International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Real-time LLM-driven telephony conversational agent for healthcare insurance workflows

Sudharson D [1], Shruthi P [1,*], Afsheen Zaahrah A [2], Tejaswini SS [2] and Karthick A [2]

[1] *Head of Department, Department of AI and DS, Kumaraguru College of Technology, Tamil Nadu, India.*
[2] *Department of AI and DS, Kumaraguru College of Technology, Tamil Nadu, India.*

International Journal of Science and Research Archive, 2026, 19(01), 414-422

Publication history: Received on 27 February 2026; revised on 06 April 2026; accepted on 08 April 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.19.1.0687

Abstract

This research paper gives a voice bot for healthcare domain which utilizes the pipe cat framework for agent orchestration along with Twilio for calling provider / configuration for telephony that encapsulates STT, for conversation management LLM, TTS layer for the conversational AI in real time, state management. At the media layer, stands the Livekit WebRTC for reduced latency and the streaming of audio in bidirectional that achieves about sub 500 ms with an end-to-end response time through SIP trunking.

and Twilio programmable voice API. This bridges the legacy PSTN along with cloud native telephone configuration to optimize the workflow of clinical and the engagement of patients.

Keywords: Conversational AI; Healthcare Communication; Telephony Systems; Large Language Models (LLM); Speech Recognition; Dialogue Management

1. Introduction

This article represents a conversational workflow system also known as real-time conversational phone-call AI agent that can communicate about the healthcare insurance claim to the patients. This AI agent helps reduce the workload of the patients by calling a hospital or insurance office and waits in a queue instead of a human in line. The proposed AI agent calls the users, primarily the patients in the hospital and initiates a natural conversation with them, like a human executive who can assist the patients with the issues faced by them. The agent can identify the patients, collecting details about the insurance claims and updates the claim status to the patients. It can also request missing documents and details from the patients and finally get a heads-up from the users to keep the documents updated. The key innovation is to build a structured innovative AI caller agent that behaves like a healthcare support executive over a real phone call with the users. Also, the entire interaction happens over a normal phone call for which the patient does not need an app, internet, or smartphone. The AI Agent was built with a multi layered architecture with the flow-based dialogue management which is key in tracking the conversational AI agent to address the workflow failures and so on.

2. Related work

2.1. Healthcare Conversational Agents

Conversational agents have been extensively used in improving patient interaction and access to information regarding specifics in healthcare. Prior studies have shown exceptional accuracy and have reduced the workload in healthcare, resolving patient clarifications and medical Question-Answering [1]. Some systems such as MarIA have also been

* Corresponding author: Shruthi P

supportive in monitoring chronic diseases such as diabetes [2]. These systems enable asynchronous communication, needless of the physician's availability [3]. However, most of the architectures support only message-driven platforms and chatbot-style clarifications. These solutions require digital literacy, where elderly or individuals with lower familiarity in technology are deprived of such solutions.

2.2. Voice Assistants in Healthcare

Voice based assistants are proposed as an enhanced and natural form of human-digital assistant interaction. Systems with voice enabled features are seen in medical self-diagnosis, robust with identifying intent recognition errors and intent misfires [4]. Generative AI Voice agents have showcased prominent accuracy in multiple fields of healthcare including appointment scheduling, insurance verification and handling billing questions [5]. Although VA proved their readiness during pandemic, supporting symptom screening and healthcare delivery by reducing effort and enabling accessibility during crisis [6], these kinds of systems operate over mobile applications or smart speakers.

2.3. Telephony and IVR in Healthcare

Telephone communication stands as a primary mode for administrative tasks such as appointment reminders, follow-ups and insurance clarifications in healthcare. Automated IVR with a long haul of choosing options with a predefined menu, severely lacks contextual understanding [7]. IVR and phone-based survey systems exist with a structured health questionnaire, enabled with audio recordings from patients making it highly autonomous in the field [8]. Prior studies have shown introduction of virtual psychological assistants, with solid technical frameworks such as Twilio for call routing and WaveNet for speech-to-text, which although enables internetless patient access, has limitations in terms of latency and noise suppression which can lead to misunderstanding of words [8].

