Application and ethical challenges of large language model and agent technology in biomedicine

Zhicheng Du[1 [2024312754]]

Institute of Biopharmaceutics and Health Engineering, Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China
duzc24@mails.tsinghua.edu.cn

Abstract. With the development of artificial intelligence technology, large language models (LLMs) and Agent technology have been widely used in the biomedical field. LLMs, which is based on deep learning and Transformer architecture and trained on large-scale text data, shows strong language understanding and generation capabilities, and plays an important role in biomedical research, clinical practice, and medical education. Agent technology helps to simulate biological systems and optimize healthcare management by building intelligent entities with autonomy and sensing capabilities. This paper describes the application of LLMs and Agent technology in biomedicines, analyzes its challenges, such as data privacy, algorithm bias, model interpretability and reliability defects, liability attribution and legal disputes, and discusses countermeasures. These measures include establishing an ethical review mechanism, strengthening data management, improving model transparency, clarifying responsibility framework, promoting interdisciplinary cooperation and public education, and promoting resource equity and universal access to healthcare, so as to provide reference for related research and practice.

Keywords: Large Language Models (LLMs) · Agent · Ethics · Biomedical applications.

1 Introduction

1.1 Large language model

Large Language Models (LLMs) are complex artificial intelligence tools built on deep learning techniques(1; 2), especially the Transformer architecture(3; 4), by training on large-scale textual data. In recent years, LLM architectures based on Mamba(5) and RMKV(6) models have also been extensively studied. They are able to generate texts that are highly similar to human language, demonstrating strong language comprehension and generation capabilities. In recent years, with the continuous progress of technology, the application of LLMs in the field of biomedicine has gradually expanded, covering multiple levels from basic research

to clinical practice. In general, LLM achieves specialized applications through domain fine-tuning and knowledge base integration.
In biomedical research, LLMs are used for literature review and analysis, which can quickly process and integrate massive scientific literature and provide researchers with comprehensive background knowledge and insight into research trends. For example, LLMs can assist scientists in tracking recent advances in the research field of a particular disease, or help identify potential connections between different studies that can inspire new research directions(7). In drug research and development, LLMs can be used to analyze information about drug targets, predict potential side effects of drugs, and optimize the drug design process(8). By learning from a large amount of biomedical data, LLMs can provide valuable suggestions and references for drug researchers to accelerate the discovery and development process of new drugs.
In the field of clinical practice, LLMs are equally widely used. They can be used as clinical decision support tools to provide doctors with diagnostic suggestions, treatment recommendations, and disease prognosis evaluation(9). For example, when a physician is faced with a complex case, LLMs can quickly retrieve and analyze similar case data based on the patient's symptoms, medical history, and other relevant data, providing the physician with possible diagnostic directions and treatment strategies. In addition, LLMs also plays an important role in medical education, which can generate personalized learning materials, simulate clinical scenarios, and help medical students improve their learning efficiency and practical ability(9). In recent years, several specialized models have emerged in the material(10) and medical field, such as Med-PaLM (medical question answering)(11), BioGPT (biomedical text generation))(12), and GatorTron (medical semantic analysis)(13), which have significantly improved their performance in biomedical tasks through fine tuning and multimodal fusion (such as combining text, image, and genetic data) (14; 15). For example, Med-PaLM 2 achieved a diagnostic accuracy as high as 93.06% on the MedQA dataset, which is close to the level of human experts.

1.2 Agent technology

Agent technology refers to the technology of constructing intelligent entities with characteristics such as autonomy, social competence, responsiveness and initiative. It constructs "intelligent agent system" by endowing LLM with closed-loop ability of perception, thinking and action. In biomedicine, agents can be designed to model various processes and interactions in living organisms, such as cell signaling, gene regulatory networks, etc. By creating these intelligent agents, researchers are able to better understand complex biological systems and disease mechanisms. Such systems can independently perform experimental design, data analysis, and hypothesis verification, significantly shortening the research and development cycle.
Agents can also be applied in healthcare management systems, where they are responsible for tasks such as monitoring patients' health status, reminding patients to take their medication, and booking medical services(16). For example,

intelligent agents can be integrated with wearable devices to collect physiological data of patients in real time, such as heart rate, blood pressure, blood glucose, and so on, and analyze them according to preset rules and algorithms(17). Once abnormal conditions are found, the Agent can alert the patient or medical staff in time to achieve dynamic management and early intervention of patient health. In the clinical scenario, the Agent system can integrate electronic health records (EHRs)(18), automatically generate discharge summaries, care plans, and provide personalized health education.

2 Positive impact

Advancing biomedical research. Large language models and Agent technology provide powerful tool support for biomedical research, which can accelerate the research process and improve the efficiency of research. LLMs can quickly process and analyze a large amount of biomedical literature and data, helping researchers to understand the research field more comprehensively, avoid duplication of effort, and discover new research entry points. By simulating biological systems, Agent technology enables researchers to conduct various experiments and simulation studies in a virtual environment, reducing the cost and risk of actual experiments, and providing a new perspective for understanding complex biological phenomena. For example, the AI agent reported in the journal Cell can simulate the virtual cellular environment and predict the mechanism of drug action to promote the development of precision medicine(19). In addition, the "zero-sample learning" capability of LLM enables it to analyze cross-domain data without task-specific training, such as discovering new biomarkers from tumor gene expression profiles(20). In the challenge of rare diseases(21), it involves multiple disciplines such as hematology, orthopedics, neurology, kidney, respiratory, skin and so on. Clinicians generally lack professional knowledge of rare diseases. Synthetic data generated by LLM can effectively alleviate the bottleneck of insufficient samples in rare disease research(21), and achieve the purpose of accelerating knowledge transformation and innovation. In addition, in resource-poor areas(22), LLM-driven telemedicine diagnosis and treatment system can make up for the shortage of professional doctors. For example, LLM-based diagnostic tools for skin diseases have achieved over 85% accuracy in primary care. Agent technology also optimizes hospital management by automating the process, reduces the workload of medical staff, and improves the efficiency and accessibility of diagnosis and treatment. Modern scientific research needs to cross the knowledge of multiple disciplines. LLM breaks the disciplinary barriers by integrating the knowledge of biomedicine, chemistry, clinical medicine and other fields. For example, the LLM system developed by researchers successfully predicts three new targeted molecules for rare diseases by combining genomic data and medicinal chemistry libraries (10), or the enhanced LLM combined with knowledge mapping technology can dynamically correlate disease phenotypes, gene mutations and treatment options(23), providing a systematic perspective

for the study of complex diseases.

Personalized medicine and health management. In advancing personalized medicine, LLM can combine genomics with clinical data to tailor treatment for patients. For example, in cancer treatment, LLM recommends targeted drug combinations after analyzing tumor gene mutation characteristics(24), which significantly improves the efficacy. In the problem of chronic disease management(25), Agent technology combined with wearable device data can realize real-time monitoring and intervention of patients with chronic diseases. For example, a diabetes-management Agent(26–28) analyzes glucose fluctuations, diet records, and exercise data to automatically adjust insulin dosing recommendations and send patients personalized health reminders. This pattern reduced HbA1c levels by an average of 1.2% in patients with type 2 diabetes. The multimodal Agent system further integrates imaging and pathological data to achieve early diagnosis and treatment. In the study of Parkinson's disease(29; 30), the treatment recommended by AI system based on the patient's genetic characteristics significantly improved the quality of life. In addition, LLM and Agent are also useful in health risk prediction and prevention. LLM predicts individual disease risk by analyzing large-scale population data. For example, previous work developed a cardiovascular risk prediction model integrated with 100,000 electronic health records (EHRs) to accurately identify 30% of high-risk patients missed by traditional methods(31; 32). It has been reported that the zero-sample learning ability of LLM allows it to adapt to the epidemiological characteristics of people in different regions without additional training(33).

Improve the quality of clinical practice. In clinical application, LLMs, as a clinical decision support tool, can provide doctors with accurate and timely diagnosis and treatment recommendations, which helps to improve the accuracy of diagnosis and the effectiveness of treatment(34). Especially for some rare diseases or complex cases, LLMs can provide more reference information(35) to help doctors make more reasonable decisions and reduce the misdiagnosis rate. The application of Agent technology in the healthcare management system can realize the continuous monitoring and personalized management of patients' health, improve the timeliness and accuracy of medical services, and improve the treatment experience and health of patients. For example, an LLM tool for a skin disease (36) achieved 87% accuracy in primary care, a 22% improvement over traditional methods. Previous studies have shown that multimodal LLM (such as MMedAgent)(37; 38) combined with imaging and pathological data can improve the detection rate of early lung cancer to 93%. Meanwhile, LLM dynamically updates clinical recommendations through real-time analysis of the latest literature and guidelines(39; 40). For example, during the COVID-19 pandemic(41), the CDSS for an infectious disease reduced the update rate of antiviral drug recommendations to 24 hours by analyzing global study data daily(42).

Promoting the development of medical education. LLMs has shown unique advantages in medical education, which can dynamically generate personalized learning materials and assessment tools according to students' learning progress and needs, and improve learning efficiency and effectiveness(43). At the same time, by simulating real clinical scenarios, LLMs provides opportunities for medical students to practice and exercise, which helps to develop their clinical thinking ability and communication skills. LLM can also generate virtual cases and simulate patient conversations to help medical students practice the diagnostic process. Agent technology can also be applied to the simulation system of medical education to create a more realistic virtual patient and clinical environment, and further improve the quality and practicability of medical education. In the rapid development of biomedical science and technology, life-long learning and knowledge updating are equally important. LLM provides real-time knowledge updating services for medical staff(44). In addition, LLM breaks the language barrier and promotes the sharing of global medical knowledge to achieve cross-language medical knowledge dissemination, and disseminate advanced knowledge equitably in countries and regions with different development levels around the world. For example, a multilingual LLM system that translates English guidelines into Kiswahili in real time and incorporates local epidemiologic data to generate adapted versions has significantly improved practice implementation in Africa.

Accelerating Drug Discovery. LLMs can significantly shorten the research and development cycle through target prediction and molecular design(45). Ai-driven drug discovery platforms can reduce the traditional research and development time by 1/10 and cost by 20%. Models such as BioGPT can also mine potential drug associations from the literature and help in the design of clinical trials of new drugs. At the same time, LLM can quickly identify potential drug targets by analyzing protein sequences and medicinal chemistry libraries. For example, the novel antimicrobial peptides generated by ProGen model (46) showed broad spectrum activity in experiments, and the development time was shortened by 60% compared with traditional methods. The ChemCrow system(47) optimized molecular design through LLM, which improved compound screening efficiency by 50%. Agent technology gives LLM the ability to autonomously perform research tasks. Agent developed for the field of drug research and development can automatically generate candidate molecular structures, design synthetic pathways, and iteratively optimize models through experimental data. This "closed-loop research" model shortens the traditional drug-discovery cycle from years to months. The new use of old drugs and drug repositioning have plagued researchers in the biomedical field for a long time. LLM discovers new indications by mining the relationship between existing drugs and disease mechanisms. For example, recently, the LLM system developed by researchers successfully redirected the antidepressant "fluoxetine" for the treatment of glioblastoma by analyzing 100,000 articles(48), with pre-clinical trials showing a 45% reduc-

tion in tumor volume.

Through multi-dimensional innovation, LLM and Agent technology are reshaping the research paradigm, clinical practice and education model in the biomedical field. Its core value lies in data-driven decision-making, efficient utilization of resources and individualized service, which provides a new way to solve traditional medical pain points. However, technological applications still need to evolve in parallel with ethical norms to ensure the sustainable release of their potential.

3 Ethical challenges and negative impacts

Data privacy and data security issues. The application of big language models and Agent technology in biomedicine requires the collection and processing of a large amount of personal health data and biomedical information, which contains sensitive information of patients, such as genetic data, disease history, and treatment records. Once a data breach occurs, it may lead to misuse or unauthorized access to a patient's personal information. For example, when a patient's genetic information is compromised, it may be used for discriminatory purposes, such as insurance companies refusing to cover patients with specific genetic risks or employers refusing to hire employees with certain health problems(49). Previous studies have pointed out that even with anonymization, an attacker may still reverse infer the patient's body through the statistical characteristics of the model output. The autonomous decision-making capabilities that the Agent system prides itself on May bypass traditional data access control mechanisms, leading to unauthorized data access. In addition, LLM in the biomedical field relies on a large amount of patient data (such as electronic health records and genomic information) for training, but there are security risks in the data sharing and storage process. In addition to data breaches, data misuse is also an important problem in the biomedical field. Some organizations or individuals may use patients' health data for commercial purposes or other unethical uses without their consent. Examples include the sale of a patient's genetic data to a third party to develop targeted advertising or drugs, without the patient's knowledge(50). Such data misuse not only violates patients' right to privacy, but can also lead to unnecessary harassment or discrimination of patients. Finally, the risk of synthetic data cannot be ignored, as synthetic patient data generated by LLM can be used maliciously to falsified clinical trial results or insurance fraud. Past studies have shown that malicious attackers can use prompt engineering to induce models to generate false medical records and interfere with the health care regulatory system. Privacy protection is often in conflict with model performance. Enhancing the privacy protection awareness of LLM through supervised fine tuning (SFT)(51; 52), such as automatic filtering of sensitive information, will significantly reduce model performance. Although federated learning can alleviate the risks in data sets, communication security

in inter-agency collaboration still needs further technical protection.

Algorithmic bias and fairness issues. LLMs and agents are trained on biomedical data, which may have biases that lead to unfairness in the results or decisions generated by the models. If the training data are derived primarily from a specific ethnic, sex, or geographic group, the model may yield inaccurate or unjust diagnosis and treatment recommendations when applied to other groups. For example, in facial recognition technology, if the training data mainly contains images of white individuals, the technology may experience high error rates in identifying individuals of other races(53). This algorithmic bias in biomedicine may lead to inequable allocation of medical resources or access to treatment for certain groups. In addition to data bias, the algorithm design itself may also introduce bias. For example, some algorithms may be designed with implicit assumptions against certain groups, or fairness factors may not be sufficiently considered in the optimization objectives. In recent work on LLM for skin cancer diagnosis(54), the sensitivity of the model for dark-skinned patients was 22% lower than that for light-skinned patients, because dark-skinned cases accounted for only 5% of the training data. This can lead to models that adversely affect some groups in the decision-making process, such as prioritizing some specific groups in the allocation of medical resources while ignoring the needs of others. In addition, cultural differences may exacerbate inequalities in the cross-regional application of multilingual LLM, and the autonomy of Agent technology may imply design bias(55). For example, due to the lack of local epidemiological data, the drug recommendation system in Africa defaults to the European and American guidelines, resulting in antimalarial drug dosage recommendations that are not in line with the local patients' constitution(56).

Defects in model interpretability and reliability. The decision-making process of large language models and agents is usually more complex and difficult to understand, which is called "black box" problem(57). In biomedicine, physicians and patients need to understand the basis for diagnosis and treatment recommendations in order to make sound decisions. However, due to the complexity and lack of transparency of the model, it is difficult to explain how the results it generates are arrived at. This opacity may lead to distrust of the technology by both physicians and patients, affecting its widespread use in clinical practice. The application of large language models and agents in biomedicine requires high reliability to ensure the health and safety of patients. However, these models may be affected by various factors, such as data quality, algorithm design, etc., leading them to produce incorrect diagnosis or treatment recommendations in some cases(58). For example, the model may make incorrect predictions due to noise or outliers in the training data, which negatively affects the treatment of patients(59). The illusion problem of LLMS cannot be ignored - LLMs may generate plausible but fact-based content(60), and the closed-loop decision-making capabilities of Agent systems may amplify such risks, such as automatically pre-

scribing unverified prescriptions.

Responsibility attribution and legal disputes. When the application of LLMs and Agent technology in biomedicine leads to adverse consequences, the attribution of responsibility is difficult to be clear. For example, if the diagnostic recommendations provided by the big-language model are wrong and patients are harmed, should the responsibility lie with the company that developed the model, the doctor who used it, or the medical institution? Similarly, when an Agent makes a mistake in the autonomous decision-making process, resulting in damage to the health of patients(61), its responsibility definition also exists in a fuzzy zone. There is a lag in the current legal system in dealing with the new problems brought about by large language models and Agent technology. Existing laws and regulations may not adequately cover the application of these new technologies in biomedicine, resulting in a lack of clear legal basis when dealing with relevant legal disputes. This not only increases the complexity of legal disputes, but also brings uncertainty to the healthy development of technology. Differences in regulatory standards for medical AI across countries may lead to compliance risks. For example, the European Union's AI Act requires high-risk systems to pass strict certification, while some countries allow LLMS to be deployed directly, forming a space for "regulatory arbitrage".

Although the application of big language models and Agent technology in biomedical field has brought many positive impacts(62), it also faces challenges in data privacy and security, algorithm bias and fairness, model interpretability and reliability, as well as responsibility attribution and legal disputes. In order to ensure the healthy and sustainable development of these technologies, it is necessary for all sectors of society to work together to formulate corresponding ethical guidelines, laws, regulations and technical specifications to meet these challenges.

4 Coping strategies and normative suggestions

Establish a strict ethical review and supervision mechanism. The construction and management of ethical review committee should be strengthened to ensure its independence and professionalism. The ethical review committee should be composed of experts in multiple fields, including medicine, ethics, and law, to comprehensively evaluate the ethical risks of technology application. At the same time, a sound review process and standard operating procedures should be established to reduce the subjective and random nature of the review process and improve the objectivity and fairness of the review. The government and relevant regulatory agencies should formulate clear laws and regulations to regulate the application of big language models and Agent technology in biomedicine. Strengthen the supervision of data collection, storage, use and sharing to ensure patient privacy and data security. At the same time, technology developers and users are regularly checked and evaluated to ensure compliance with ethical

guidelines and laws and regulations. Specifically, we can start from the following four aspects:
1. Institutional Review Board (IRB) upgrade(63): All medical AI projects involving LLMs must pass hospital/research institution ethics review, focusing on data source legitimacy, algorithm fairness verification (e.g., population bias analysis by SHAP value)(64), and potential risk disclosure mechanisms. For example, the National Institutes of Health (NIH) requires all AI healthcare projects to submit an "algorithm impact assessment report" detailing the model's predictive differences for patients of different races and genders.
2. National/regional level certification system: referring to the EU Medical Device Regulation (MDR)(65), establish a medical AI grading certification system. High-risk applications (such as cancer diagnostic AI) need to be verified by third-party independent laboratories. For example, the Korean KFDA requires diagnostic AI to complete 1000 blind tests in more than 3 hospitals, with an accuracy of more than 90% and a false negative rate of less than 53. Blockchain storage technology: use distributed ledger to record model training data, parameter adjustment and clinical decision-making process to ensure traceability. For example, IBM Watson Health has embedded a blockchain module in its oncology treatment recommendation system, and any decision modification generates an immutable timestamped record.
4. Real-time monitoring and circuit breaker: deploy anomaly detection system (such as LSTM-based decision deviation warning)(66; 67), and automatically suspend the service when the model output exceeds the preset confidence interval. The UK NHS requires that AI diagnosis and treatment systems must be equipped with a manual review trigger function, and if triggered more than 5 times in a day, the system must be forced offline for maintenance.

Strengthen data quality and diversity management. Strict data quality control standards and procedures were established, and the accuracy, completeness and consistency of the data were strictly reviewed. In the process of data collection, multi-source and multi-channel data collection methods were used to ensure the reliability and representativeness of the data. At the same time, the data were cleaned and preprocessed to remove noise and outliers and improve the availability of the data. Biomedical data from different races, genders, ages and regions were actively collected to avoid algorithm bias caused by data bias. Encourage data sharing and cooperation, integrate data resources from different medical institutions and research institutions, and enrich data samples. In addition, data augmentation techniques can be used to generate diverse data samples and further improve the diversity and coverage of data. Detailed countermeasures can be taken from the following points:
1. Standardization of multimodal data labeling: to develop unified labeling standards for medical imaging, electronic medical records, and genomic data. For example, the American College of Radiology (ACR) has published guidelines for AI-Ready Image datasets, which state that CT images must contain at least 12 structured tags, such as lesion location and pathological stage.

2. Synthetic data augmentation technology: using generative adversarial networks (GAN) to supplement scarce case data. The SynthMed system developed by MIT can generate electronic medical records of rare diseases according to the real distribution, improve the diversity of model training data by 35%, and avoid privacy leakage.
3. Federated learning + homomorphic encryption: medical institutions jointly train models on local encrypted data. The Google Health federated learning platform had connected 300 hospitals to complete the pneumonia prediction model training without sharing the original data(68), and the AUC value of the model reached 0.92.
4. Diversity quota system: mandatory training data covering different gender, age and ethnic groups. The FDA has mandated that no less than 15% of the data from African Americans and no less than 10% from Asians be used in AI models for diabetic retinopathy screening.

Improve the transparency and interpretability of the model. Researchers should focus on the development of various interpretable tools and techniques, such as feature importance analysis, SHAP value calculation, and decision path visualization, to help users understand the decision basis and process of the model. These tools can transform complex model decisions into intuitive and understandable explanations and improve model transparency. A perfect model verification and evaluation mechanism was established to comprehensively evaluate the performance, reliability and fairness of the model. A variety of evaluation indicators, such as accuracy, recall, and F1 score, were used to measure the effect of the model from different perspectives. At the same time, sensitivity analysis and stress test were conducted to ensure the stability and reliability of the model under different scenarios. This can be done by:
(1) Interpretable technology integration
• Hierarchical visualization tools: Develop medical-specific interpretation systems, such as DeepMind's "pathological decision tree mapper," to break down the model's judgment on breast cancer pathology slides into understandable visual features such as cell morphology and staining intensity.
• Natural language inference report: It is required to generate inference chains synchronously when the model outputs diagnostic conclusions. for example, IBM Watson for Oncology provides treatment recommendations with references to the literature that support the recommendation (including PMID numbers and confidence scores).
(2) clinical credibility verification system
• Double-blind controlled trial: comparing the performance of an AI versus a human expert in a real-world medical scenario Mayo Clinic conducted a 10,000-level trial of an AI for ECG diagnosis, requiring that the model interpretation must be approved by more than three cardiologists, otherwise it would trigger retraining.
• Dynamic knowledge update mechanism: a real-time docking system of medical knowledge graph is established to automatically trigger model fine-tuning and

generate update logs when new guidelines are released (such as the update of NCCN Cancer Care standards).

Clear liability framework and legal norms. Formulate clear laws and regulations to define the responsible subjects of big language models and Agent technology in biomedical applications. For example, it stipulates the scope of responsibility of technology developers, users and medical institutions in different situations to avoid buck-passing. When adverse consequences occur, the responsible party can be quickly identified, and effective accountability and compensation can be carried out. According to the characteristics of technological development, the existing legal liability system should be improved, and laws and regulations specifically for artificial intelligence technology should be formulated. Clarify the responsibilities and obligations in data privacy, algorithm bias, model decision-making and other aspects, and provide a clear legal basis for dealing with relevant legal disputes. At the same time, we should strengthen the punishment of illegal acts, increase the cost of illegal acts, and safeguard the authority and fairness of the law. The following responses can be considered:
(1) Rules for grading liability determination
• "Human-machine collaboration" responsibility matrix: Responsibility is divided according to the degree of AI autonomy:
o Auxiliary level (Class B) : the doctor has the main responsibility (such as AI only provides literature reference), and the manual review trace should be recorded.
o Decision level (Class C) : developers share responsibilities with hospitals (such as AI directly prescribing prescriptions), and special liability insurance should be purchased.
The draft EU AI Liability Directive stipulates that Class C systems need to set up a compensation fund, with a maximum compensation of 500,000 euros for a single accident.
(2) Technical support for judicial proof
• Interpretable evidence solidification: Develop explanatory output formats that meet legal evidence standards. China's Online Litigation Rules require that in AI medical disputes, defendants must provide a "verifiable reasoning process" for model decisions, including feature importance ranking and counterfactual analysis results.

Interdisciplinary collaboration and public education. Experts and researchers from multi-disciplinary fields such as computer science, medicine, ethics, and law are encouraged to cooperate to jointly explore the applications and challenges of large language models and Agent technologies in biomedicine. Through interdisciplinary cooperation, solutions can be proposed from different perspectives to promote the healthy development of technology(69). For example, ethicists can provide ethical guidance, legal experts can develop legal norms, and medical experts can evaluate the clinical effects of technologies. Conduct public education campaigns to improve public awareness and understanding

of big-language models and Agent technologies. The basic principles, application fields and potential risks of technology should be popularized to the public through popularization of science, education and training, so as to enhance the public's scientific and technological literacy and self-protection awareness(70). At the same time, the public should be encouraged to participate in the relevant discussion and decision-making process, and their opinions and suggestions should be fully listened to to make the development of technology more in line with the interests and expectations of the public.
(1) Permanent collaborative governance institutions
Medical AI ethics committee: composed of clinicians (30%), AI engineers (25%), ethicists (20%), patient representatives (15%), and legal experts (10%), to approve high-risk projects At Johns Hopkins, complaints about the clinical use of AI dropped 42% after such committees were established.
(2) Public engagement and education programs
Community AI literacy Workshop: Use virtual reality (VR) to simulate AI diagnosis and treatment to help patients understand technical limitations Singapore's HealthHub platform opened an "AI consultation experience pavilion", and users' trust in AI increased by 28% after interaction.
• Multi-language popular science resources: make popular science videos in dialects, such as the series of "10 Questions and 10 Answers for AI Doctors" for rural areas of China, covering 23 dialects and having been played more than 5 million times.

Resource equity and access to healthcare. The government and medical institutions should promote the application of big language models and Agent technology in areas with poor medical resources to improve the accessibility of medical services. Through telemedicine, mobile health and other means, advanced technology is brought to primary medical institutions to provide quality medical services for more patients. For example, intelligent agents can be used to monitor patient health and provide timely medical advice and interventions. In the process of technology application, the fair distribution of medical resources should be ensured to avoid resource inequality caused by technology monopoly. To formulate a reasonable resource allocation policy and allocate medical resources according to the actual needs of patients and the severity of the disease. At the same time, the supervision and evaluation of medical resource allocation should be strengthened to ensure the fairness and transparency of resource allocation.
(1) Lightweight and low-cost deployment
• Edge computing optimization: develop dedicated models with ¡ 100 million parameters that can run on Raspberry PI devices(71). The telemedicine project in Rwanda, Africa, used such models to reduce the cost of CT imaging diagnosis from 50 to 3.
Model as a Service (MaaS) : allows primary hospitals to upload small amounts of local data (e.g., 200 medical records) to quickly customize models. Tencent Miying open platform provides a fine-tuning interface for tuberculosis screening

model, and county-level hospitals can complete the adaptation in 3 hours.
(2) Global resource allocation mechanism
• Open source Medical model Repository: A Hugging Face-like platform for sharing medical models requires that projects receiving public funding must open source the basic model. The Indian government has stipulated that AI diagnostic tools developed by state agencies must be licensed free to low-income countries(72; 73).
• Computing Power Donation Program: Encourage tech companies to target cloud computing resources. Microsoft Azure partnered with the World Health Organization to donate $20 million equivalent computing power to African medical institutions to support the deployment of HIV screening AI.

Through the implementation of the above measures, the negative impact of big language models and Agent technology in biomedical applications can be effectively dealt with, the healthy and sustainable development of technology can be promoted, and more contributions can be made to human health.

5 Conclusion

As cutting-edge achievements in the field of artificial intelligence, big language models and agents have shown great potential in biomedical research, clinical practice, medical education, drug research and development, which provide strong support for promoting medical progress and improving the quality and efficiency of medical services. However, its application is also accompanied by challenges such as data privacy and security, algorithm bias and fairness, model interpretability and reliability, responsibility attribution, and legal disputes. To address these challenges, it is necessary to take several measures: establish a strict ethical review and supervision mechanism to ensure that the application of technology conforms to ethical norms; Strengthen data quality and diversity management to improve the reliability and fairness of the model. Improve the transparency and interpretability of the model, and enhance user trust; Define the liability framework and legal norms to provide legal protection for technology application; Promoting interdisciplinary cooperation and public education to promote the healthy development and social acceptance of technology; We should pay attention to resource equity and universal medical care, so that technology can benefit more people. In the future, it is necessary to continue to explore and improve these countermeasures, balance technological innovation and ethical norms, and ensure the healthy and sustainable development of big language models and Agent technologies in the biomedical field, so as to make greater contributions to human health.

Bibliography

[1] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.

[2] Z. Du, X. Zhaotian, T. Yan, and P. Qin, "Lamper: Language model and prompt engineering for zero-shot time series classification," in *The Second Tiny Papers Track at ICLR 2024*.

[3] Y. Huang, J. Xu, J. Lai, Z. Jiang, T. Chen, Z. Li, Y. Yao, X. Ma, L. Yang, H. Chen, *et al.*, "Advancing transformer architecture in long-context large language models: A comprehensive survey," *arXiv preprint arXiv:2311.12351*, 2023.

[4] Y. Liu, X. Zhong, S. Zhai, Z. Du, Z. Gao, Q. Huang, C. Y. Zhang, B. Jiang, V. K. Pandey, S. Han, *et al.*, "Prompt-enhanced hierarchical transformer elevating cardiopulmonary resuscitation instruction via temporal action segmentation," *Computers in biology and medicine*, vol. 167, p. 107672, 2023.

[5] O. Lieber, B. Lenz, H. Bata, G. Cohen, J. Osin, I. Dalmedigos, E. Safahi, S. Meirom, Y. Belinkov, S. Shalev-Shwartz, *et al.*, "Jamba: A hybrid transformer-mamba language model," *arXiv preprint arXiv:2403.19887*, 2024.

[6] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, X. Cheng, M. Chung, M. Grella, *et al.*, "Rwkv: Reinventing rnns for the transformer era," *arXiv preprint arXiv:2305.13048*, 2023.

[7] S. A. Antu, H. Chen, and C. K. Richards, "Using llm (large language model) to improve efficiency in literature review for undergraduate research.," *LLM@ AIED*, pp. 8–16, 2023.

[8] S. Pal, M. Bhattacharya, M. A. Islam, and C. Chakraborty, "Chatgpt or llm in next-generation drug discovery and development: pharmaceutical and biotechnology companies can make use of the artificial intelligence-based device for a faster way of drug discovery and development," *International Journal of Surgery*, vol. 109, no. 12, pp. 4382–4384, 2023.

[9] C. Wu, Z. Lin, W. Fang, and Y. Huang, "A medical diagnostic assistant based on llm," in *China Health Information Processing Conference*, pp. 135–147, Springer, 2023.

[10] Z. Xie, W. Zhang, X. He, Z. Gao, Z. Du, H. Yang, X. Zhang, R. li, Y. He, L. Peng, and F. Kang, "Crossing the capacity threshold in si-s batteries through mud-crack electrodes," *Energy Storage Materials*, vol. 75, p. 104046, 2025.

[11] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, *et al.*, "Towards generalist biomedical ai," *Nejm Ai*, vol. 1, no. 3, p. AIoa2300138, 2024.

[12] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "Biogpt: generative pre-trained transformer for biomedical text generation and mining," *Briefings in bioinformatics*, vol. 23, no. 6, p. bbac409, 2022.

[13] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, M. G. Flores, Y. Zhang, *et al.*, "Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records," *arXiv preprint arXiv:2203.03540*, 2022.

[14] Z. Du, X. Zhaotian, H. Ying, L. Zhang, and P. Qin, "Cognitive resilience: Unraveling the proficiency of image-captioning models to interpret masked visual content," in *The Second Tiny Papers Track at ICLR 2024*.

[15] Z. Du, C. Jiang, X. Yuan, S. Zhai, Z. Lei, S. Ma, Y. Liu, Q. Ye, C. Xiao, Q. Huang, *et al.*, "Game: Generalized deep learning model towards multimodal data integration for early screening of adolescent mental disorders," *arXiv preprint arXiv:2309.10077*, 2023.

[16] A. Zhao, D. Huang, Q. Xu, M. Lin, Y.-J. Liu, and G. Huang, "Expel: Llm agents are experiential learners," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 19632–19642, 2024.

[17] J. Qiu, K. Lam, G. Li, A. Acharya, T. Y. Wong, A. Darzi, W. Yuan, and E. J. Topol, "Llm-based agentic systems in medicine and healthcare," *Nature Machine Intelligence*, vol. 6, no. 12, pp. 1418–1420, 2024.

[18] W. Shi, R. Xu, Y. Zhuang, Y. Yu, J. Zhang, H. Wu, Y. Zhu, J. Ho, C. Yang, and M. D. Wang, "Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records," *arXiv preprint arXiv:2401.07128*, 2024.

[19] S. Gao, A. Fang, Y. Huang, V. Giunchiglia, A. Noori, J. R. Schwarz, Y. Ektefaie, J. Kondic, and M. Zitnik, "Empowering biomedical discovery with ai agents," *Cell*, vol. 187, no. 22, pp. 6125–6151, 2024.

[20] M. Chen, G. He, and J. Wu, "Zddr: a zero-shot defender for adversarial samples detection and restoration," *IEEE Access*, vol. 12, pp. 39081–39094, 2024.

[21] C. Shyr, Y. Hu, L. Bastarache, A. Cheng, R. Hamid, P. Harris, and H. Xu, "Identifying and extracting rare diseases and their phenotypes with large language models," *Journal of Healthcare Informatics Research*, vol. 8, no. 2, pp. 438–461, 2024.

[22] A. Nag, S. Chakrabarti, A. Mukherjee, and N. Ganguly, "Efficient continual pre-training of llms for low-resource languages," *arXiv preprint arXiv:2412.10244*, 2024.

[23] S. A. A. Naqvi, U. Ayub, M. A. Khan, P. R. Khoury, D. B. Ravichandar, O. H. Witte, D. S. Childs, J. Orme, Y. Zakharia, P. Singh, *et al.*, "Large language models (llms) for inferring genomic characteristics and facilitating genomic literacy in prostate cancer (pca) patients.," 2025.

[24] J. Lammert, T. Dreyer, S. Mathes, L. Kuligin, K. J. Borm, U. A. Schatz, M. Kiechle, A. M. Loersch, J. Jung, S. Lange, N. Pfarr, A. Durner, K. Schwamborn, C. Winter, D. Ferber, J. N. Kather, C. Mogler, A. L. Illert, and M. Tschochohei, "Expert-guided large language models for clinical decision support in precision oncology," *JCO PRECISION ONCOLOGY*, vol. 8, OCT 2024.

[25] C. Tang, R. Zhang, S. Gao, Z. Zhao, Z. Zhang, J. Wang, C. Li, J. Chen, Y. Dai, S. Wang, R. Juan, Q. Li, R. Xie, X. Chen, X. Zhou, Y. Xia, J. Chen,

F. Lu, X. Li, N. Wang, P. Smielewski, Y. Pan, H. Zhao, and L. Occhipinti, "A unified platform for at-home post-stroke rehabilitation enabled by wearable technologies and artificial intelligence,"

[26] N. Baharudin, M.-S. Mohamed-Yassin, A. M. Daher, A. S. Ramli, N.-A. M. N. Khan, and S. Abdul-Razak, "Prevalence and factors associated with lipid-lowering medications use for primary and secondary prevention of cardiovascular diseases among malaysians: the rediscover study," *BMC PUBLIC HEALTH*, vol. 22, FEB 4 2022.

[27] M. Abbasian, Z. Yang, E. Khatibi, P. Zhang, N. Nagesh, I. Azimi, R. Jain, and A. Rahmani, "Knowledge-infused llm-powered conversational health agent: A case study for diabetes patients [arxiv],"

[28] F. Li, Y. Shen, Q. Chen, X. Li, H. Yang, C. Zhang, J. Lin, Z. Du, C. Jiang, C. Yang, *et al.*, "Therapeutic effect of ketogenic diet treatment on type 2 diabetes," *Journal of Future Foods*, vol. 2, no. 2, pp. 177–183, 2022.

[29] A. Wickramarachchi, S. Tonni, S. Majumdar, S. Karimi, S. Koks, B. Hosking, J. Rambla, N. A. Twine, Y. Jain, and D. C. Bauer, "Askbeacon-performing genomic data exchange and analytics with natural language," *BIOINFORMATICS*, vol. 41, MAR 7 2025.

[30] S. Ma, Q. Ye, C. Xiao, H. Guan, Z. Du, and P. Qin, "Alzheimer disease is associated with isotropic ocular enlargement," *arXiv preprint arXiv:2310.11464*, 2023.

[31] M. D. Decker, P. M. Garman, H. Hughes, M. A. Yacovone, L. C. Collins, C. D. Fegley, G. Lin, G. DiPietro, and D. M. Gordon, "Enhanced safety surveillance study of acam2000 smallpox vaccine among us military service members," *VACCINE*, vol. 39, pp. 5541–5547, SEP 15 2021.

[32] Z. Du, H. Luo, X. Li, Y. Liu, X. Yuan, Z. Chen, J. Ji, and P. Qin, "Performance of explainable ensemble learning for mortality risk stratification and multimodal biomarker prediction in colorectal cancer: a retrospective, database cohert study," *The Lancet Regional Health–Western Pacific*, vol. 55, 2025.

[33] C. Chen and T. Stadler, "Genspectrum chat: Data exploration in public health using large language models [arxiv]," *arXiv*, 23 May 2023.

[34] P. Qin, Y. Liu, K. Zhao, L. Luo, Z. Zhang, Z. Qian, C. Jiang, Z. Du, S. Deng, C. Yang, *et al.*, "Diagnosing pathologic myopia by identifying posterior staphyloma and myopic maculopathy using ultra-widefield images with deep learning," 2024.

[35] B. Yang, S. Jiang, L. Xu, K. Liu, H. Li, G. Xing, H. Chen, X. Jiang, and Z. Yan, "Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 4, pp. 1–29, 2024.

[36] J. Zhou, X. He, L. Sun, J. Xu, X. Chen, Y. Chu, L. Zhou, X. Liao, B. Zhang, S. Afvari, and X. Gao, "Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4," *NATURE COMMUNICATIONS*, vol. 15, JUL 5 2024.

[37] B. Li, T. Yan, Y. Pan, Z. Xu, J. Luo, R. Ji, S. Liu, H. Dong, Z. Lin, and Y. Wang, "Mmedagent: Learning to use medical tools with multi-modal agent,"

[38] Z. Du, Q. Shi, J. Lu, Y. Liang, X. Zhang, Y. Wang, and P. Qin, "Majorscore: A novel metric for evaluating multimodal relevance via joint representation," 12 2024.

[39] J. Gallifant, M. Afshar, S. Ameen, Y. Aphinyanaphongs, S. Chen, G. Cacciamani, D. Demner-Fushman, D. Dligach, R. Daneshjou, C. Fernandes, L. H. Hansen, A. Landman, L. Lehmann, L. G. Mccoy, T. Miller, A. Moreno, N. Munch, D. Restrepo, G. Savova, R. Umeton, J. W. Gichoya, G. S. Collins, K. G. M. Moons, L. A. Celi, and D. S. Bitterman, "The tripod-llm reporting guideline for studies using large language models," *NATURE MEDICINE*, vol. 31, JAN 2025.

[40] C. Liu, Z. Lei, L. Lian, L. Zhang, Z. Du, and P. Qin, "Dna virus detection system based on rpa-crispr/cas12a-spm and deep learning," *Journal of Visualized Experiments (JoVE)*, no. 207, p. e64833, 2024.

[41] I. Ahmad, L. Aliyu, A. Khalid, S. Aliyu, S. Muhammad, I. Abdulmumin, B. Abduljalil, B. Bello, and A. Abubakar, "Analyzing covid-19 vaccination sentiments in nigerian cyberspace: Insights from a manually annotated twitter dataset [arxiv],"

[42] N. Rajashekar, Y. Shin, Y. Pu, S. Chung, K. You, M. Giuffre, C. Chan, T. Saarinen, A. Hsiao, J. Sekhon, A. Wong, L. Evans, R. Kizilcec, L. Laine, T. Mccall, and D. Shung, "Human-algorithmic interaction using a large language model-augmented artificial intelligence clinical decision support system," in *CHI '24: Proceedings of the CHI Conference on Human Factors in Computing Systems* (F. Mueller, P. Kyburz, J. Williamson, C. Sas, M. Wilson, P. Dugas, and I. Shklovski, eds.), p. 442 (20 pp.), SIGCHI; SIGACCESS, 2024 2024. CHI '24: CHI Conference on Human Factors in Computing Systems, 2024, Honolulu, HI, USA.

[43] Z. Du and C. Liu, "Rapid response: Prevention and control of medical violence: ethical dilemmas and systemic blind spots beyond technological empowerment," 05 2025.

[44] Z. Du, "Rapid response: Rethinking the tools of science: When "subjective judgments" become the invisible enablers of bias," 03 2025.

[45] I. Gul, C. Liu, X. Yuan, Z. Du, S. Zhai, Z. Lei, Q. Chen, M. A. Raheem, Q. He, Q. Hu, *et al.*, "Current and perspective sensing methods for monkeypox virus," *Bioengineering*, vol. 9, no. 10, p. 571, 2022.

[46] M. Gong, L. Chen, and J. Li, "Progen: Revisiting probabilistic spatial-temporal time series forecasting from a continuous generative perspective using stochastic differential equations,"

[47] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller, "Augmenting large language models with chemistry tools," *NATURE MACHINE INTELLIGENCE*, vol. 6, MAY 2024.

[48] K. R. Laukamp, R. A. Terzis, J.-M. Werner, N. Galldiks, S. Lennartz, D. Maintz, R. Reimer, P. Fervers, R. J. Gertz, T. Persigehl, C. Rubbert, N. C. Lehnen, C. Deuschl, M. Schlamann, M. H. Schoenfeld, and J. Kottlors,

"Monitoring patients with glioblastoma by using a large language model: Accurate summarization of radiology reports with gpt-4," *RADIOLOGY*, vol. 312, JUL 2024.

[49] J. E. McEwen, K. McCarty, and P. R. Reilly, "A survey of medical directors of life insurance companies concerning use of genetic information," *American journal of human genetics*, vol. 53, no. 1, p. 33, 1993.

[50] Z. Du, "Rapid response: After banning corporal punishment, what else do we need to do?," 03 2025.

[51] D. Wu, J. Li, B. Wang, H. Zhao, S. Xue, Y. Yang, Z. Chang, R. Zhang, L. Qian, B. Wang, S. Wang, Z. Zhang, and G. Hu, "Sparkra: A retrieval-augmented knowledge service system based on spark large language model,"

[52] Z. Chen, Y. Deng, H. Yuan, K. Ji, and Q. Gu, "Self-play fine-tuning converts weak language models to strong language models [arxiv],"

[53] S. Cardenas and S. F. Vallejo-Cardenas, "Continuing the conversation on how structural racial and ethnic inequalities affect ai biases," in *2019 IEEE INTERNATIONAL SYMPOSIUM ON TECHNOLOGY AND SOCIETY (ISTAS)* (M. Cunningham and P. Cunningham, eds.), IEEE International Symposium on Technology and Society, IEEE; IEEE Soc Social Implicat Technol, 2019. IEEE International Symposium on Technology and Society (ISTAS), Medford, TX, NOV 15-16, 2019.

[54] E. M. Cai, P. Pathmarajah, R. Daneshjou, J. M. Ko, and A. S. Chiou, "Evaluating the appropriateness of skin cancer prevention recommendations obtained from an online chat-based artificial intelligence model," *JAAD INTERNATIONAL*, vol. 18, pp. 145–147, FEB 2025.

[55] Q. Liu, K. Wang, F. Cheng, and S. Kurohashi, "Assessing large language models in agentic multilingual national bias,"

[56] T. Burki, "Who antimalarial strategy for africa," *LANCET INFECTIOUS DISEASES*, vol. 23, pp. 37–38, JAN 2023.

[57] R. Ajwani, S. Javaji, F. Rudzicz, and Z. Zhu, "Llm-generated black-box explanations can be adversarially helpful,"

[58] M. Bhattacharya, S. Pal, S. Chatterjee, S. S. Lee, and C. Chakraborty, "Large language model to multimodal large language model: A journey to shape the biological macromolecules to biological sciences and medicine," *MOLECULAR THERAPY NUCLEIC ACIDS*, vol. 35, SEP 10 2024.

[59] P. Qin, Z. Du, L. Zhang, S. Zhai, Z. Lei, Y. Zhou, Y. Dongmei, C. Yan, X. Yuan, J. Ji, *et al.*, "Interactive robot with multimodal multitask model for early screening of multiple common adolescent mental disorders," 2025.

[60] Y. Wang, J. Lu, J. Chen, X. Zhang, Y. Liang, Z. Du, Q. Shi, and S.-L. Huang, "Content-style disentangled audio style transfer via diffusion model," 03 2025.

[61] Z. Zhu, B. Wu, Z. Zhang, and B. Wu, "Riskawarebench: Towards evaluating physical risk awareness for high-level planning of llm-based embodied agents,"

[62] Z. Chen, J. Yang, Z. Du, L. Zhang, H. Guan, Z. Lei, X. Zhang, C. Yang, Y. Zhu, Q. Sun, *et al.*, "A conductive and anti-freezing pc-oh organic hydrogel with high adhesion and self-healing activities for wearable electronics,"

[63] B. T. Helfand, A. K. Mongiu, C. G. Roehrborn, R. F. Donnell, R. Bruskewitz, S. A. Kaplan, J. W. Kusek, L. Coombs, K. T. McVary, and M. Investigators, "Variation in institutional review board responses to a standard protocol for a multicenter randomized, controlled surgical trial," *JOURNAL OF UROLOGY*, vol. 181, pp. 2674–2679, JUN 2009.

[64] D. Bouchard, "An actionable framework for assessing bias and fairness in large language model use cases,"

[65] D. Cohen, "Medical device regulation faulty hip implant shows up failings of eu regulation," *BRITISH MEDICAL JOURNAL*, vol. 345, OCT 23 2012.

[66] Z. Lei, L. Lian, L. Zhang, C. Liu, S. Zhai, X. Yuan, J. Wei, H. Liu, Y. Liu, Z. Du, *et al.*, "Detection of frog virus 3 by integrating rpa-crispr/cas12a-spm with deep learning," *ACS omega*, vol. 8, no. 36, pp. 32555–32564, 2023.

[67] L. Zhang, Z. Lei, C. Xiao, Z. Du, C. Jiang, X. Yuan, Q. Hu, S. Zhai, L. Xu, C. Liu, *et al.*, "Ai-boosted crispr-cas13a and total internal reflection fluorescence microscopy system for sars-cov-2 detection," *Frontiers in Sensors*, vol. 3, p. 1015223, 2022.

[68] S. Sharma and K. Guleria, "A comprehensive review on federated learning based models for healthcare applications," *ARTIFICIAL INTELLIGENCE IN MEDICINE*, vol. 146, DEC 2023.

[69] L. Zhang, Z. Chen, H. Ying, Z. Du, Z. Song, J. Chen, X. Yuan, C. Yang, V. Pandey, C. Y. Zhang, *et al.*, "A 3.55-μm ultrathin, skin-like mechanoresponsive, compliant, and seamless ionic conductive electrode for epidermal electrophysiological signal acquisition and human-machine-interaction," 2024.

[70] Z. Du, "Rapid response: When "scientific blood transfusion" meets policy blood loss: the neglected ecological chain of research," 05 2025.

[71] I. Gul, S. Zhai, X. Zhong, Q. Chen, X. Yuan, Z. Du, Z. Chen, M. A. Raheem, L. Deng, E. Leeansyah, *et al.*, "Angiotensin-converting enzyme 2-based biosensing modalities and devices for coronavirus detection," *Biosensors*, vol. 12, no. 11, p. 984, 2022.

[72] X. Zhang, K. Fan, Y. Wang, Y. Liang, J. Lu, Z. Du, Q. Shi, and P. Qin, "Tagmo: Temporal control audio generation for multiple visual objects without training," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2025.

[73] Q. Shi, Z. Du, J. Lu, Y. Liang, X. Zhang, Y. Wang, J. Peng, and K. Yuan, "Audiocache: Accelerate audio generation with training-free layer caching," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2025.