Grenze International Journal of Engineering and Technology, June Issue

CUBA: A Conversational AI for Medical Education

Somesh Sharma[1], Nisha Wankhade[2], Divyansh Katakwar[3], Yash Kedar[4], Chaitrali Despande[5] and Hardik Thakre[6]
[1-6]Yeshwantrao Chavan College of Engineering, Nagpur, India
Email: sharmasomesh2611@gmail.com, nisha.wankhade@gmail.com, divyanshkatakwar8@gmail.com, yashkedar19@gmail.com, chaitralid1704@gmail.com, hardikthakre830@gmail.com

***Abstract*— CUBA is a Large Language Model based Virtual Patient Simulator designed as a training system for medical students to improve their communication skills and practice on their diagnostic reasoning skills through interactive conversations with the virtual patients. The system is build on two strong base models namely ClinicalBERT and Phi-3-mini. ClinicalBERT, fined tuned on medical questions, which is used to check the relevance of questions asked by the interns. It reached an accuracy of 97.8% and F1 score of 98.7%. Only relevant queries are forwarded to the Phi-3-mini model, which is fine-tuned using LoRA (training loss 0.29) to generate realistic patient responses. The System integrates voice-based interaction and multilingual support using the Google Translate API. After each conversation, interns are required to predict the patient primary diagnosis, and the system then provides feedback regarding the correctness of prediction and gives a explanatory guidance. Overall, CUBA offers a safe and interactive environment for improving AETCOM, clinical questioning, and diagnostic reasoning skills among medical trainees.**

***Index Term*— API, Chatbot, ClinicalBERT, CUBA, Diagnostic Feedback, Firebase Authentication, Fine-Tuning, LLM, Medical Interns, Multilingual, Patient-Doctor Interaction, Phi-3-mini, VPS.**

I. INTRODUCTION

The development of clinical expertise in medical education is heavily dependent on repeated patient interaction, where students have to learn to gather symptoms, communicate effectively, and make diagnostic decisions. However, real-world clinical exposure is often limited due to constraints such as patient availability, time restrictions, and ethical concerns involving inexperienced trainees. Consequently, many students struggle to practice structured questioning and clinical reasoning at a good level. Virtual patient-based training has been shown to address these limitations by providing a controlled and repeatable learning environment [1]. Such systems also introduces a wider variety of clinical cases that may not be encountered during routine hospital training.

Recent improvement in Artificial Intelligence (AI) and Natural Language Processing (NLP) have enabled conversational systems to simulate human-like dialogue. Large Language Models (LLMs), in particular, can maintain context and generate coherent responses, making them suitable for interactive healthcare applications [2], [3]. However, most existing systems are designed for patient assistance rather than medical training and often lack mechanisms to ensure clinical accuracy and relevance [4], [5].

Grenze ID: 01.GIJET.12.2.213_24

Additionally, LLMs are prone to generating inaccurate or fabricated responses, known as hallucinations, and generally do not evaluate whether user queries are clinically appropriate [6]. These limitations highlight the need for systems that combine conversational intelligence with domain-specific validation.

To address these challenges, this work proposes a Virtual Patient Simulator that integrates conversational generation with clinical validation. The system combines Phi-3-mini, a lightweight language model for generating context-aware responses [7], with ClinicalBERT, a domain-specific model for identifying medically relevant queries [8]. Furthermore, parameter-efficient fine-tuning using Low-Rank Adaptation enables effective domain adaptation with reduced computational cost [9]. The proposed system provides an end-to-end training workflow, including real-time interaction, automated patient summarization, and diagnostic feedback, thereby supporting the development of clinical communication and reasoning skills in a scalable and reliable manner. This approach ensures that the system not only generates realistic interactions but also maintains clinical relevance throughout the training process.

II. Related Work

The use of Artificial Intelligence in healthcare has progressed from simple rule-based systems to advanced conversational agents capable of generating dynamic responses. Early healthcare chatbots were primarily designed to handle basic tasks such as symptom checking and appointment management, but they lacked the ability to understand context or respond empathetically, which limited their effectiveness in realistic clinical scenarios [4].

