lambda allows running BioGPT functions without provisioning or managing servers. It offers automatic scaling, fine-grained billing based on usage, and seamless integration with other AWS services.

b) Azure Functions: Azure Functions provides serverless computing capabilities similar to AWS Lambda, enabling BioGPT to execute code in response to events with automatic scaling and pay-per-use pricing.

B. Hardware Requirements

1) CPU

- BioGPT requires a powerful CPU to perform the complex computations involved in natural language processing (NLP) and machine learning tasks. A multicore CPU with high clock speeds is recommended to handle parallel processing efficiently.
- The specific CPU requirements depend on the size of the dataset, the complexity of the models, and the expected workload. A modern multi-core processor, such as an Intel Corei7 or AMD Ryzen processor, would provide sufficient processing power for BioGPT.

2) Memory (RAM)

- Sufficient RAM is essential for BioGPT to load and process large language models and datasets efficiently. The amount of memory required depends on the size of the model and the size of the input data.
- As a general guideline, BioGPT typically requires several gigabytes of RAM to operate effectively. For optimal performance, a minimum of 16 GB of RAM is recommended, butlarger models or datasets may require even more memory.

3) Storage

- BioGPT requires storage space to store model files, datasets, and other auxiliary data. The storage capacity needed depends on the size of the models, the volume of data being processed, and the desired retention period for stored data.
- It is recommended to use fast solid-state drives (SSDs) for storage to ensure quick access to data and faster processing times. The storage capacity required varies based on the specific use case and dataset size but could range from hundreds of gigabytes to multiple terabytes.

4) Networking

- BioGPT may require a stable and high-speed network connection to access external data sources, APIs, or cloud services, especially in distributed computing environments or when processing data stored remotely.
- It is recommended to have a network connection with sufficient bandwidth to handle data transfer requirements effectively, ensuring smooth operation and timely access to externalresources.

5) Operating System

- BioGPT can run on various operating systems, including Linux, Windows, and macOS. The choice of operating system depends on compatibility requirements, developer preferences, and organizational policies.

V. ALGORITHM DETAILS

A. Transformer Architecture

1) The Transformer architecture is a type of neural network architecture that is well-suited for natural language processing (NLP) tasks. It uses a self-attention mechanism to learn long-range dependencies in the input sequence.

2) The Transformer architecture is used in BioGPT to learn the relationships between words and phrases in biomedical text. This allows BioGPT to generate text, answer questions, and extract relationships between entities in a comprehensive and informative way.

B. Masked Language Modeling(MLM)

1) MLM is a pre-training objective that involves predicting masked tokens in a sentence. This helps the model to learn the relationships between words and phrases in the language.

2) MLM is used in BioGPT to pre-train the model on a large dataset of unlabeled biomedical text. This helps the model to learn the patterns and relationships in biomedical text, which enables it to perform a variety of tasks, such as text generation, question answering, and relation extraction.

C. Next Sentence Prediction(NSP)

1) NSP is a pre-training objective that involves predicting whether two sentences are consecutive in a document. This helps the model to learn the relationships between sentences and paragraphs in the text.

2) NSP is used in BioGPT to pre-train the model on a large dataset of biomedical documents. This helps the model to learn the relationships between sentences and paragraphs in biomedical text, which enables it to perform tasks such as question answering and document classification.

D. Contrastive Learning of Biomedical Entites(CLBE)

1) CLBE is a pre-training objective that involves learning to distinguish between positive and negative pairs of biomedical entities. This helps the model to learn the relationships between different biomedical entities, such as genes, diseases, and drugs.

2) CLBE is used in BioGPT to pre-train the model on a large dataset of biomedical entities. This helps the model to learn the relationships between different biomedical entities, such as genes, diseases, and drugs, which enables it to perform tasks such as relation extraction and document classification.