2.4. LLM Medical Communication

The recent advancements in large language models have increasingly shown improvement in conversational reasoning and context setting in patient-computer healthcare related conversations. These systems work exceptionally well in understanding intent, often showcasing significant accuracies up to 85.2%, highlighted in a study for understanding patient calls in cancer centres [9]. Voice calls with LLM integrated have shown that LLM ensemble can handle complex, multi-turn medical dialogues while keeping patient safety a priority [10].

2.5. Real-Time Voice AI

Studies as of late have started to explore real-time conversational voice agents where telephony infrastructure is integrated with large language models. The latest work on emergency communication systems established that LLM-powered workflows can facilitate live call audio , reconstruct speech content and classify call urgency[11] . Patient monitoring and routine check-ins in health care settings are being done by telephone based agents . Initial deployments indicate that patients are open to engaging with AI-based callers for follow up communications [12].Subsequent implementations have demonstrated telephony pipelines which combines call routing along with speech recognition and speech synthesis to allow voice interaction with patients directly [13].

2.6. Research Gap

Existing research has explored healthcare conversational agents, telephony interfaces, large language model - based systems. These systems show enhanced patient engagement, accessibility and autonomous assistance in healthcare communication. However, these architectures exist in silos, working in isolation and don't operate as a solid unified solution.

Healthcare admin-based workflows such as insurance claim communication needs a structured, planned framework involving verification, capturing information and clarifying context. While existing voice agents primarily work in conducting surveys, engaging in medical dialogues, this paper manages a solution which emphasizes goal-oriented conversational flow, especially in the field of healthcare insurance claims.

3. System architecture and conversational flow

3.1. Twilio Configuration for Telephony

Communication starts with an outbound call is initiated from the system to the receiver's phone with the help of Twilio's programmable voice API, that gives a concrete endpoint for the telecom carriage and the infrastructure of PSTN. The calling provider is responsible for setting up the call , signaling of SIP and ensure if the call is established across networks of landline and mobiles.

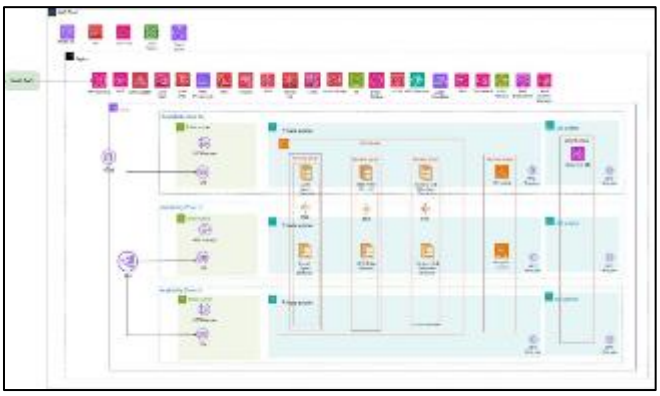

Figure 1 Voice bot AWS architecture

3.2. SIP trunking

After the call has been established, the RTP audio packets are sent to the calling provider (Twilio) via the sip configuration that relates to the telephony provider and the media server/ WebRTC. The Sip Trunk is used to interconnect for signaling and sending RTP packets for bidirectional transport of audio, handle codec configuration (use G.711) and securing the transport with TLS/SRTP whenever required using SIP. To ensure that the RTP packets are sent to the media layer , proper configuration of ports , codecs , traversal of NAT are required.

3.3. WebRTC - Livekit

At the media layer, the SIP/RTP is handled by the WebRTC media layer - Livekit , it serves as a low-latency bridge for audio in between the agent (voicebot) and the telephony configuration. The RTP audio packets are ingested by the Livekit and is encapsulated by the livekit with the codec that is compatible with webRTC sudio-tracks and it is exposed to a channel of signalling for the agent to publish audio streams .During the return of audio packets , the voicebot's voice that is synthesized is sent to the livekit as a stream of WebRTC , then converted into RTP and it is then forwarded over the trunk of SIP to Twilio , to close the loop of media

3.4. Pipecat agent orchestrator for conversation and management of state

The agent framework is above the WebRTC media layer and it is responsible for orchestrating the conversation, from receiving the incoming track of audio packets, coordinating on what the system should process, think and communicate. It is a agent pipeline that handles the continuous audio streams for publishing it to Speech - To-Text.

, managing the endpoint for the user availability and brings up a behavior where the agent cannot speak if interrupted by the user. This agent framework is responsible for managing the single state of session which includes history management, extracts sip headers and main the level of workflow that is carried out now and using that state to create input for the language model as prompt. The LLM's output is used for applying processing the text that is sent to the TTS layer. Pipecat agent framework can also maintain the per conversation turn, listen to calls and handle the events proactively , silence handling , interruption handling, hangup call when participant disconnected and transition the agent into a relevant state.

4. Implementation

The system has proposed an AI Voice Bot system built in python 3.10+ using an asynchronous, event-driven, architecture. Fundamentally we have six principal layers - API Layer, Bot Pipeline, Service Logic, External Clients, Data Models and Shared Utilities. We are using FastAPI as the HTTP layer; real time audio processing pipelines are migrated from Pipecat. All the inter component communication is non-blocking, realized through Python's asyncio primitives to facilitate low latency environments.

4.1. Voice Processing Pipelines

The voice pipeline is built using VoicePipelineFactory, a component that assembles provider specific STT(Speech to text), LLM(Large Language Models), and TTS(Text to speech) processors into a unified Pipecat Pipeline. Audio frames are captured and processed using LiveKit, There are 5 stages: Inbound audio reception, STT Transcription, LLM Inference, TTS Synthesis and Outbound Audio Delivery.

Deepgram and Sarvam AI are the tools used to facilitate speech to text, We switch between both of them depending on the latency and language constrains. Sarvam AI is the best tool for Indian native languages. Our large language model is hosted in AWS bedrock which has access to Claude 3 Sonnet, Meta Llama 3, enabling users to choose their desired models flexibly. There are four different TTS providers namely elevenlabs, AWS polly, Sarvam AI and smallest AI. This architecture allows the users to configure their desire through a no code platform.

Full pipeline lifecycle is coordinated by BotRunner, initiates the Pipecat pipeline, event call backs gets registered, and allows smooth transition upon call termination. Multiple processes are inserted into the pipeline which is used to identify silence in the audio wavelength, barge-in interruption and implement turn talking logic.

4.2. Live Kit and SIP Integration

LiveKit is a WebRTC-based media server which provides low-latency, full duplex audio stream transport. The livekit client wrapper exposes APIs for room creation, participant token issuance, and participant management. For outbound calls, the system programmatically creates a LiveKit room connects the bot participant before dialing the destination endpoint.

Telephony integration is achieved through SIP trunking, managed by the SIPService component. Inbound SIP trunks are configured with dispatch rules that route incoming calls to designated LiveKit rooms, where a waiting bot instance accepts the audio session. For outbound calls, the service provisions SIP trunk credentials and initiates calls via the LiveKit SIP API. This architecture abstracts the telephony layer, allowing the bot pipeline to operate identically for both WebRTC-native and PSTN-originated calls.

4.3. Recording and Transcript Management:

We have also added the ability to record calls using LiveKit express. This captured room puts together and sends out audio streams. The recording service controls a state machine with four states (Idle, Starting, Active, and Stopped) to make sure that each recording session is unique and that there are no duplicates. Completed recordings are then uploaded to Amazon S3.

Conversation transcripts are accumulated in real time by the bot pipline's transcript processor. On call termination the transcript manager serializes the turn by turn exchange including speaker role, utterance text and timestamp analytics into a structured JSON document.

4.4. API Design and Webhook System

The proposed system has RESTful HTTP API built with FastAPI, which is split into four router modules health, calls, webhooks, and sip. FastAPI's Depends feature for dependency injection ensures that service instances are created only once during the application's lifetime and are reused across requests. This reduces overhead. Pydantic v2 models validate all request and response payloads, meaning that types are automatically converted and errors are reported in a clear way.

A special webhook endpoint gets Live Kit events that happen, like when a room is made, when someone joins or leaves, and when the egress state changes. Before processing, an HMAC-SHA256 signature checks each incoming webhook payload to make sure it came from the LiveKit server.

The CallService connects webhook event types to state changes and initiates actions like starting a recording, flushing the transcript, and shutting down the bot. This event-driven design ensures that the system remains responsive to asynchronous call lifecycle events without needing to poll.

5. Experimental results and evaluation

Table 1 Latency metrics

Metric	Count	Average	Minimum	Maximum
TTS TTFB	11	0.368	0.000	3.185
LLM TTB	6	0.829	0.000	1.083
Processing	18	0.339	0.000	1.510

The latency values which are measured for the analysis shows that the system has the capacity to provide a responsive real time voice interaction. Low average delay is showed by TTS Time to First Byte proving that the speech responses will begin quickly once the text is generated thought the random high values shows variability under few conditions. A higher average is given by LLM Time To Begin which indicates the time needed for generating contextual responses, but it stays is a considerable range for the conversational uses. The latency across multiple observations shows a relatively low latency , indicating the pipeline is minimal. Overall the results suggest that the proposed system gives an efficient response time and is very appropriate for an interactive voice based healthcare assistance scenarios.

6. Future work

It is aimed to focus on improving patient accessibility through multilingual support, researching in the direction to enable agents to actively communicate in regional languages. This requires multi-voice speech synthesis and multi-language speech recognition. The agent will also incorporate knowledge-base integrated within the dialogue nodes, which will allow access of claim information from backend sources, enabling the agent to perform in both administrative and procedural way. For safety, an escalation system which will transfer the call to a human operator will be implemented in cases of repeated failures and low confidence. Additionally, the system will consist of a sentiment analysis mechanism and emotion detection feature to adapt responses and identify specific sensitive situations.

7. Conclusion

This AI-Agent based telecommunication system acts as a low latency system that helps users in various capabilities. This is very efficient in healthcare domains that provide customer support to maintain quality response mechanisms towards the users. This system bridges the major gaps between the communications and support systems between the users especially in healthcare. Since this system is prepared with multiple Knowledge bases like FAQ's, Docs and Data that makes the model proficient and can be fine-tuned according to the use cases to give more accurate results. The Healthcare systems globally struggle with administrative burden, since administrative communication consumes hospital staff, Insurance staff, patient time and etc. This system handles these through handling of repetitive communication, reduces call centre load, provides 24/7 assistance which automatically results in improvement of patient response rates. This system is majorly impactful in places like the users or patients prefers phone calls instead of apps.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Dorai et al., "A novel machine learning approach for software reliability growth modelling with pareto distribution function," *Soft Computing*, vol. 23.0, pp. 8379-8387, 2019.

[2] Dr Prabha D, "A contemporary survey on the aspects of software reliability growth models via size of the software," *International Journal of Pure and Applied Mathematics*, vol. 119.0, pp. 1253-1271, 2018.

[3] Dr Prabha, "A REVIEW ON THE PROMINENT FACTORS OF SOFTWARE TESTING FOR OPTIMIZED OUTPUT THROUGH DATA ANALYTICS," *International Journal of Pure and Applied Mathematics*, vol. 115.0, pp. 95*106, 2017.

[4] Kavinraj A.S, "Facial image noise removal via a trained dictionary," *International Journal of Computer Trends and Technology*, vol. 9.0, pp. 14, 2014.

[5] "A LITERATURE SURVEY OF FUZZY BASED FACIAL EXPRESSION CLASSIFICATION," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 5.0, pp. 5519-5523, 2017.

[6] D et al., "Improved EM algorithm in software reliability growth models," *International Journal of Powertrains*, vol. 9.0, pp. 186-199, 2020.

[7] Priya et al., "Reversible information hiding in videos," *International Journal Of Innovative Research in Computer and Communication Engineering*, vol. 2.0, pp. nan, 2014.

[8] Dr. et al., "A NOVEL AI AND RF TUTORED STUDENT LOCATING SYSTEM VIA UNSUPERVISED DATASET," *Turkish Journal of Physiotherapy and Rehabilitation*, vol. 32.0, pp. 882-887, 2021.

[9] Dr.B.Arunkumar, "A Novel Approach for Boundary Line Detection using IOT During Tennis Matches," *Advancement of Electrical, Information and Communication Technologies for Life Application*, vol. 13.0, pp. 243-246, 2020.

[10] Dr. et al., "Performance Analysis of Enhanced Adaboost Framework in Multifacet medical dataset," *Natural Volatiles & Essential Oils*, vol. 8.0, pp. 1752 - 1756, 2021.

[11] Arun Kumar B et al., "AN OVERVIEW OF CLOUD SCHEDULING ALGORITHMS," *Vidyabharati International Interdisciplinary Research Journal*, vol. nan, pp. 2778 - 2782, 2021.

[12] S. Ashfia Fathima et al., "Performance Evaluation of Improved Adaboost Framework in Randomized Phases Through Stumps," *IEEE Xplore 2021*, vol. nan, pp. 1-6, 2021.

[13] D et al., "Self-Reliant Dimensionality Reduction That Uses Improved Pareto Distribution PCA Framework," *IEEE Xplore 2021*, vol. nan, pp. 1-5, 2021.

[14] Mr. Naren S R et al., "An Optimized Machine Learning Framework For Extracting Suicide Factors Using K-Means++ Clustering," *https://www.interscience.in/cgi/viewcontent.cgi?article=1197&context=ijcsi*, vol. 4.0, pp. 2231 –5292, 2022.

[15] Vignesh K et al., "Classification of Diabetics and Cardiovascular Diseases using Machine Learning Frameworks," *International Journal of Research Publication and Reviews (IJRPR)*, vol. 3.0, pp. 1848-1853, 2022.

[16] D et al., "Enhancing the efficiency of lung disease prediction using catboost and expectation maximization algorithms," *2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA)*, vol. nan, pp. 57-61, 2022.

[17] K. Suganya et al., "Data Communication Using Cryptography Encryption," *Asian Journal of Computer Science Engineering*, vol. 7.0, pp. 1-5, 2022.

[18] Arun Kumar B et al., "A Novel Ai Framework for Personalisation and Customization of Product Prices through Bigdata Analytics," *2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, vol. 1.0, pp. 1-6, 2023.

[19] Bhuvaneshwaran A et al., "A Multimodal AI Framework for Hyper Automation in Industry 5.0," *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, vol. 1.0, pp. 282-286, 2023.

[20] Vignesh et al., "Impact of Classification Algorithms on Cardiotocography Dataset for Fetal State Prediction," *Asian Journal of Computer Science Engineering*, vol. 7.0, pp. 71-76, 2023.

[21] Sri Thrishna J et al., "A Novel AI Framework for Anomaly Detection and Predictive Maintenance in Heterogenous Networks," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 11.0, pp. 9083-9086, 2023.

[22] Saranya et al., "Twitter Based Complaint Management System for Rail Transit," *2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, vol. nan, pp. 1-6, 2023.

[23] Saranya et al., "Emotion recognition using EEG signal classification of SEED dataset," *2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, vol. nan, pp. 1-6, 2023.

[24] D et al., "A PD ANN Machine Learning Framework for Reliability Optimization in Application Software," *2022 Smart Technologies, Communication and Robotics (STCR)*, vol. nan, pp. 1-4, 2022.

[25] D et al., "Performance Evaluation of Improved Adaboost Framework in Randomized Phases Through Stumps," *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, vol. nan, pp. 1-6, 2021.

[26] D, "A novel Sentimental Analysis framework using Gated Recurrent Units for Text Transliteration," *2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, vol. nan, pp. 1-6, 2023.

[27] P. S. Kailas et al., "Software Quality Prediction by CatBoostFeed-Forward Neural Network in Software Engineering," *System Reliability and Security*, vol. nan, pp. 272-284, 2023.

[28] D, "A novel Pareto Distribution based CatBoost framework for Software Reliability Estimation," *nan*, vol. nan, pp. nan, 2022.

[29] D et al., "Software Quality Prediction by CatBoostFeed-Forward Neural Network in Software Engineering," *System Reliability and Security*, vol. nan, pp. 207-218, 2023.

[30] D. et al., "A comparative study in predictive analytic frameworks in big data," *INTERNATIONAL CONFERENCE ON INNOVATIONS IN ROBOTICS, INTELLIGENT AUTOMATION AND CONTROL*, vol. 2914.0, pp. 10-18, 2023.

[31] D et al., "A novel adaptive framework for immersive learning using VR in education," *Transforming education with virtual reality*, vol. nan, pp. 1-26, 2024.

[32] D et al., "Implementing Catboost Algorithm for Allergen Cross Contamination Detection in Food Industry," *2023 International Conference on Emerging Research in Computational Science (ICERCS)*, vol. nan, pp. 1-4, 2024.

[33] D et al., "Proactive headcount and suspicious activity detection using YOLOv8," *Procedia Computer Science*, vol. 230.0, pp. 61-69, 2023.

[34] D et al., "An Abstractive Summarization and Conversation Bot using T5 and its Variants," *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, vol. nan, pp. 432-437, 2024.

[35] D et al., "A novel AI framework for assuring data sustainability in health care dataset," *2023 Third International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, vol. nan, pp. 161-166, 2023.

[36] D et al., "SOFTWARE RELIABILITY ANALYSIS BY USING THE BIDIRECTIONAL ATTENTION BASED ZEILER-FERGUS CONVOLUTIONAL NEURAL NETWORK.," *Neural Network World*, vol. nan, pp. nan, 2024.

[37] D et al., "Cloud-Enabled Predictive Modeling of Cancer Progression in Digital Twins: A LightGBM Classification Approach," *Smart Data Intelligence*, vol. nan, pp. 519, 2024.

[38] Kumar et al., "A Paretto Distribution based Gradient Boosting for Sustainable Agriculture," *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, vol. nan, pp. 1307-1311, 2024.

[39] D et al., "Data Synergizing by Behavioral Cloning," *Innovations in Cybersecurity and Data Science: Proceedings of ICICDS 2024*, vol. nan, pp. 171, 2024.

[40] D et al., "Data Synergizing by Behavioral Cloning and RNN for Autonomous Vehicles," *International Conference on Innovations in Cybersecurity and Data Science Proceedings of ICICDS*, vol. nan, pp. 171-184, 2024.

[41] D et al., "Parallelism in Cloud Architecture using Implicit Partition Clustering Framework," *2024 4th International Conference on Soft Computing for Security Applications (ICSCSA)*, vol. nan, pp. 278-282, 2024.

[42] D et al., "Sustainable Urban Street Lighting through Predictive Maintenance using IoT and AI," *2024 International Conference on Emerging Research in Computational Science (ICERCS)*, vol. nan, pp. 1-5, 2025.

[43] D et al., “AI Powered Monitoring and Risk Prediction for Maternal Health to Ensure Fetal Well-Being,” *2025 3rd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, vol. nan, pp. 1-5, 2025.

[44] Kanagaraj et al., “Deep Learning Techniques to Analyze Chest X-Ray and Providing Accurate Predictions Regarding the Presence of Lung Disease,” *2025 3rd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, vol. nan, pp. 1-6, 2025.

[45] D et al., “Implementing Vision Transformers for Precision Disease Detection in Spinach Cultivation,” *International Conference on Smart Data Intelligence*, vol. nan, pp. 339-349, 2025.

[46] D et al., “Designing Intelligent and Integrated Analytics Frameworks for Personalized Healthcare,” *2025 4th International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, vol. nan, pp. 1307-1313, 2025.