The introduction of Large Language Models has significantly improved the capabilities of conversational systems by enabling more natural and context-aware interactions. However, recent research highlights several limitations associated with these models, including the generation of inaccurate information, lack of interpretability, and concerns related to safety and trust in medical applications [2], [3], [6]. These challenges indicate that standalone LLMs are insufficient for reliable healthcare use without additional control mechanisms.

To improve efficiency and scalability, lightweight models such as Phi-3-mini have been explored as alternatives to large-scale architectures. These models offer competitive performance while reducing computational requirements, making them suitable for real-time applications [7]. Additionally, alignment strategies have been proposed to improve model reliability, and parameter-efficient tuning approaches such as LoRA and P-Tuning v2 enable effective adaptation to domain-specific tasks without full retraining [9], [10], [11].

At the same time, domain-specific models like ClinicalBERT have demonstrated strong performance in processing medical text by leveraging pretraining on clinical datasets. These models are particularly effective for tasks such as classification and information extraction, as they capture the nuances of medical terminology and context [8], [12], [13]. However, their application is typically limited to isolated NLP tasks and they are not commonly integrated into conversational frameworks. In the context of medical education, virtual patient systems have been shown to support the development of communication skills and clinical reasoning by providing simulated interaction environments [1]. Despite their benefits, many existing systems lack adaptive conversational capabilities and do not provide structured feedback based on user performance. Similarly, recent chatbot systems designed for healthcare applications focus on accessibility and scalability but often fail to incorporate mechanisms for validating user input or guiding diagnostic reasoning [5], [14]. Ethical and practical considerations also play a critical role in the deployment of AI systems in healthcare. Issues related to transparency, safety, and responsible use have been widely discussed, emphasizing the importance of designing systems that balance automation with reliability and user control [6], [15].

Overall, while significant progress has been made in both conversational AI and clinical NLP, existing approaches remain fragmented. Current systems either focus on response generation without validation or perform classification without interaction. The proposed work addresses this gap by integrating both capabilities into a unified system that supports structured, realistic, and feedback-driven medical training.

III. Methodology

The proposed Virtual Patient Simulator (CUBA) is designed as an end-to-end AI-driven system that enables interactive and relevance-aware doctor–patient conversations for medical training. The system integrates speech processing, natural language understanding, conversational response generation, and diagnostic evaluation into a unified pipeline. The overall system architecture is shown in ‘Fig 1’.

The system operates in a sequential manner. The user provides input through speech, which is converted into text and processed by a relevance classification module. The ClinicalBERT model determines whether the query is

medically relevant. If the query is valid, it is passed to the Phi-3-mini model for response generation; otherwise, the system prompts the user to rephrase the input. The generated response is converted back into speech to simulate a realistic patient interaction. At the end of the session, the system extracts clinical information and evaluates the user's diagnostic prediction to provide feedback.

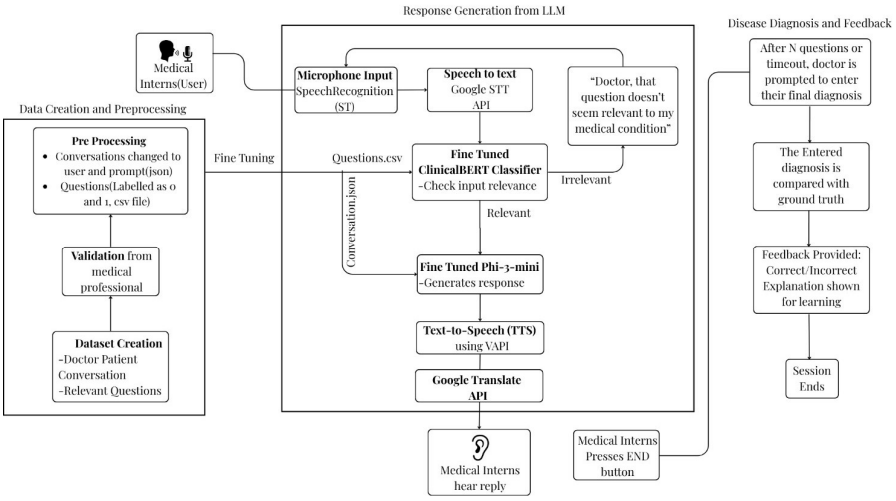

Fig. 1. System Architecture

A. Data Creation and Validation

Two datasets were used to train the system. The first dataset consists of 915 labeled doctor-style questions used for training the relevance classifier. The second dataset contains multi-turn doctor–patient conversations, which are converted into prompt–response pairs for conversational fine-tuning. Since publicly available datasets were not fully aligned with the required clinical context, a synthetic dataset was created and validated by medical experts. This ensures that the training data reflects realistic clinical scenarios and improves the reliability of the system.