VI. CONCLUSION

BioGPT emerges as an indispensable asset for fostering inspired innovations across diverse sectors. Its capacity to swiftly generate creative solutions to existing challenges empowers individuals and organizations to maintain a competitive edge within their industries. Moreover, BioGPT's versatility transcends boundaries, rendering it applicable across various fields of study and industries. Whether seeking novel ideas or solutions, BioGPT stands out as the ultimate tool for innovation. Its rapid ideation capabilities revolutionize conventional problem-solving approaches, promising transformative outcomes. Furthermore, BioGPT contributes to advancing science communication and education by simplifying intricate biological concepts, thereby facilitating broader accessibility and understanding. As a catalyst for innovation and knowledge dissemination, BioGPT epitomizes the potential of AI in driving positive change and progress.

VII. FUTURE SCOPE

1) To develop a language model capable of producing text in different biomedical formats, including drug discovery reports, clinical trial protocols, and scholarly publications.

2) To develop a language model capable of providing thorough and enlightening responses to inquiries regarding biomedical subjects To develop a language model capable of deriving connections between entities in biomedical text, such as those involving drugs and their targets or genes and diseases.

3) To develop a language model capable of categorizing biological texts according to various attributes, like medication classes and disease kinds.

4) To create thorough documentation for BioGPT that explains how to use the tool and how to interpret its findings.

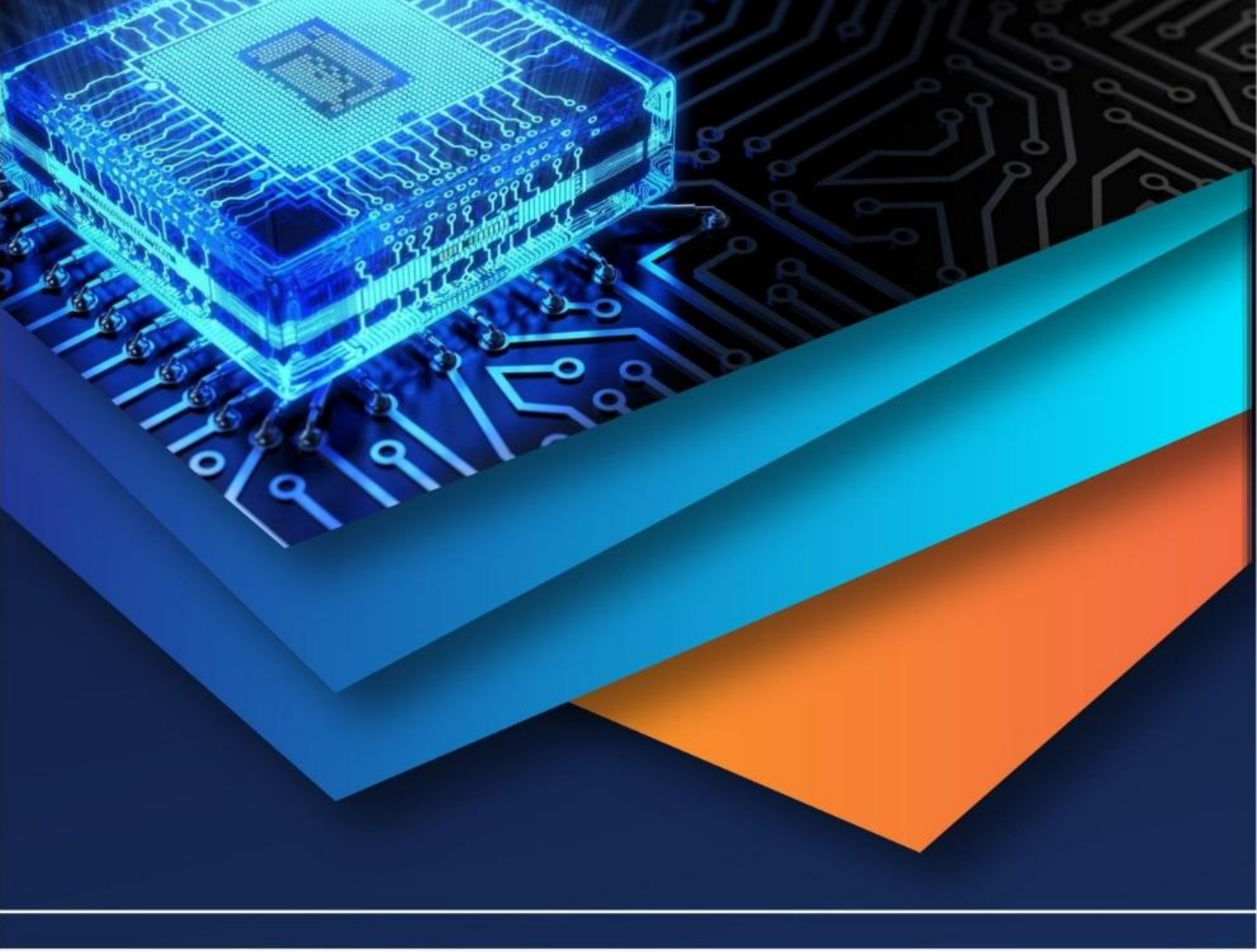
5) To create a collection of benchmark datasets for BioGPT so that its efficacy on various biomedical activities may be assessed.