B. Model Selection

After thorough analysis of multiple large language model it was found that Phi-3-mini and ClinicalBERT best satisfies the requirements of the Virtual Patient System. Both these models are freely available and are compatible with each other. Phi-3-mini is a small language model that with 3.8 billion parameters. Even though small it delivers performance com- parable to large models and is effective and efficient in terms of response generation and is easy to run on mobile devices. ClinicalBERT on the other hand is trained on clinical and medial text data to improve NLP task in healthcare. It inherits the BERT architecture but is fine tuned on medical dataset. It is used as classifier to segregate the input as relevant and irrelevant. Both of these models are very useful and can be fine tuned freely, hence is used in the patient simulation software.

C. Relevacnce Classification using ClinicalBERT

The input query is represented as a token sequence:

$$X=(x_1, x_2, x_3, ..., x_n) \quad \text{....(1)}$$

where X denotes the tokenized representation of the input query. The sequence is then encoded using ClinicalBERT to obtain contextual embeddings:

$$H= BERT(X) \quad \text{....(2)}$$

From this representation, the embedding corresponding to the [CLS] token is used for classification. The relevance score y is computed using (3):

$$y=\sigma(W\cdot h_{[CLS]}+b) \quad \text{....(3)}$$

where σ is the sigmoid function. The final prediction $\hat{y}$ is obtained using the thresholding rule in (4):

$$\hat{y}=\begin{cases}1, & if\ y\geq\tau\\ 0, & otherwise\end{cases} \quad \ldots.(4)$$

This formulation ensures that only medically relevant queries are passed to the response generation module.

D. Response Generation (Phi-3-mini)

For a relevant query Q and conversation context C, the response R is generated by maximizing the conditional probability, as expressed in (5):

$$R=\arg\max_{r} P(r|Q,C) \quad \ldots.(5)$$

The probability distribution is modeled autoregressively according to (6):

$$P(r|Q,C)=\prod_{t=1}^{T} P(r_t|r_{<t},Q,C) \quad \ldots.(6)$$

This formulation enables the generation of coherent and context-aware multi-turn responses.

E. Parameter Efficient Fine Tuning using LoRA

Fine-tuning is the process of adapting pre-trained machine learning models using smaller, task-specific datasets so they perform better in a particular domain. In this system, fine-tuning helps the models understand medical conversations more accurately. To efficiently adapt the model, LoRA modifies the weight matrix as $W'=W+\Delta W$ where the update term is given by (7):

$$\Delta W=A\cdot B \quad \ldots.(7)$$

Here, $A\in R^{(d\times r)} and B\in R^{(r\times k)}$, with $r\ll d$, enabling efficient parameter updates with reduced computational overhead.

F. Conversational Pipeline

The conversational pipeline of the system operates through a structured, multi-step process designed to support seamless interaction. Initially, the acoustic speech input "{S}" is transcribed into text "{T}" . This is followed by an optional translation phase, where the initial text may be translated to English "{" "T" _"en" "}" if necessary. To ensure the system accurately interprets domain-specific clinical nuances, relevance classification is then performed utilizing ClinicalBERT. Subsequently, the Phi-3-mini model processes the classified data to generate an appropriate, context-aware response. Finally, text-to-speech synthesis converts the generated text back into an audible format. Overall, this robust pipeline ensures continuous interaction while successfully maintaining contextual consistency across multiple dialogue turns.

G. Feedback Mechanism

The Virtual Patient Simulator helps students evaluate their diagnostic reasoning after interacting with the virtual patient. When the student feels enough questions have been asked, they can press the “End” button, after which the system automatically processes the conversation and stores it in the database with a unique patient ID. Important patient details such as name, age, gender, and symptoms are extracted from the conversation and displayed on a patient summary page. Based on this information, the student is asked to predict the most likely disease. The system then compares the predicted diagnosis with a clinical diagnosis–symptom database and immediately informs the student whether the answer is correct or incorrect. Each attempt is recorded in the database so that the student’s performance can be monitored and feedback can be provided to improve their diagnostic reasoning skills.

H. Details System Implementation and Integration

The system is implemented using a modular architecture that integrates authentication, conversational processing, and data management. A secure authentication layer is implemented using Firebase, where each user is assigned a unique identifier and validated through token-based verification. This ensures controlled access and secure storage of user-specific session data. To support real-time interaction, multiple APIs are integrated within the system. Speech-to-text and text-to-speech modules enable natural voice-based communication, while translation services allow multilingual interaction. Additional APIs handle conversation storage, patient data extraction, and session retrieval. The modular API-based design ensures scalability and flexibility, allowing independent components to interact seamlessly. By integrating these components into a unified pipeline, the system achieves efficient, secure, and real-time operation of the Virtual Patient Simulator.

IV. RESULT

A. Model Performance and Analysis

Both language models were evaluated using standard training and classification metrics. As illustrated in 'Fig. 3', the Phi-3-mini model achieved a final training loss of 0.29, indicating effective learning of conversational patterns from doctor–patient dialogues. The steady decline in training loss demonstrates stable convergence without significant overfitting, suggesting that the model can generalize well to unseen clinical conversations.

Fig. 3. Training loss during fine tuning Phi-3-mini

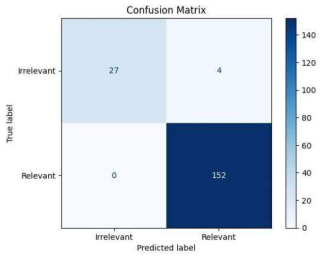

Fig. 4. Confusion Matrix of ClinicalBERT Relevance Classifier

The confusion matrix in 'Fig. 4' and 'Fig. 5' shows that the ClinicalBERT relevance classifier correctly identified 152 relevant and 27 irrelevant questions, no relevant questions were misclassified as irrelevant (zero false negatives), which is critical in medical training scenarios where missing a clinically valid query could negatively impact learning. Only 4 irrelevant questions were incorrectly classified as relevant, indicating a very low false positive rate. The model achieved an accuracy of 97.8% and an F1-score of 98.7%, demonstrating strong classification performance. Compared to a baseline TF-IDF + Logistic Regression model with an F1-score of approximately 86%, the proposed approach shows significant improvement, highlighting the effectiveness of contextual embeddings and domain-specific pretraining.

Error analysis indicates that misclassifications mainly occur in ambiguous or partially relevant queries, suggesting scope for further improvement in handling borderline conversational inputs.

In addition to quantitative results, qualitative evaluation shows that the fine-tuned Phi-3-mini model generates context-aware and clinically relevant responses, maintaining conversational continuity across multiple dialogue turns. Compared to the base model, it produces more structured responses with improved symptom coherence and reduced irrelevant outputs, demonstrating its effectiveness in simulating realistic patient interactions.

B. User Interface Workflow

The implemented Virtual Patient Simulator provides a secure user workflow that begins with a registration interface where users create accounts by entering their name, email, and password. The backend validates and stores these credentials to ensure secure access. After registration, users can log into the system and access the simulation environment. Once authenticated, they are directed to the main dashboard as shown in 'Fig. 7', where they can start a simulation session. During the simulation, the real-time conversation interface in 'Fig. 8' allows trainees to interact naturally with the virtual patient through conversational dialogue. At the end of the interaction, the system automatically generates a patient summary page as shown in 'Fig. 6', which displays clinical information extracted from the session such as symptoms, severity, and diagnosis. The "Check Prediction" feature enables users to verify their diagnosis and receive instant feedback, helping them improve their diagnostic reasoning skills.

		Predicted: True	Predicted: False	
Actual	True	True Positive(TP) 152	False Negative(FN) 0	Recall=Sensitivity TP/(TP+FN) 1.0
	False	False Positive(FP) 4	True Negative(TN) 27	Specificity TN/(TN+FP) 0.871
		Precision TP/(TP+FP) 0.9744		Accuracy (TP+TN)/(TP+TN+FP+FN) 0.978 F1 (2*(Precision x Recall))/ (Precision+Recall) 0.987

Fig. 5. Performance Metrics of fine-tuned ClinicalBERT

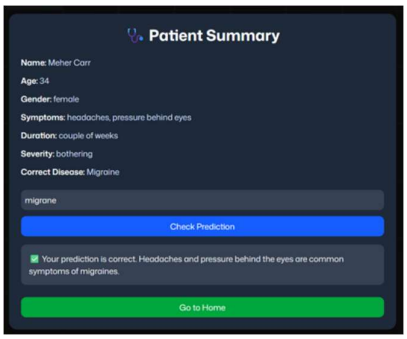

Fig.6. Patient Summary Page

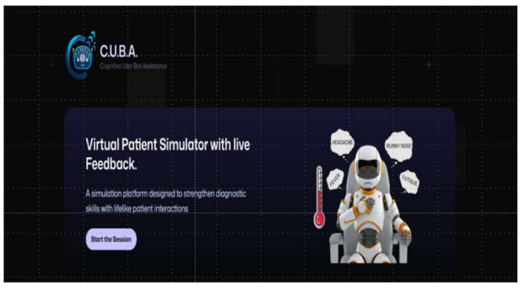

Fig. 7. Home Page

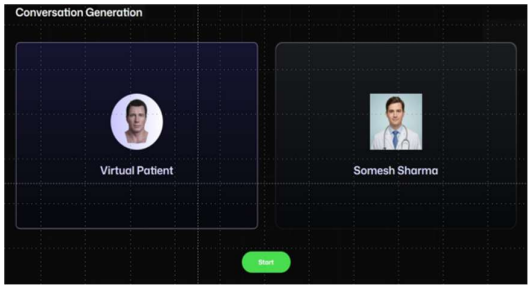

Fig. 8. Conversation Page

C. User Feedback and System Validation

The Virtual Patient Simulator was evaluated by medical interns and reviewed by medical professionals to assess its usability and effectiveness. Around 80% of interns provided positive feedback, stating that the system offered realistic patient interactions and context-aware responses during simulations. The user interface was considered simple and easy to navigate, allowing interns to use the system without difficulty. Interns also appreciated the instant feedback mechanism, which helped them analyze their diagnostic reasoning and improve their clinical thinking skills. Some users suggested adding more diverse clinical cases, including rare diseases, and incorporating performance analytics or a progress tracker to monitor learning across multiple sessions.

Medical professionals also reviewed the system to evaluate its clinical accuracy. They confirmed that the responses generated by the virtual patient were logical, medically relevant, and consistent with the conversation context. The symptoms provided by the virtual patient were appropriately linked to related diseases, helping students think about possible diagnoses. Experts noted that the system can be particularly useful for early-year medical students to improve their communication and patient questioning skills. They suggested referring to the predicted diagnosis as a “query” rather than a final diagnosis, since real diagnoses require medical tests and physical examination. They also recommended expanding the system with subject-specific cases such as cardiology and orthopedics. Overall, both interns and medical professionals agreed that the simulator is a useful and realistic training tool for developing communication and diagnostic reasoning skills in a safe learning environment.

V. Conclusion

This paper presented a Virtual Patient Simulator, an AI-driven educational system designed to enhance clinical training for medical students. The proposed approach integrates Phi-3-mini for generating context-aware patient responses with ClinicalBERT for validating the clinical relevance of user queries, ensuring structured and meaningful interactions. In addition, the system incorporates secure user authentication using Firebase and a feedback mechanism that evaluates diagnostic predictions against a medical knowledge base. The simulator supports real-time, multilingual voice-based interaction, enabling an immersive and accessible learning experience. By combining conversational intelligence with domain-specific validation and feedback, the system provides a safe and scalable environment for practicing clinical questioning, communication, and diagnostic reasoning. The results demonstrate that the proposed system effectively improves interaction quality and supports the development of essential clinical skills, making it a valuable tool for modern medical education.

FUTURE SCOPE

Future work will focus on enhancing the system's ability to evaluate communication skills by incorporating real-time emotion and behavior analysis. Techniques such as voice tone analysis, response pattern evaluation, and facial expression recognition can be integrated to assess empathy, clarity, and patient engagement during interactions. Additionally, adaptive learning mechanisms can be introduced to generate personalized patient cases based on a student's performance and interaction history, enabling targeted improvement. The system can also be extended to include specialized medical domains such as psychiatry, pediatrics, cardiology, and emergency medicine, allowing users to practice a broader range of clinical scenarios. Further improvements may include enhanced handling of ambiguous queries and the integration of performance analytics dashboards to track user progress over time.

REFERENCES

[1] T. Poulton and C. Balasubramaniam, "The effectiveness of virtual patients in medical education: A meta-analysis," Journal of Medical Internet Research, vol. 19, no. 5, pp. 1–10, 2017.

[2] M. Khan, M. Hossain, K. Fatema, N. Fahad, S. Sakib, M. Jannat, J. Ahmad, M. Ali, and S. Azam, "A review on large language models: Architectures, applications, taxonomies, open issues and challenges," IEEE Access, Feb. 2024, doi: 10.1109/ACCESS.2024.3365742..

[3] S. Montagna, G. Aguzzi, S. Ferretti, and M. Pengo, "LLM-based solutions for healthcare chatbots: A comparative analysis," in Proceedings of the IEEE International Conference on Telemedicine and E-health Evolution, 2024, doi: 10.1109/PerComWorkshops59983.2024.10503257.

[4] Y. Xu et al., "Systematic review of chatbot applications in healthcare and oncology," Journal of Medical Internet Research, vol. 23, no. 11, pp. 1–15, 2021.

[5] R. Kumar, P. Mehta, and S. Ghosh, "Conversational AI chatbot for healthcare," in Proceedings of the IEEE International Conference on Smart Healthcare Innovations, 2023.

[6] Y. Huo et al., "Large language models for chatbot health advice studies: A systematic review," JAMA Network Open, vol. 8, no. 2, p. e2457879, Feb. 2025.

[7] A. Brogly, J. Hong, S. Guo, and J. Carroll, "Evaluating the small language model Phi-3-mini on medicine, health, and sports injury news headlines," arXiv preprint, 2024.

[8] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in Proceedings of the 2nd Clinical Natural Language Processing Workshop, Minneapolis, MN, USA, Jun. 2019, pp. 72–78, doi: 10.18653/v1/W19-1909.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in Proceedings of the International Conference on Learning Representations (ICLR), Apr. 2022, doi: 10.48550/arXiv.2106.09685.

[10] E. Haider et al., "Phi-3 safety post-training: Aligning language models with a 'break-fix' cycle," arXiv preprint arXiv:2407.13833, 2024, doi: 10.48550/arXiv.2407.13833.

[11] X. Liu, K. Ji, Y. Fu, W. L. Tam, Z. Du, Z. Yang, and J. Tang, "P-Tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks," arXiv preprint arXiv:2110.07602, Oct. 2021, doi: 10.48550/arXiv.2110.07602.

[12] D. Mithun, U. B. Sherkhane, A. K. Jha, and S. Shah, "Transfer learning with BERT and ClinicalBERT models for multiclass classification of radiology imaging reports," Research Square Preprint, 2024.

[13] J. Varghese, "ClinicalBERT—A tool for processing clinical texts," GYANVESHAN Journal, Muthoot Institute of Technology and Science, 2024.

[14] S. Shetty, A. K. Verma, and R. Desai, "AI-based healthcare chatbot for resource-constrained settings," in Proceedings of the International Conference on Artificial Intelligence and Applications, 2022.

[15] N. Saadat, K. Raza, and M. Farooq, "Responsible AI in telehealth: Balancing autonomy and patient agency," IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3381125.

[16] A. Pourkeyvan, R. Safa, and A. Sorourkhah, "Harnessing the power of Hugging Face transformers for predicting mental health disorders in social networks," IEEE Access, vol. 12, pp. 1–10, Feb. 2024, doi: 10.1109/ACCESS.2024.3366653.

[17] J. Binfet, A. Kumar, L. Zhang, and M. Davis, "Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation," npj Mental Health Research, vol. 5, no. 2, pp. 101–115, 2024, doi: 10.1038/s44184-024-00097-4.