6) To make BioGPT available to the general public so that physicians and researchers can use it to enhance patient care and progress scientific research.

REFERENCES

[1] Fuji Ren, (Senior Member, IEEE), and Yangyang Zho. CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering. 2020
[2] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. 201
[3] Shamsi Daneshi, Anthony Gitter. Attention Is All You Need: A Review of Attention Mechanisms in NLP and their Application in Genomicc.2020
[4] Andrew M Jones, Max Bileschi, Gokhan Tur, A. Gilad Kusne. Applications of deep learning and reinforcement learning to biological data. 2020
[5] Georgior Balikas, Prodromos Malakasiotis, Ioannis Partalas, et al. BioASQ at 7: Large-scale Biomedical Semantic Indexing and Question Answering. 2020

[6] Guo J, Huang X, Dou L, Yan M, Shen T, Tang W, Li J. Aging and aging-related diseases: from molecular mechanisms to interventions and treatments. SignalTransductTargetTher. 2022; 7:391.
[7] Aging: Molecular Pathways and Implications on the Cardiovascular System. Oxid Med Cell Longev. 2017; 2017:794156
[8] The hallmarks of aging. Cell. 2013; 153:1194–217. . BioGPT-The ChatGpt of life sciences 20
[9] Galkin F, Mamoshina P, Aliper A, Putin E, Moskalev V, Gladyshev VN, Zhavoronkov A. Human Gut Microbiome Aging Clock Based on Taxonomic Profiling and Deep Learning. iScience. 2020; 23:101199.
[10] Zhavoronkov A. Generation of Novel Chemistry. Mol Pharm. 2018; 15:4311–3.
[11] Pun FW, Liu BHM, Long X, Leung HW, Leung GHD, Mewborne QT, Gao J, Shneyderman A, Ozerov IV, Wang J, Ren F, Aliper A, Bischof E, et al. Identification of Therapeutic Targets for Amyotrophic Lateral Sclerosis Using PandaOmics - An AIEnabled Biological Target Discovery Platform. FrontAging Neurosci. 2022; 14:914017
[12] Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel). 2023; 11:887
[13] Luo R,BioGPT: generative pretrained transformer for biomedical text generation and mining. Brief Bioinform. 2022; 23:bbac409
[14] Kandhaya-Pillai R, Yang X, Tchkonia T, Martin GM, Kirkland JL, Oshima J. TNFα/IFN-γ synergy amplifies senescence-associated inflammation and SARS-CoV-2 receptor expression via hyper-activated JAK/STAT1. Aging Cell. 2022; 21:e13646
[15] Tchkonia T, Niedernhofer LJ. Senolytic Drugs: Reducing Senescent Cell Viability to Extend Health Span.Annu Rev PharmacolToxicol. 2021; 61:779–803



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY