informatics

MDPI

Article

An AIoT-Based Framework for Automated English-Speaking Assessment: Architecture, Benchmarking, and Reliability Analysis of Open-Source ASR

Paniti Netinant [1], Rerkchai Fooprateepsiri [2], Ajjima Rukhiran [3] and Meennapa Rukhiran [4,*]

[1] Faculty of Business Administration and Information Technology, Rajamangala University of Technology Tawan-ok, Bangkok 10400, Thailand; paniti_ne@rmutto.ac.th
[2] Institute for Innovative Education and Lifelong Learning, Rajamangala University of Technology Tawan-ok, Chonburi 20110, Thailand
[3] Department of Provincial Administration, Ministry of Interior, Chanthaburi 22000, Thailand
[4] Faculty of Social Technology, Rajamangala University of Technology Tawan-ok, Chanthaburi 22210, Thailand
* Correspondence: meennapa_ru@rmutto.ac.th

Abstract

The emergence of low-cost edge devices has enabled the integration of automatic speech recognition (ASR) into IoT environments, creating new opportunities for real-time language assessment. However, achieving reliable performance on resource-constrained hardware remains a significant challenge, especially on the Artificial Internet of Things (AIoT). This study presents an AIoT-based framework for automated English-speaking assessment that integrates architecture and system design, ASR benchmarking, and reliability analysis on edge devices. The proposed AIoT-oriented architecture incorporates a lightweight scoring framework capable of analyzing pronunciation, fluency, prosody, and CEFR-aligned speaking proficiency within an automated assessment system. Seven open-source ASR models—four Whisper variants (tiny, base, small, and medium) and three Vosk models—were systematically benchmarked in terms of recognition accuracy, inference latency, and computational efficiency. Experimental results indicate that Whisper-medium deployed on the Raspberry Pi 5 achieved the strongest overall performance, reducing inference latency by 42–48% compared with the Raspberry Pi 4 and attaining the lowest Word Error Rate (WER) of 6.8%. In contrast, smaller models such as Whisper-tiny, with a WER of 26.7%, exhibited two- to threefold higher scoring variability, demonstrating how recognition errors propagate into automated assessment reliability. System-level testing revealed that the Raspberry Pi 5 can sustain near real-time processing with approximately 58% CPU utilization and around 1.2 GB of memory, whereas the Raspberry Pi 4 frequently approaches practical operational limits under comparable workloads. Validation using real learner speech data (approximately 100 sessions) confirmed that the proposed system delivers accurate, portable, and privacy-preserving speaking assessment using low-power edge hardware. Overall, this work introduces a practical AIoT-based assessment framework, provides a comprehensive benchmark of open-source ASR models on edge platforms, and offers empirical insights into the trade-offs among recognition accuracy, inference latency, and scoring stability in edge-based ASR deployments.

Keywords: automatic speech recognition; benchmark; hybrid learning; Internet of Things; Artificial Internet of Things; open source; system development; system testing

Academic Editor: Guangjie Han

Received: 18 December 2025
Revised: 17 January 2026
Accepted: 19 January 2026
Published: 26 January 2026

1. Introduction

Pressure is placed on educators, students, and their families to identify effective preparation methods for examinations, given their significant influence on admission decisions for undergraduate programs and institutions [1]. Especially in English-language instructions, exams usually test the four main skills of reading, writing, listening, and speaking to assess individual learning capabilities.

English learning in a traditional classroom is generally effective at helping students acquire new skills, but it often struggles to provide personalized feedback, targeted support, and long-term motivation for students of different skill levels. One-on-one tutoring is often an effective way to deliver and improve individualized learning. By talking to them directly, tutors can determine what students need to work on, build on what they already know, and adapt lessons to fit each student's needs [2,3].

In the last few years, hybrid teaching models have become more popular. They combine in-person interaction with online learning activities that utilize communication technology. This approach makes learning more accessible and flexible and makes it easier to manage resources in terms of time, place, and money while giving students more opportunities to practice [4,5]. Gudoniene et al.'s [6] study identified hybrid teaching systems that not only leverage the advantages of both physical and digital environments but also offer learners diverse modalities, convenient and consistent learning and practice, adaptable scheduling, access to a wide range of resources, and immediate feedback. Many studies on English language education have investigated blended methodologies for specific skills, including reading [7], writing [8], listening [9], and speaking [10]. Also, research on individualized one-on-one instruction emphasizes the importance of interpersonal communication in encouraging learner confidence and addressing specific learning weaknesses [11].

Blended learning and individualized tutoring have been extensively studied [12]; however, research has rarely examined hybrid one-on-one tutoring models that merge the individualized attention of tutoring with the flexibility and technological features of hybrid instructional delivery. Research gaps have been identified in understanding the impacts of combining models for long-term skill development and sustained student engagement in preparing for high-stakes examinations. Addressing this gap is critical to enhancing educational strategies and informing effective improvement practices.

In recent years, Artificial Intelligence (AI) has provided various tools that promise to address these challenges. Many AI-based systems have been studied to deliver language education support by utilizing Natural Language Processing (NLP), Automatic Speech Recognition (ASR), and machine learning algorithms to provide objective and real-time feedback across multiple skills [13]. In addition, AI-based scoring systems have been developed to evaluate essay writing for grammar [14], coherence [15], and lexical sophistication [16]; to assess speaking for fluency [17] and pronunciation [18]; and to analyze reading comprehension accuracy [19]. AI feedback capabilities enable continuous formative assessment, allowing both learners and instructors to identify progress trends, determine areas requiring skills-targeted attention, and optimize instructional time. They also facilitate personalized learning pathways to adjust content, difficulty levels, resource types, and feedback methods based on a learner's performance history, such as through adaptive e-learning systems [20]. Numerous prior studies on AI efficacy have identified a limitation: the lack of real-time feedback on contextually relevant data [21]. Numerous assessment systems utilizing AI exclusively depend on user inputs, such as recorded audio or typed text. Generally, user inputs inadequately address essential components of learning processes, including engagement levels, behavioral indicators, and contextual factors influencing learning outcomes.

Existing studies on ASR systems highlight the challenges and opportunities in low-resource, real-world, and embedded environments, particularly for educational and assistive applications. Yadava et al. [22] focused on ASR development for a low-resource language in noisy, uncontrolled environments. By integrating noise reduction techniques and time-delay neural network (TDNN) models, the system achieved measurable improvements in recognition accuracy. The importance of robust preprocessing, noise handling, and lightweight modeling strategies for real-time ASR deployment in practical applications was also highlighted. Mulfari et al. [23] demonstrated that convolutional neural network (CNN)-based ASR models using MFCC features can be effectively deployed on resource-constrained devices while maintaining acceptable word error rates, even for atypical speech such as dysarthria. The study emphasized the importance of low computational overhead, reduced latency, and real-time feasibility, positioning edge-based ASR as an alternative to cloud-dependent systems. Fatehi et al. [24] conducted a survey and experimental evaluation of ASR models originally trained in high-resource environments. The results showed that models optimized for large-scale datasets often perform poorly when directly applied to low-resource settings. They argued that pre-training on high-resource data followed by domain-specific fine-tuning is more effective than increasing model complexity, underlining the need to balance recognition accuracy with data availability and computational efficiency. Ahlawat et al. [25] presented an overview of ASR evolution, covering hybrid models, end-to-end deep learning approaches, Transformers, and multilingual systems. While the survey highlighted significant advances in recognition accuracy, it also identified limitations related to data dependency, generalization across domains, and real-time deployment constraints. The study concluded that accuracy alone is insufficient for practical ASR deployment and that latency, scalability, and performance efficiency must be considered, especially for embedded and IoT-based systems.

The Internet of Things (IoT), when used as edge computing, enables real-time data collection through interconnected sensors and actuators, fostering confidence in its ability to deliver continuous, immediate insights in educational and monitoring environments [26]. Examples of IoT-based learning include facial expression analysis [27] and the monitoring of interaction frequency and learner activity engagement [28], which not only provide real-time [29] datasets and promptly inform instructional decision-making but also enable the individualization of language-learning conditions and strategies to meet the specific improvement needs of each student.

Artificial Internet of Things (AIoT), which combines AI and IoT, supports personalized, learner-focused experiences by enabling adaptive mobility, real-time interactions, and efficient device connectivity. Previous studies have demonstrated that AIoT is effective in meeting students' needs, which can increase student engagement and academic performance [13,30,31], as well as in developing an awareness of the learning environment in both primary schools and post-secondary education [32], introducing students to digital skills, supporting learning equity, and preparing students for the many challenges of a rapidly changing digital world. Another study included issues concerning data privacy, infrastructure limitations, resource scarcity, and scalable solutions [33].

Our study proposes an AIoT hybrid tutoring framework built on ASR AI running on IoT platforms to provide an affordable, accurate, flexible, energy-efficient, and mobile solution for English language learning and exam preparation. The AI speech testing exam is implemented on Raspberry Pi 4 and Pi 5 architectures. The system provides a centralized information dashboard and compiles individual and group historical data on learners' performance in AI-based English speech assessments. Although the initial development stage emphasizes the English speech-test module, the system architecture promises to extend the design to support all four language skills: listening, speaking, reading, and

writing. The architectural framework offers several benefits by creating new learning opportunities for students and innovative teaching tools for educators.

The research questions of this study seek to bridge the gap between individualized one-on-one learning improvement and automated real-time AIoT English skill assessments.

(1) How can a hybrid one-on-one tutoring framework based on AIoT be designed to support real-time evaluation of English speech performance using cost-effective embedded hardware?
(2) How do different ASR engines, specifically the open-source Whisper and Vosk, perform in terms of accuracy, latency, and efficiency within the AIoT architectural platforms for the development of English speech assessments?
(3) To what extent can the AIoT hybrid tutoring framework enhance learner engagement within hybrid one-on-one tutoring, motivation, and measurable progress in English speech proficiency?

The contributions of this study are threefold.

(1) The creation of a hybrid tutoring system that is AIoT-based and combines real-time English speech assessments with language performance analytics to facilitate individualized hybrid one-on-one English learning.
(2) The speaking test module is evaluated in the empirical study of AIoT architectural platforms, which includes the comparative benchmarking of ASR engines, the analysis of performance metrics on AIoT devices, and the evaluation of English speech assessment accuracy.
(3) The system establishes technological and pedagogical standards for the implementation of AIoT frameworks in hybrid one-on-one tutoring to enhance learner engagement, motivation, and achievement.

There are five parts to this article. The Related Work Section examines earlier studies on IoT in education, AI-driven language testing, hybrid learning environments, and personalized tutoring systems. The Materials and Methods Section described the suggested AIoT tutoring framework in detail, including hardware and software setups, ASR models, the AI processing pipeline, the system architecture, and evaluation methods. The Results and Discussion Section presents the most important results and how they should be interpreted. The Conclusions Section summarizes the study's main contributions, discusses the real-life implications of the proposed system, and suggests directions for further research.

2. Related Work

The development of an AIoT-based hybrid tutoring framework draws upon five major research streams: IoT-enabled learning environments, AI-driven language proficiency assessment, hybrid and blended language instruction, personalized one-on-one tutoring and scaffolding, and the gaps that remain in integrating these domains. This section synthesizes these areas to establish the foundation and justification for the proposed system.

2.1. IoT-Enabled Learning and Context-Aware Instruction

IoT has become increasingly significant in educational contexts due to its ability to collect multimodal, real-time data that can enhance learner engagement, environmental monitoring, and instructional responsiveness [34]. The early development of smart classroom technology focused on using sensors to track temperature, humidity, lighting conditions, and motion to enhance the educational environment, improve overall energy use, and encourage students' active participation in the learning process [35]. The IoT-enabled system demonstrates how IoT devices may provide an adaptable, environmentally aware learning environment.

In addition, the study of students' IoT-based interactions enables extended support by providing environmental awareness through remote video streaming. IoT devices have been used to assess student involvement/attendance during remote learning sessions by analyzing sensor-generated activity patterns [36]. Voice-activated assistants can also be integrated with IoT to assist students in completing tasks, retrieving course materials, and practicing languages through multimodal input methods that combine speech recognition and sensor data [37]. Thus, the potential of IoT extends beyond providing additional support for the educational process to laying the foundation for ongoing assessment and the development of personalized learning areas.

Overall, IoT-enabled learning environments add value to traditional assessments by incorporating multimodal signals, connecting physical and virtual learning environments, and enabling educators to gain actionable insights into learner participation, performance trends, and situational constraints, all of which are relevant to hybrid and individualized tutoring systems.

2.2. AI for Self-Learning in English Proficiency

ASR with automated scoring has significantly increased the reliability and scalability of language assessment in reading, writing, listening, and speaking. AI-based systems can assess many linguistic features, such as lexical diversity, syntactic complexity, grammatical accuracy, fluency, and pronunciation quality, which often align well with human scoring standards [38–43].

Automated Essay Scoring (AES) systems, such as the International Energy Agency (IEA) and other machine-learning-based approaches, are highly valid in distinguishing writing performance by assessing both micro- and macro-level features [44]. Regarding speaking assessments, ASR systems have shown the ability to evaluate intonation, rhythm, articulation, and prosody. Hidden Markov model (HMM) has achieved recognition rates exceeding 90% in controlled conditions [45], supporting their use in learning-oriented feedback environments [46].

Recent studies further demonstrate strong ASR, human scoring correlations (e.g., ICC $\approx$ 0.97), improved accuracy via fine-tuning, and effectiveness across diverse age groups and speech corpora [47–50]. Despite this, a vast majority of applications rely on cloud servers, with very high computational resource requirements. They also typically focus on one single skill at a time, thus limiting accessibility in low-resource environments, as well as environments where students may be using a smartphone or need an application they can use for individualized tutoring.

The potential to embed ASR engines (Whisper and Vosk) on low-power (IoT) devices offers new opportunities to create cost-effective Common European Framework of Reference for Languages (CEFR) assessments that do not require access to external computing resources, an area that has been relatively unexplored to date and represents a major developing trend in learning technology.

2.3. Hybrid and Blended Learning Models in Language Education

Face-to-face and digital learning environments, known as blended learning, combine face-to-face instruction with digital tools to enable students to learn in flexible ways while still maintaining some of the social benefits that support second-language acquisition [51,52]. Hybrid learning environments have been found to increase learners' autonomy, create more opportunities for learners to receive input from a variety of sources, and allow learners to practice their language skills in different modes [53–55]. Additionally, sociocultural perspectives on learning view hybrid learning environments as spaces for

mediated learning, including how tools and technology can influence learners' experiences and knowledge construction [51].

Research has also indicated that hybrid learning models can include both synchronous (in real time) and asynchronous (at your own pace) types of learning and provide learners with opportunities to develop long-form discourse, review material at leisure, and apply what they have learned to real-world tasks outside of the classroom environment [56,57]. The hybrid model is particularly effective for language learning, as it allows learners to have repeated exposure to a topic and/or skill and to practice language skills independently.

One challenge of hybrid learning is encouraging learners to remain engaged over an extended period; this can be difficult when they lack access to consistent, guided instruction and/or timely feedback [57]. However, one way to address these issues is through AI- and IoT-based instructional systems that monitor students' progress, adaptively scaffold learning, and promote learner motivation.

2.4. Personalized One-on-One Tutoring, SLA Principles, and Scaffolding

For many years, researchers have identified one-on-one, personalized tutoring as an especially effective method to improve student outcomes. Both types of tutoring provide instruction based on a student's preferences, goals, or interests, and adapt instruction to each learner's abilities and developmental needs. One-on-one tutoring is tailored to students' ability levels, academic performance, and developmental needs, while personalized tutoring is tailored to students' interests and preferences [58–60].

Both forms of tutoring are supported theoretically by second-language acquisition (SLA) theories, specifically the interaction hypothesis, which focuses on negotiation of meaning, corrective feedback, and opportunities for student output. SLA theorists also use Vygotsky's Zone of Proximal Development (ZPD) to emphasize the importance of scaffolding to help students develop skills beyond those they can accomplish independently [59].

Newer digital tools now make it possible to deliver tutoring in a variety of formats online. Digital tutoring tools enable instructors to provide synchronous and asynchronous feedback, allow students to view and listen to recorded sessions, allow for real-time interaction between instructors and students, and include a wide range of embedded learning supports [61–64]. In recent studies, large language models (LLMs) have shown the potential to scaffold complex thought processes and produce structured feedback in the form of dialogic interaction [65,66]. These newer tools and platforms enable the creation of even more adaptive, rich scaffolding within online tutoring environments.

However, research indicates that students who receive only asynchronous feedback may lack a deeper understanding of concepts unless they have the opportunity to ask questions and clarify information in real time [67]; additionally, research shows that students learn best when they can interact in real time with their instructor [68]; therefore, hybrid tutoring models that incorporate both modes may represent the greatest potential.

IoT-enabled platforms build upon prior scaffolding models by incorporating contextual data beyond environmental noise and student engagement indicators to adaptively guide students. This type of data is critical for high-stakes speaking performance, where a variety of contextual factors can significantly affect a student's speech quality.

Using both synchronous and asynchronous forms of tutoring allows students to receive timely, relevant assistance [69]. Additionally, using AI and IoT-based features will enhance scaffolding by allowing instructors to provide students with individually tailored linguistic feedback and monitor contextual issues that may negatively impact student performance. This combination of AI and IoT features enables the proposed hybrid tutoring model to be designed for personalized, adaptive, and context-aware learning.

2.5. Research Gap

While much research has shown that hybrid learning and one-to-one tutoring are effective methods for increasing English proficiency, they have generally been explored separately, with little attention to integrating AI and IoT technologies. Hybrid learning models have been shown to increase student participation, allow students to use various learning tools, and offer flexibility to meet the changing needs of the learner population [57,70]. Similarly, one-to-one tutoring has consistently improved learner outcomes by using varied instructional paces, providing individualized feedback, and offering targeted instructional support to meet learner needs based on SLA [4,59]. However, while many hybrid learning studies have focused on improving group instruction in college or professional development settings, few have examined individualized hybrid learning models.

Similarly, while blended learning environments [56] and stand-alone digital tutoring systems [69] have been extensively researched, relatively few studies have developed models that combine AI-based assessment with IoT-enabled contextual sensing to build real-time, adaptive, and personalized tutoring systems [71,72]. Emerging research on AI-driven feedback and ITS has shown the value of data-driven adaptive instruction [73–75]; however, these systems generally do not include environmental and contextual elements that impact speaking performance. The use of IoT monitoring can help capture these missing contextual elements by monitoring ambient noise, connectivity stability, and engagement patterns, thereby enabling more accurate, contextually aware scaffolding. The use of IoT monitoring is particularly important in high-stakes speaking assessments, as environmental conditions directly affect learner output. While there is great potential to develop a system that integrates AI and IoT, currently, there is a lack of individualized hybrid tutoring frameworks that leverage both to support one-to-one language assessment and instruction in real time.

To address this gap, the current study will develop an AIoT-enhanced hybrid tutoring framework that uses ASR-based speaking assessment via low-cost IoT devices and connects real-time performance metrics to contextual data. The system will provide immediate, CEFR-aligned feedback to learners; support ongoing skill tracking; and deliver individualized tutoring that meets each learner's speaking proficiency and exam-preparation needs. Speaking was chosen as the primary skill area because it plays a major role in high-stakes language testing and requires immediate, accurate, and environment-sensitive feedback.

3. Materials and Methods

In this section, the development and validation of an AIoT hybrid tutoring system for real-time English speech proficiency assessment, along with the methodology, are described. For this study, only speaking ability was evaluated due to its potential for automated processing via ASR, its direct relationship to communicative competence, and its high complexity. The hybrid framework uses both ASR models and an IoT-enabled web platform to allow students to complete a speaking task and receive immediate, data-driven feedback. To compare two model families of ASR capabilities, two ASR models were used: Whisper (version v20250625—four variants in multiple quantization formats—tiny, base, small, and medium English models) [76,77] and Vosk (version commit a428d65—small and large English models) [78,79]. To measure recognition accuracy, fluency, and computational efficiency for each model, 100 repeated trials were run under identical test conditions. In addition to measuring the performance metrics of the models using word error rate (WER), accuracy percentage, recall, words per minute (WPM), pause ratio, fillers per minute, band fluency, band pronunciation, band prosody, overall speaking band, processing time, CPU utilization, memory consumption, etc., results were also visualized as a dashboard to enable learners and instructors to view results immediately following the completion of the test.

Students chose from different levels of test difficulties and time allocations, after which the system generated a random test passage. Students then completed a test within 5–7 min while speech was recorded at 16 kHz, mono, in WebM format. Following completion of the test, the resulting audio was converted into MP4 format and processed by the selected ASR model. Immediately after the test, results were displayed on the dashboard for the students and instructors to view.

Since this study did not involve any medical or psychological interventions, it did not require formal ethical approval. However, in compliance with data privacy regulations, the study implemented all necessary anonymization procedures. The study removed identifiable information from all student responses, and all audio recordings were stored on encrypted servers. All source code, configuration files, and datasets will be made available for replication and further research, except for third-party-licensed models.

3.1. AIoT Hybrid Tutoring Framework

The proposed hybrid tutoring model based on AI and IoT consists of a multilayer structure [80]. It has been developed to provide an assessment and feedback process concerning real-time, English-speaking assessment in one-to-one tutoring settings. As depicted in Figure 1, the system architecture was designed to achieve modularity, with each layer performing a distinct role and working together to ensure a seamless flow of data from the hardware (data capture) to the application (personalized feedback).

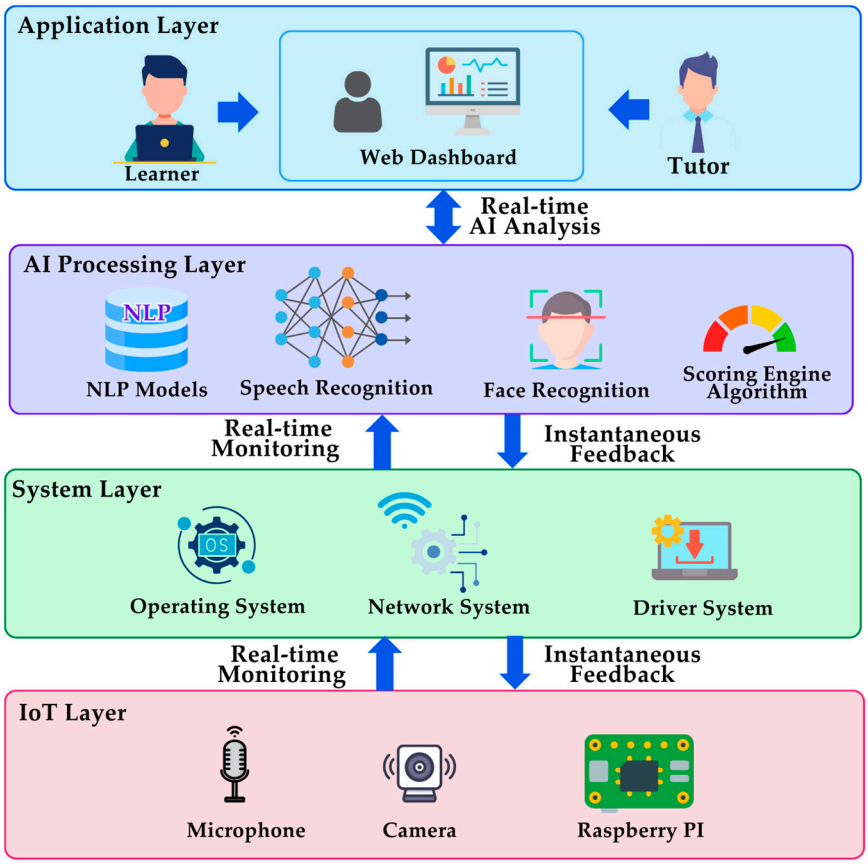


Figure 1. Layered AIoT architecture for automated real-time ASR-based speaking assessment.

The base layer of the structure is the IoT layer, which captures the student's input via a physical interface. Each layer of the structure functions separately yet together to ensure a seamless flow of data captured from the hardware to the application, providing students with personalized feedback. At the bottom of the structure, the IoT layer captures the learner's physical input. The central hub is a Raspberry Pi equipped with a camera and

a microphone to collect the learner's spoken responses and interactions. The benefits of the layered structure are low data collection latency and minimal energy use. Since the IoT layer gathers data locally, the learner does not need to rely on an external network; thus, the learner's experience is unaffected regardless of whether sufficient bandwidth is available.

In the provided example structure, the IoT layer is the system layer. The system layer is between the low-level hardware and the high-level processing layers. The system layer serves as a coordination layer that enables communication among processing units, microphones, and cameras through the operating system, network stack, and device drivers. A major benefit of the system layer is its coordination of resource distribution. Therefore, the system layer will provide seamless synchronization of audio and video during live assessments.

Above the system layer is the AI processing layer, which takes the raw data and converts it into computational data, information helpful for something. The NLP algorithms examine the syntax, fluency, and pronunciation of what the learner has said. Also, voice recognition engines like Vosk and Whisper transcribe the learner's spoken response. This layer is also modular; therefore, smaller models (such as Whisper-tiny) can be used for speed when hardware allows, and larger models (such as Whisper-medium) can be used if accuracy is the most critical factor. However, the output is not just a number; it is compared to global standards because the grading system aligns with CEFR descriptions. In addition, learners receive continuous feedback through the system, allowing them to immediately see their exact mistakes and adjust their responses in real time.

Finally, the application layer is the top layer of the structure and displays the results using a web-based dashboard. The key feature of the application layer interface is the dashboard's functional usability. The dashboards display the learners' overall performance, and teachers can see both current and historical results. The visual dashboard shows learners' progress over time and helps teachers refine their teaching methods. The platform offers many display options across various devices (desktop computers, tablets, and mobile phones) and can be used in a wide range of educational settings.

The framework's three main strengths stem from its four-layer design. This multilayer approach is adaptable, flexible, and scalable, with components and interfaces that can be upgraded without reinventing the whole architecture or its functional systems; it is also robust and tolerant enough to allow each layer to be optimized independently. By consistently capturing, analyzing, and providing actionable feedback based on learner input, these qualities make the tutoring process both reliable and personalized.

3.2. AIoT Hardware and Software Environment

Due to Python's ability to run on different hardware and software libraries, it was chosen to provide accurate information, quick processing, and sufficient user feedback. Two single-board computer systems, the Raspberry Pi 4 Model B and Raspberry Pi 5 (both equipped with 8 GB RAM), were selected as edge-AI computing platforms for their mobility, lower power consumption than larger computers, and demonstrated ability to handle AI workloads. The Raspberry Pi 4 is the main device, while the Raspberry Pi 5 is used for more computationally intensive tasks, such as higher-quality ASR. Table 1 lists the hardware and software elements of the hybrid AI-IoT tutoring system. A Logitech C922 camera with a built-in stereo microphone captures 1080p video and 16 kHz audio and stores them locally on a 32 GB microSD card with SSD support for redundant storage. A stable 5V/5A USB-C power supply enables continuous use. Both the Raspberry Pi 4 and 5 run Raspbian Bookworm 64-bit (Linux Kernel 6.12) and thus support Python-based programming across all layers of the system.

Table 1. Summary of hardware and software components in the AIoT automated ASR system.

Category	Component	Specifications
Hardware	Raspberry Pi 4 Model B (8 GB RAM) and Raspberry Pi 5 (8 GB RAM)	Single-board processing units for AI execution, inference, and system integration, supporting local storage and multiple connectivity.
	Logitech C922 Pro Stream USB Camera (with built-in microphone)	Captures images in high-definition 1080p resolution and stereo audio at 16 kHz for speaking assessments.
	Local Storage 64 GB (microSD)	Primary storage for the operating system, AI models, recorded audio data, and assessment results; SSD used for caching larger files.
	Power Supply Unit	Provides stable 5V/5A regulated power for Raspberry Pi boards and connected peripherals.
Software	ASR Models (Whisper, Vosk)	Whisper (tiny, base, small, and medium variants) and Vosk (small and large English models)
	NLP Pipeline	Tokenization, grammar, and syntax analysis; vocabulary richness assessment; error detection for automated feedback.
	Pronunciation and Fluency Scoring Engine	Evaluates phoneme accuracy, prosody, pause ratio, fillers, and overall fluency against CEFR-aligned benchmarks.
	Scoring Engine	Implements automated scoring algorithms calibrated to human grading standards.
	Web Server Framework (FastAPI/Flask with Uvicorn)	Hosts APIs and manages asynchronous communication between the Raspberry Pi and the dashboard.
	Front-End Dashboard (HTML5, CSS3, JavaScript, Chart.js)	Interactive interface for students and tutors; visualizes performance metrics, CEFR progression, test history, and comparative analysis.
	Database (SQLite with SQLAlchemy 2.0)	Stores learner profiles, test logs, and session results for retrieval and analysis.

Whisper, a transformer-based ASR model developed by OpenAI, converts 16 kHz audio to an 80-channel log-Mel spectrogram. The encoder in Whisper uses multiple self-attention layers to contextualize temporal frames, while the decoder produces subword tokens via probabilistic beam search. In this research, Whisper was implemented using CTranslate2, reducing the model's memory requirements from gigabytes to tens of megabytes by applying INT8 or hybrid quantization, thus enabling real-time inference on ARMv8 processors of Raspberry Pi 4 and 5.

Vosk utilizes TDNN acoustic models and WFST decoding similar to the Kaldi speech recognition toolkit. Vosk captures 13-dimensional MFCC features, optionally using i-vectors for adaptation, and supports inference directly on embedded platforms without a GPU. Vosk achieved inference latency of under 120 s per test on both Raspberry Pi platforms, demonstrating consistent, predictable performance in an interactive classroom environment. Vosk enables rapid responses during live sessions, providing complementary high-fidelity transformer-based recognition capabilities for Whisper.

Eleven Whisper configurations are tested (tiny, base, small, and medium variants) using three quantization schemes (int8, hybrid int8_float32, and float32). A uniform set of parameters is used for the preprocessing stage of Whisper (i.e., 16 kHz sampling rate, 20 ms frame length, and a beam size of 5) to manage the preprocessing load and balance the trade-off between inference time and accuracy. Hybrid quantization is used to stabilize performance in a variety of real-world scenarios.

Three Vosk configurations (small-en-us-0.15, 0.22, and 0.22-lgraph) are tested in float32 precision using an FST decoding pipeline. A 16 kHz sampling rate and a 25 ms frame length are used to enable robust word boundary detection, while a beam size of 10 improves the recognition and recall of continuous speech segments.

The IoT layer captures audio at 16 kHz mono using either sounddevice or PyAudio. When needed, video streams can also be accessed through OpenCV. The ffmpeg-python library is used to convert WebM audio into MP4 format so that the downstream ASR models can utilize the data.

The system layer enables real-time communication between the client and server via Uvicorn, with Flask managing the API endpoints. Throughout live sessions, the system layer continuously monitors system health metrics (CPU, memory, and process load) using psutil.

The AI processing layer consists of Whisper (via PyTorch) and Vosk (via Python/NumPy/SciPy) for transcribing the input audio, followed by a custom scoring module for computing WER, WPM, pause ratio, filler rate, and CEFR-aligned fluency and pronunciation bands.

The dashboard and visualization layer utilize HTML5, CSS3, JavaScript, and Chart.js to render session logs and performance metrics via Jinja2 templates. Performance metrics and session logs are stored in SQLite using SQLAlchemy.

The combined hardware–software framework allows students to record speech directly through the dashboard. The audio is then processed in near-real-time by the dual-ASR architecture and scored, resulting in the instantaneous display of the results. This feature demonstrates that complex AI processing can be hosted on Raspberry Pi devices while enabling scalable user interaction and deployment for educational purposes.

3.3. AIoT Flow and Processing Pipeline

The AIoT hybrid tutoring system data flow and processing pipeline were established to seamlessly integrate the learner's spoken input through speech recognition, model analysis, and real-time feedback delivery, as illustrated in Figure 2. The figure shows how the architecture provides continuous assessment and improvement of the system by processing data from the beginning of the primary process, capturing a learner's voice, through to the end of the primary process. The learner's speaking proficiency is represented visually on the dashboard.

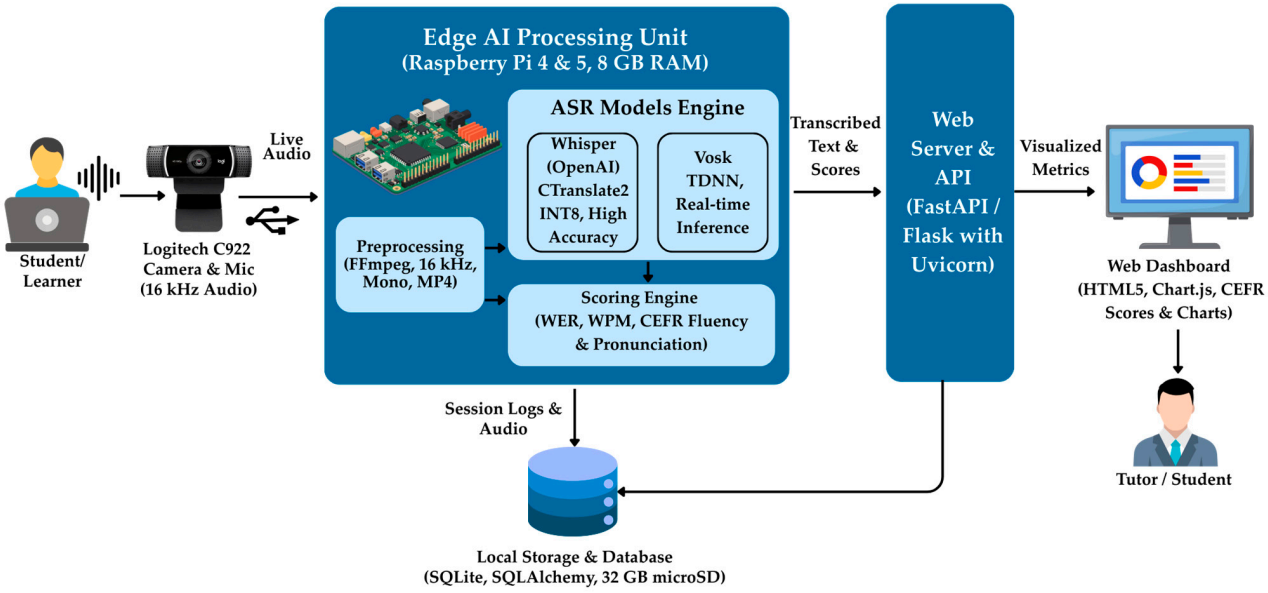


Figure 2. Edge-AI processing architecture for ASR inference, scoring, and web-based feedback delivery on Raspberry Pi platforms.

The processing pipeline for the framework consists of four main stages—data capture, AI processing, feedback generation, and dashboard visualization—which operate in a continuous cycle.

The primary process begins when a learner initiates a speaking test via the web interface. The speech is recorded by the embedded microphone in the Logitech C922 camera and saved at 16 kHz in mono to meet the requirements of the ASR models. Before processing, the raw audio is stored in WebM format and subsequently converted to MP4 format using the ffmpeg library. This conversion is completed to balance storage efficiency with compatibility with the processing pipeline. At this stage, clarity and precision of the recording are critical. Clarity and precision of the recording are important because the quality of the captured audio will determine the accuracy of transcription and scoring of the student's performance at later stages of the system.

After capture, the audio is passed to the AI processing layer running on Raspberry Pi 4 and Pi 5 devices. Two different kinds of automatic speech recognition models are employed to convert spoken English voice passages into text. Although the inference complexity is greater than in the previous example, Whisper (implemented in PyTorch) is used to support the development of the more complex inference process. At the same time, the Vosk version provides a simpler model that functions well in real time on low-end devices. Following completion of the transcription, a customized evaluation component assesses all of the performance metrics from the transcription output (accuracy percentage, WER, WPM, pause ratio, filler rate, CEFR-aligned fluency, CEFR-aligned pronunciation, CEFR-aligned prosody, and CEFR-aligned overall speaking band). The above metrics provide both a numerical representation of student performance and a qualitative description.

Students' scores are then converted into usable feedback. The automated scoring algorithm maps students' CEFR descriptors to provide a standardized method for comparing their proficiency levels at any point in time. At the same time, the system indicates to the instructor which areas could be improved (e.g., excessive pause ratio, insufficient clarity in speech articulation, poor fluency, etc.). Since this system was designed to be an instantaneous technique, students and learners would receive continuous feedback. This feedback would presumably help students and learners practice, correct, and develop their spoken language skills.

The results are then displayed on a web dashboard for students, learners, and tutors to view. The tutors or educators can observe the details of the analytics used to assess students' or learners' performance over a time interval, the CEFR levels of the students or learners, and the results of pre- and post-tests for each student or learner. Further, the dashboard enables tutors to record notes, intervention activities, and other details and observe their effects across multiple sustained tutoring sessions. By utilizing instantaneous feedback alongside students' or learners' historical data, the dashboard provides the principal location where data analytics and human involvement meet to support acquisition learning.

The pipeline is structured as a continuous loop. The dashboard output from the current iteration is connected to the next iteration's data collection process. As a result, closed-loop design allows the system to evolve as learners' competency increases, providing a customized, unique way for students to practice their English-speaking skills.

3.4. Dashboard Features

A second prominent feature of the AIoT hybrid tutoring framework's dashboard is its facilitation of teacher–student dialogue. Using this online interface, teachers and students can communicate about each other's development, share feedback on each other's responses to feedback, and support each other in decision-making about education. When creating the dashboard for this project, the authors focused on designing an accessible, transparent,

and interactive experience. This provides users with organized, easily understandable AI-generated data analysis and ensures that users find the data useful. This was accomplished through the inclusion of automated grading, visual analytics, and tutor feedback.

Figure 3 outlines the dashboard's general layout. The filters at the top enable users to organize data by student, date, subject, etc. Below the filter section, scorecards display the top-level results of the assessment (i.e., total score, CEFR level, and individual scores in reading, writing, speaking, and listening). The interactive graphics below the scorecards detail the student's overall performance before and after tutoring (time series plots, CEFR progression charts, and comparison bars). One final panel connects levels of background noise to the student's ability to speak, enabling users to contextualize variation in results. By combining top-level indicators and performance trends, the AIoT hybrid tutoring framework provides educators and learners with a multi-layered perspective on student development.

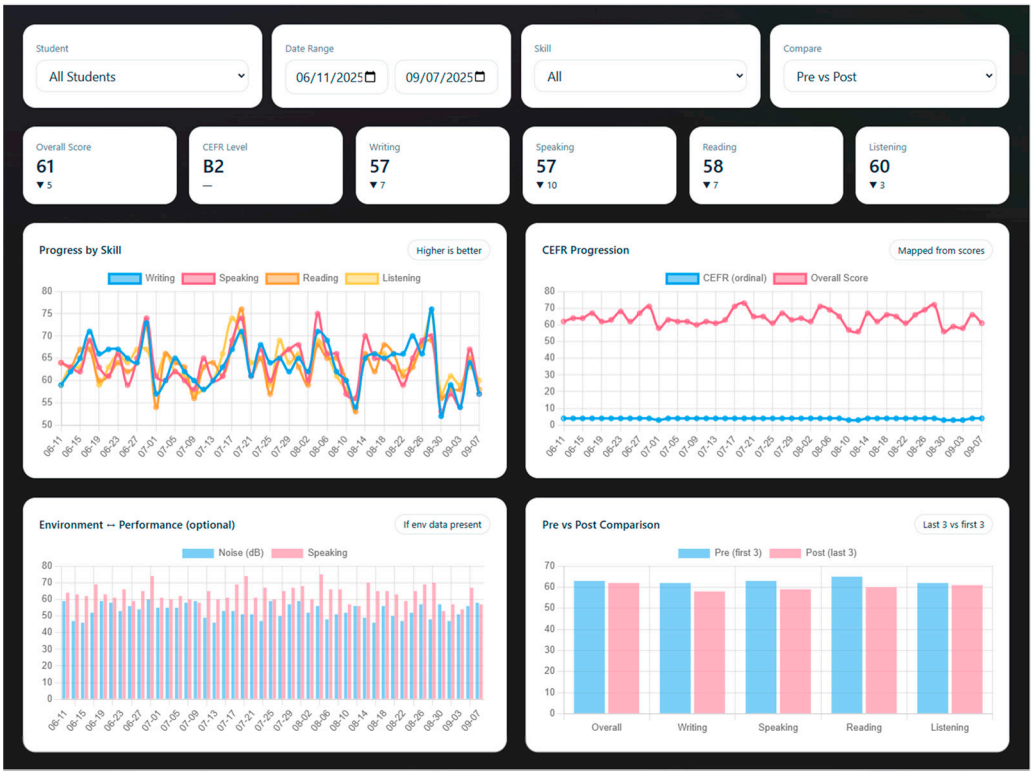
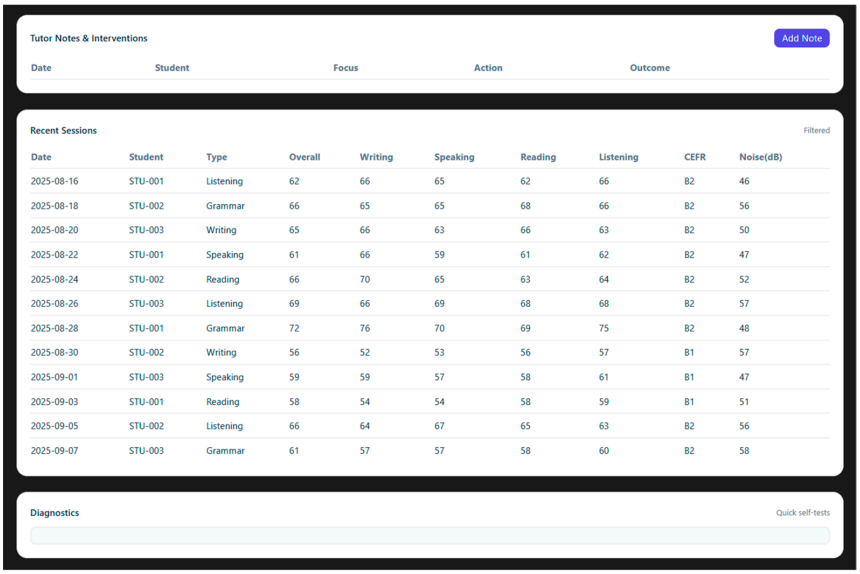


Figure 3. An overall dashboard of the AI-IoT hybrid tutoring framework.

The tutor notes and intervention panel are shown in Figure 4. Teachers can keep track of their observations, lesson plans, and subsequent results in this section of the dashboard. An up-to-date record of sessions is available, detailing the test type, total score, sub-skill outcomes, CEFR level, and recorded noise levels. By keeping track of previous sessions, both the instructor and the student can identify what went well and where they may need improvement. This enhances transparency. The system finds a balance between a dispassionate analysis of the metrics and the teacher's personal experience, including the use of automated metrics in the procedure for assessing progress alongside professional input, so that improvement is based first on numbers and then on professional experience.

A complete test interface for the real-time English-speaking test is illustrated in Figure 5. Learners begin by choosing their test type, passage length, preparatory time, and test time. Once the test has started, a passage of talking text is presented, which the learner reads out during the available test time. The speech input is recorded in mono at 16 kHz and stored in WebM format for subsequent conversion to ASR processing.



Tutor Notes & Interventions — Add Note

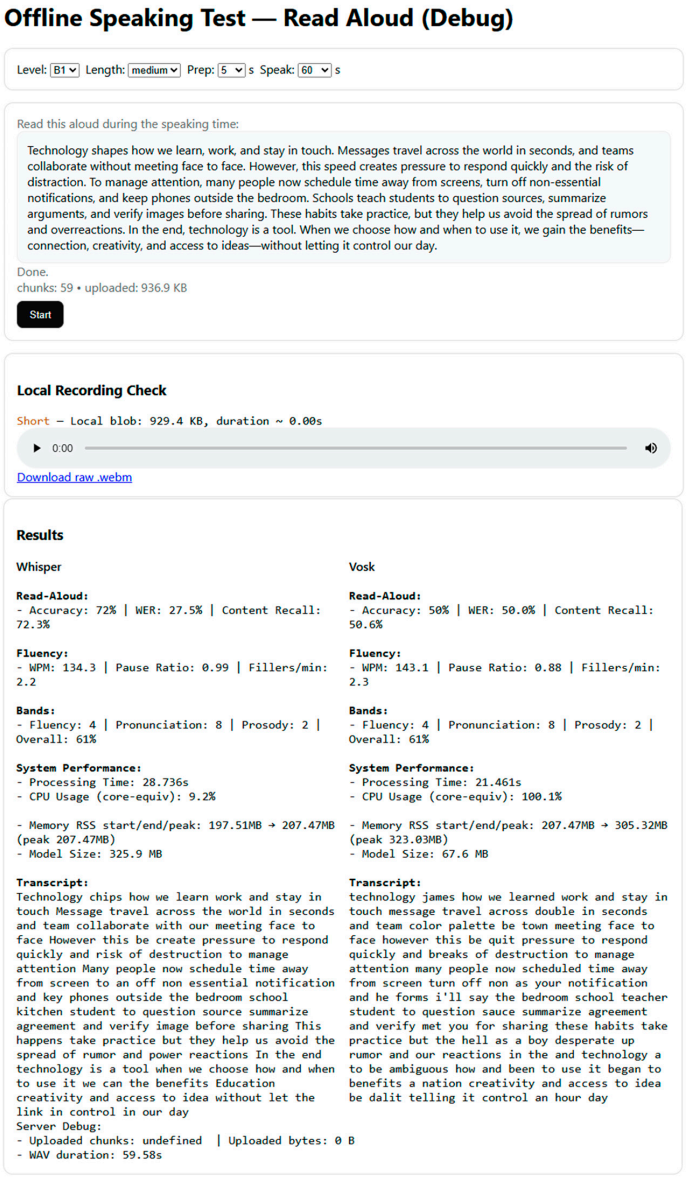
Date	Student	Focus	Action	Outcome

Recent Sessions — Filtered

Date	Student	Type	Overall	Writing	Speaking	Reading	Listening	CEFR	Noise(dB)
2025-08-16	STU-001	Listening	62	66	65	62	66	B2	46
2025-08-18	STU-002	Grammar	66	65	65	68	66	B2	56
2025-08-20	STU-003	Writing	65	66	63	66	63	B2	50
2025-08-22	STU-001	Speaking	61	66	59	61	62	B2	47
2025-08-24	STU-002	Reading	66	70	65	63	64	B2	52
2025-08-26	STU-003	Listening	69	66	69	68	68	B2	57
2025-08-28	STU-001	Grammar	72	76	70	69	75	B2	48
2025-08-30	STU-002	Writing	56	52	53	56	57	B1	57
2025-09-01	STU-003	Speaking	59	59	57	58	61	B1	47
2025-09-03	STU-001	Reading	58	54	54	58	59	B1	51
2025-09-05	STU-002	Listening	66	64	67	65	63	B2	56
2025-09-07	STU-003	Grammar	61	57	57	58	60	B2	58

Diagnostics — Quick self-tests

Figure 4. Tutor notes, interventions, and recent session records with scores, CEFR levels, and noise data.



Offline Speaking Test — Read Aloud (Debug)

Level: B1 Length: medium Prep: 5 s Speak: 60 s

Read this aloud during the speaking time:

Technology shapes how we learn, work, and stay in touch. Messages travel across the world in seconds, and teams collaborate without meeting face to face. However, this speed creates pressure to respond quickly and the risk of distraction. To manage attention, many people now schedule time away from screens, turn off non-essential notifications, and keep phones outside the bedroom. Schools teach students to question sources, summarize arguments, and verify images before sharing. These habits take practice, but they help us avoid the spread of rumors and overreactions. In the end, technology is a tool. When we choose how and when to use it, we gain the benefits—connection, creativity, and access to ideas—without letting it control our day.

Done.
chunks: 59 • uploaded: 936.9 KB

Start

Local Recording Check

Short — Local blob: 929.4 KB, duration ~ 0.00s

0:00

Download raw .webm

Results

Whisper

Read-Aloud:
- Accuracy: 72% | WER: 27.5% | Content Recall: 72.3%

Fluency:
- WPM: 134.3 | Pause Ratio: 0.99 | Fillers/min: 2.2

Bands:
- Fluency: 4 | Pronunciation: 8 | Prosody: 2 | Overall: 61%

System Performance:
- Processing Time: 28.736s
- CPU Usage (core-equiv): 9.2%

- Memory RSS start/end/peak: 197.51MB → 207.47MB (peak 207.47MB)
- Model Size: 325.9 MB

Transcript:
Technology chips how we learn work and stay in touch Message travel across the world in seconds and team collaborate with our meeting face to face However this be create pressure to respond quickly and risk of destruction to manage attention Many people now schedule time away from screen to an off non essential notification and key phones outside the bedroom school kitchen student to question source summarize agreement and verify image before sharing This happens take practice but they help us avoid the spread of rumor and power reactions In the end technology is a tool when we choose how and when to use it we can the benefits Education creativity and access to idea without let the link in control in our day
Server Debug:
- Uploaded chunks: undefined | Uploaded bytes: 0 B
- WAV duration: 59.58s

Vosk

Read-Aloud:
- Accuracy: 50% | WER: 50.0% | Content Recall: 50.6%

Fluency:
- WPM: 143.1 | Pause Ratio: 0.88 | Fillers/min: 2.3

Bands:
- Fluency: 4 | Pronunciation: 8 | Prosody: 2 | Overall: 61%

System Performance:
- Processing Time: 21.461s
- CPU Usage (core-equiv): 100.1%

- Memory RSS start/end/peak: 207.47MB → 305.32MB (peak 323.03MB)
- Model Size: 67.6 MB

Transcript:
technology james how we learned work and stay in touch message travel across double in seconds and team color palette be town meeting face to face however this be quit pressure to respond quickly and breaks of destruction to manage attention many people now scheduled time away from screen turn off non as your notification and he forms i'll say the bedroom school teacher student to question sauce summarize agreement and verify met you for sharing these habits take practice but the hell as a boy desperate up rumor and our reactions in the and technology a to be ambiguous how and been to use it began to benefits a nation creativity and access to idea be dalit telling it control an hour day

Figure 5. The real-time English-speaking test webpage.

A local recording check is performed to ensure that some input is detected at least before the speech analysis phase. This test interface is seamlessly linked to the dashboard environment, with test results streamed directly to the analysis pipeline and ultimately displayed in the visualization panels.

3.5. Speaking Test Design and Passage Generation System

The spoken component of the hybrid tutoring framework uses artificial intelligence to generate and administer a series of tests that assess a student's oral proficiency according to predetermined test protocols (e.g., time-limited and/or sequential). These tasks were designed to evaluate various components of spoken language performance and fluency (e.g., pronunciation, prosody, accuracy, and lexical complexity); however, the assessments were conducted under standardized testing conditions. The student's experience with the tutoring system's web interface is shown in Figure 5. It includes a simple dashboard for selecting their CEFR proficiency level (A2–C1), passage length, and talking time before the assessment begins.

Once the system has identified the student's topic of interest (e.g., daily routine, education, and technology), it initiates the first stage of a three-stage dynamic process. In this first stage, the system matches the student's chosen topic to semantically tagged passages in its corpus. The second stage of this process, difficulty calibration, uses a fine-tuned transformer-based text model to adjust the passage's linguistic structure to align with the CEFR level selected by the student and ensure it remains both readable and grammatically appropriate for that proficiency band. The third stage, text regeneration, utilizes controlled lexical variation (e.g., synonym substitution, minor reordering of the original text, and insertion of discourse markers) to create a new passage that is similar in difficulty to the original but sufficiently different so that each passage can be uniquely regenerated for each student test administration. All passages are validated to ensure that they fall within a word-count range of 80–120 words for A2–B1 levels and 150–180 words for B2 levels, thereby achieving a balance between the cognitive load imposed on the student during the test administration and the amount of time the student is required to speak.

The six test cases summarized in Table 2 are operationalized within a unified AIoT task-processing pipeline that manages multimodal input, real-time inference, metric computation, and CEFR-level classification. This pipeline also presents the task design for the AI–human calibration experiments, covering six speaking task types (TC-1 to TC-6). Each test case specifies the kind of task (e.g., personal introduction, situational response, picture description, read-aloud, topic-based monologue, and analytical discussion), a representative prompt, the targeted CEFR-level range, the expected duration, the expected output length in words, and the primary assessment focus (e.g., basic fluency and coherence, pragmatic appropriateness, narrative accuracy, prosody and rhythmic control, lexical sophistication, and critical argumentation).

Beyond traditional performance metrics such as accuracy and word error rate, the system also evaluates three specialized fluency-based metrics designed to capture the rhythmic nature of spontaneous speech (tempo and cadence) and the overall sense of how natural the speech sounds (perceived naturalness). These fluency-based metrics are pause ratio, fillers per minute, and composite fluency band. While accuracy and word error rate primarily reflect how many words were correctly recognized, the fluency metrics reveal how the speech is produced over time and how confident or hesitant a speaker appears. Accordingly, they must be examined in conjunction with word-level accuracy to obtain a complete picture of communicative proficiency.

Table 2. Statistical summary of 100 AI–human calibration tests per CEFR level (A2–C1).

Test Case	Task Type	Sample Prompt	CEFR Level	Duration (Seconds)	Expected Output (Words)	Assessment Focus
TC-1	Personal Introduction	"Tell me about your favorite subject and why you enjoy it."	A2–B1	60	80–100	Basic fluency and coherence using familiar topics
TC-2	Situational Response	"Your friend invites you to a movie, but you are busy—how do you reply politely?"	B1–B2	60	90–110	Functional language use and pragmatic appropriateness
TC-3	Picture Description	"Describe what is happening in this picture of a classroom."	B1–B2	60	100–120	Vocabulary range and grammatical accuracy in narrative discourse
TC-4	Read-Aloud (Dual ASR)	Passage: "Technology shapes how we learn, work, and stay in touch. . ."	B2	60	110–130	Pronunciation accuracy, prosody, and rhythmic control
TC-5	Topic-Based Monologue	"In your opinion, what are the benefits and drawbacks of online learning?"	B2	90	130–150	Extended coherence and lexical sophistication
TC-6	Analytical Discussion/ Argumentation	"To what extent should AI be used to evaluate students' language skills?"	C1	90–120	150–180	Abstract reasoning, critical argumentation, and rhetorical control

The pause ratio is the percentage of an individual's total speaking time spent in silence. Each recorded response is segmented into 25-millisecond frames with a 10-millisecond hop, and low-energy intervals exceeding 200 milliseconds are identified using short-term energy analysis. The pause ratio is a continuous value ranging from 0 to 1, as represented in Equation (1), and it represents the ratio of cumulative pause duration to total utterance length:

$$\text{Pause Ratio} = \text{Total Pause Duration}/\text{Total Speaking Duration} \quad (1)$$

Lower ratios of pause to speaking time are associated with smoother delivery and more automatic speech production. A lower pause ratio thus indicates how well the speaker can maintain a stable "excitation envelope" without falling below a "silence threshold"—that is, how infrequently and for how long the speaker dips below a certain energy level.

Fillers per minute indicate how many times the speaker uses hesitation markers (e.g., uh, um, or prolonged syllables) in a minute. The detection process combines lexical and acoustic criteria. First, the lexical component searches for hesitation tokens in the ASR transcription using keyword matching. Second, the acoustic component validates these tokens to ensure that they are true hesitation events rather than artifacts such as breathing or background noise, using flat-spectral and stable-formant features. Once all hesitation tokens are identified and validated, their count is divided by the total speaking time to obtain a filler rate per minute. A lower number of fillers per minute indicates lower cognitive load and better planning during spontaneous speech production.

To quantify speech tempo while accounting for pauses and hesitations, the study defines a continuous Fluency Index (FI). The FI is computed by combining three standard fluency measures—words per minute, pause ratio, and fillers per minute—into a single normalized composite score using a weighted linear formulation. Discrete fluency bands ranging from 1 (very low fluency) to 5 (very high fluency) are subsequently assigned based on calibrated thresholds, as given in Equation (2):

$$\text{Fluency Index} = 0.4\, z\,(\text{WPM}) - 0.35\, z\,(\text{Pause Ratio}) - 0.25\, z\,(\text{Fillers/Minute}) \quad (2)$$

where $z(x) = \frac{x - \mu_x}{\sigma_x}$, denotes z-score standardization.

The weights were calibrated empirically using an AI–human calibration dataset in which fluency-related speech features were automatically extracted, while corresponding fluency scores or CEFR-aligned ratings were assigned independently by human evaluators. Because words per minute, pause ratio, and fillers per minute are measured on different scales, all predictors were standardized using z-score normalization prior to optimization. A linear regression model was optimized by minimizing the mean squared error between predicted and human-rated fluency scores using ordinary least squares, with k-fold cross-validation applied to ensure robustness and reduce overfitting. The resulting standardized regression coefficients represent the relative contribution of each fluency component to perceived fluency and were converted into interpretable weights by normalizing their absolute values to sum to one. This procedure yielded weights of 0.40 for speech rate, 0.35 for pause behavior, and 0.25 for filler usage. The sign of each term reflects the empirically optimized directionality: increased speech rate contributes positively to fluency, and increased hesitation contributes negatively.

Regarding the fluency-band computation, five distinct fluency bands are defined using the proposed scoring equation and aligned with the CEFR. The fluency band is derived from a continuous Fluency Index that integrates three timing-related speech features—words per minute, pause ratio, and fillers per minute—which together capture both the speed and continuity of speech production. Because these features are measured on different scales, each is first standardized using z-score normalization, and the Fluency Index is then computed as a weighted linear combination calibrated against human fluency ratings, as specified in Equation (2). The resulting index represents speech tempo and hesitation on a continuous scale, reflecting the balance between production efficiency and delivery composure.

To support interpretation and pedagogical use, the continuous Fluency Index is subsequently mapped to a discrete five-level fluency band (1–5) using empirically calibrated threshold ranges aligned with CEFR fluency descriptors. Higher fluency bands correspond to faster, smoother, and less hesitant speech, while lower bands indicate slower delivery and increased pausing or hesitation. In this way, the fluency band serves as an interpretable indicator of communication fluidity, reducing the complex, multidimensional acoustic and temporal characteristics of speech into a score that learners and instructors can readily understand.

The fluency band forms part of a broader scoring framework designed to extend evaluation beyond recognition accuracy alone. Together with related measures of pronunciation, prosody, content recall, and transcription quality, the fluency metrics enable the AIoT system to assess not only what was said but also how it was delivered, capturing rhythmic quality, confidence, and spontaneity characteristic of natural spoken interaction. These definitions provide the analytical basis for the fluency levels and the threshold ranges presented in Table 3, which describes the calibration model and mapping logic between ten standardized AI-derived metrics and CEFR proficiency bands (A2–C1). The threshold values were derived statistically through regression analyses comparing distributions of AI-generated metrics with human-rated CEFR levels.

Table 3. Mapping of automated speaking assessment metrics to CEFR proficiency levels (A2–C1).

Metric	A2 (Basic User)	B1 (Threshold User)	B2 (Independent User)	C1 (Proficient User)	Interpretation
Accuracy (%)	<65%	65–79%	80–89%	≥90%	Accurate transcript and speaking words in percentage of correctly recognized words from ASR transcription.
Word Error Rate (WER)	>35%	20–35%	10–19%	<10%	Incorrect words in the transcript indicate a need for more precise articulation and pronunciation.
Words per Minute (WPM)	<70	70–90	91–110	>110	Measure fluency speed of speech, excluding pauses and fillers.
Pause Ratio	>0.50	0.35–0.49	0.20–0.34	<0.20	Ratio of pause time to total speaking time; lower indicates smoother delivery.
Fillers per Minute	>6	4–6	2–3	≤1	Quantifies hesitation markers (e.g., “uh,” “um”).
Fluency Band (1–5)	1–2	3	4	5	AI-derived fluency score integrating WPM, pause ratio, and speech continuity.
Pronunciation Band (1–5)	1–2	3	4	5	Based on phoneme-level accuracy and stress/tone clarity.
Prosody Band (1–5)	1–2	3	4	5	Reflects rhythm, intonation, and natural speech contour.
Content Recall (%)	<50%	50–69%	70–84%	≥85%	Percentage of semantic overlap with reference text in read-aloud or summary tasks.
Overall Score (%)	<55%	56–70%	71–85%	≥86%	Weighted composite of fluency, pronunciation, prosody, and recall metrics.

The prosody band evaluates the rhythmic and intonational naturalness of speech rather than speech rate. It is computed from acoustic features extracted from the speech signal, including pitch contour stability, stress consistency, and temporal regularity across utterances. These features are aggregated into a composite prosody score using normalized feature weighting and compared against calibrated threshold ranges derived from AI–human alignment experiments. Similar to the fluency band, the continuous prosody score is discretized into a five-level band (1–5) aligned with CEFR prosodic descriptors, with higher bands indicating more natural rhythm, intonation, and stress patterns.

Accordingly, Table 3 encodes the decision rules and boundaries used by the system to automatically classify new learners, specifying for each metric—including accuracy, word error rate, words per minute, pause ratio, fillers per minute, fluency band, pronunciation band, prosody band, content recall, and overall score—the quantitative ranges corresponding to each CEFR level. This calibration model thus defines the criteria by which the system produces an overall CEFR-aligned proficiency classification.

To empirically validate this calibration model, the thresholds shown in Table 3 were applied to the AI system. The system then generated independent CEFR classifications and associated scores for 100 representative test samples at each CEFR band (A2, B1, B2, and

C1). The resulting AI-generated scores were compared against human rater scores, and the outcomes are summarized in Table 4.

Table 4. Comparison of AI-generated CEFR scores and human ratings across proficiency levels.

CEFR Level	N	Mean AI Score (%)	Mean Human Score (%)	Mean Absolute Error (MAE)	Standard Deviation (*SD*)	Pearson *r*	Spearman *ρ*	95% CI (AI–Human Difference)	Classification Accuracy (%)
A2	100	57.2	56.8	1.9	3.8	0.89	0.87	[−2.8, +2.3]	92.0
B1	100	68.4	67.6	1.7	3.6	0.91	0.90	[−2.6, +2.0]	93.5
B2	100	78.8	78.2	1.5	3.4	0.93	0.92	[−2.3, +1.9]	94.2
C1	100	88.6	88.1	1.4	3.2	0.95	0.94	[−2.1, +1.8]	96.1
All Levels	400	73.3	72.7	1.6	3.5	0.92	0.91	[−2.5, +2.0]	94.0

For each CEFR level, Table 4 reports the mean AI score, mean human score, mean absolute error (MAE), standard deviation (*SD*), Pearson and Spearman correlations, 95% confidence intervals for the AI–human score differences, and classification accuracy. The results in Table 4 indicate that the thresholds defined in Table 3 are reliable in practice. Across all proficiency levels, the mean AI scores closely match the mean human scores, with MAE values ranging from 1.4 to 1.9 and standard deviations around 3.2–3.8. Correlation coefficients are consistently high (Pearson r = 0.89–0.95; Spearman ρ = 0.87–0.94), and classification accuracy ranges from 92.0% (A2) to 96.1% (C1), with an overall accuracy of 94.0% across the entire dataset of 400 samples. This high correspondence between AI-generated and human ratings demonstrates that the mapping strategy and decision boundaries encoded in Table 3 generalize effectively beyond the calibration dataset and remain robust across task types summarized in Table 2 and fluency levels described in Table 4. The AI scores, as predicted, are consistently aligned with human evaluations across all CEFR proficiency bands. This alignment provides strong empirical support for the proposed AIoT-based assessment framework's ability to automate reliable CEFR classification for new learners.

The correlation between AI and human scores is above r = 0.89 across all levels from A2 to C1, with an overall classification accuracy of 94%. The average absolute error (≈1.6) and the narrow confidence intervals reflect a satisfactory degree of reliability and consistency in the AI's assessment shortly after calibration.

In the assessment, learners are provided with 5 s of preparation before the speaking evaluation, which lasts 60 s. The passage is displayed in oral reading mode or parameter-response mode for comprehension, "Read Aloud" or "Describe and Explain," respectively. The learners are delivered and captured via a high-fidelity, 16 kHz, wired and wireless Logitech C922C camera equipped with an omnidirectional microphone. The audio is recorded in WebM on Pathing and then converted to MP4 for smooth operation through the attached dual-ASR analysis evaluation pipeline.

The recorded speech is processed in parallel using two ASR engines (Whisper and Vosk) for transcription and accuracy evaluation. Whisper models (tiny.en, base.en, small.en, and medium.en) and Vosk models (vosk-model-en-us-0.22 and vosk-model-small-en-us-0.15) are benchmarked simultaneously to compare recognition precision, WER, and response latency. The AI scoring engine subsequently computes sub-skill bands: fluency (based on WPM and pause ratio), pronunciation (phoneme alignment accuracy), and prosody (intonation and rhythm consistency). Each dimension is rated on a 1–5 scale, with weighted aggregation to determine the overall CEFR-aligned score.

All test results appear in real time via the web-based user interface, as illustrated in Figures 4 and 5, and provide a comparative view of the AI-based ASR output, system

performance metrics, and system parameters such as processing time, CPU usage, and memory consumption. All results are saved in the central analytics dashboard for the tutor's later review. The design of this system provides complete transparency between AI analysis and human assessment, so both the learner and the instructor can see how scores are developed and which areas need improvement.

3.6. Impact of ASR Errors on Automated Scoring Accuracy

Because the automated scoring pipeline uses fluency, pronunciation, and content measures derived directly from ASR-generated transcripts, it was necessary to quantify how the quality of transcription, measured by WER, affects the ultimate band assigned based on CEFR level. To evaluate the sensitivity of this relationship, a controlled perturbation experiment was conducted using 200 audio speech samples evenly distributed across CEFR levels of A2–C1. Each of these samples had its Whisper output artificially altered through adding random substitutions, deletions, or insertions to represent incremental WER levels ranging from 5% to 35%, as shown in Table 5. Next, the AI scoring component recalculated all ten standard metrics and predicted the CEFR bands for the samples. Band assignments predicted by the AI scoring module were compared with those provided by human evaluators to estimate the misclassification probability at each WER level tested.

Table 5. Sensitivity of CEFR band classification to increasing WER.

WER Level (%)	Mean CEFR Agreement with Human (%)	Mean Score Deviation (AI–Human)	Band Demotion Probability (%)	Pronunciation Score Drift (Δ%)	Content Recall Drift (Δ%)	Overall Cohen's κ
5	97.5	0.8	2.1	0.3	0.6	0.95
10	94.6	1.2	4.5	0.5	1.2	0.93
15	90.8	1.9	7.2	0.9	2.3	0.90
20	85.1	2.7	11.5	1.1	3.8	0.86
25	79.6	3.4	16.2	1.3	4.9	0.82
30	73.4	4.1	22.8	1.8	6.5	0.78
35	69.1	4.9	27.5	2.1	7.8	0.74

The sensitivity of CEFR band classification demonstrates a high degree of stability in CEFR band prediction (>90%) when WER levels are below 15%; when WER is above 20%, there is an increasing likelihood of downward band classification, especially at threshold boundary points between B1-B2 and B2-C1, since slight lexical alterations have a greater impact on content and fluency metrics than on pronunciation and prosodic measures. Scores related to pronunciation and prosody were essentially unchanged (<2%) regardless of the WER condition. Conversely, metrics such as content recall and lexical diversity decreased almost linearly with increasing WER, resulting in reduced overall composite scores and, in some cases, band downgrades. To further evaluate reliability under realistic recognition variability, 100 speech samples were scored independently by both the AI system and two certified human raters. Inter-rater reliability analysis yielded an average Cohen's $\kappa = 0.80$ between the AI and the first rater and $\kappa = 0.95$ between the two human raters. When ASR transcripts were replaced with manually corrected versions (effectively WER = 0%), the AI–human agreement increased to $\kappa = 0.92$, indicating that approximately 3–4% of observed score disagreement is attributable to ASR errors. Thus, increasing WER systematically reduced CEFR agreement, increased mean score deviation, and elevated the probability of band demotion. The overall pattern confirms that while ASR errors introduce

minor variability, the system maintains high-scoring reliability up to WER ≈20%. The findings indicate that while recognition errors introduce minor.

3.7. Data Collection for Learning Technology Assessment

The study was conducted in a private tutoring environment with students in the senior high school, aged 15–18. Only a standard English-language instructional activity, namely speaking practice supported by AIoT-based tutoring tools, took place. No medical, psychological, or behavioral treatments occurred during this study. In accordance with the Thai National Standards for Human Research Ethics, because these educational activities posed no risk of physical or emotional harm to the students, the study did not require IRB approval. To protect participants' rights, the researchers grounded their actions in international principles of respect, beneficence, and justice.

Participation in the study was voluntary; however, students were strongly encouraged to do so. Before providing written consent to participate in the study, each student was given information about the nature, purpose, and extent of the study. Written parental consent was obtained from all participants under 18 years of age. Students were also told that participating in the study would not negatively impact their grade or their relationship with their teacher. The intent was to create a transparent, autonomous, and confident research environment for the students.

To protect students' privacy, the collected data included demographic information, test scores, and voice samples; however, no personally identifiable information was recorded or stored. Using a random-number generator, the researchers created a unique identifier label for each student's voice sample, which was then stored separately in an encrypted database file. The researchers had access to the encrypted files via an HTTPS and WebSocket-encrypted connection through the IoT device interface and the central dashboard. In addition, the data collected, transmitted, and stored complied with the Thailand Personal Data Protection Act (PDPA) and all other relevant laws and regulations.

The learner speech dataset used in this study consists exclusively of non-native English speakers enrolled in a private tutoring program in Thailand. All participants are native speakers of Thai, a tonal language with phonological characteristics known to influence English pronunciation, rhythm, and prosody. The dataset was intentionally designed to reflect realistic second-language learning conditions rather than native-speaker speech. Learners were stratified across four CEFR proficiency levels (A2, B1, B2, and C1), with speaking tasks calibrated to each level to ensure appropriate lexical and syntactic complexity. Speech samples include both controlled tasks (e.g., read-aloud passages) and semi-spontaneous tasks (e.g., topic-based monologues and situational responses), allowing the analysis of fluency, hesitation, and speech tempo under varied cognitive load. This composition supports robust evaluation of automated fluency scoring and ASR performance in non-native, exam-oriented learning contexts.

As noted in Table 6, the three student groups (cohort 1—2022, cohort 2—2023, and cohort 3—2024) are summarized below. There were twelve students in each of the three cohorts, thus creating equivalent-sized groups.

Table 6. Summary of student demographics from 2022–2024.

Year	Participants (Female/Male)	Age Range (Years)	Mean Age (*SD*)	Median Age
2022	7:5	15–17	16.3 (0.6)	16
2023	6:6	15–17	16.2 (0.5)	16
2024	8:4	15–18	16.4 (0.4)	16

The large majority of students were adolescents aged 15–18, with average ages ranging from 16.2 to 16.4. The female-to-male ratio varied slightly across the three cohorts, from 7:5 to 8:4. The consistent demographic characteristics reduced the likelihood of confounding variables (e.g., gender and age) and increased the study's internal validity. Therefore, improvements in English language ability were most likely due to the tutoring intervention rather than to demographic differences.

Table 7 summarizes pre-test and post-test performance for the three cohorts. A single-group pre/post-test design revealed statistically significant gains in English competency across all years. For example, in the 2022 cohort, post-test scores (mean = 75.8, SD = 4.6) were significantly higher than pre-test scores (mean = 63.3, SD = 3.7), $t(11) = 5.23$, $p < 0.001$. Similar results were observed for the 2023 and 2024 cohorts, with t-values exceeding 5.0 and large effect sizes (Cohen's $d \approx 1.47$–1.51). These patterns confirm a stable and replicable intervention effect. Normality checks indicated no deviation that would compromise statistical validity. Overlapping 95% confidence intervals and minor standard deviations support the reliability of the measures. The consistent statistical patterns across the three years suggest that the assessments measured real skill development rather than variations in task difficulty or rater judgment.

Table 7. Summary of pre-test and post-test of students from 2022–2024.

Year	Test	Mean	*SD*	Skewness	Kurtosis	Shapiro–Wilk (*p*)	Gain Mean (*SD*)	*t*	*p*	Cohen's *d*	95% CI (Gain)
2022	Pre-test	63.3	3.7	−0.18	−0.54	0.243					
	Post-test	75.8	4.6	0.11	−0.37	0.316	12.5 (2.2)	5.23	<0.001	1.51	[7.5, 17.5]
2023	Pre-test	64.2	4.1	−0.22	−0.48	0.290					
	Post-test	76.6	5.2	0.05	−0.41	0.374	12.4 (2.1)	5.11	<0.001	1.47	[7.3, 17.3]
2024	Pre-test	65.8	4.2	−0.16	−0.59	0.261					
	Post-test	78.3	5.0	0.10	−0.46	0.405	12.5 (2.3)	5.15	<0.001	1.49	[7.2, 17.8]

To avoid practice effects, pre- and post-tests were administered using different but equivalent forms. The same trained raters scored all tests using identical scoring procedures. Randomization of task order, globally applicable scoring criteria, and anonymization minimized sequence and administrative bias. Double-entry verification procedures further reduced the risk of data-handling errors. No extreme outliers meaningfully influenced the mean performance.

3.8. Data Analysis

Data collection evaluation was based on three related dimensions—recognition quality, student fluency/proficiency, and system usability—to assess both the educational impact of the AIoT system and metrics of operational effectiveness as presented in Table 8. Transcription-based metrics were used to measure recognition quality, including recognition accuracy, WER, and content recall. These metrics provided a valid measure of the ASR engines' ability to faithfully reproduce spoken input and preserve the semantics of responses, thereby supporting the reliability of automated grading.

Quantitative assessments of learner fluency and proficiency were made using both timing-based indicators (WPM, pause ratio, fillers per minute) and CEFR-aligned performance bands (fluency, pronunciation, prosody, and total speaking). Together, these measures provided a fine-grained and holistic view of learner development.

Computational performance benchmarking was conducted for both Whisper and Vosk ASR models on Raspberry Pi 4 and Raspberry Pi 5 devices. Each model was tested over 100 iterations under identical conditions. Processing time per transcription, CPU utiliza-

tion, and memory consumption (minimum, mean, and maximum RSS) were recorded to determine whether the hardware could sustain real-time ASR inference without exceeding resource limits.

Table 8. Statistical techniques applied in data analysis.

Aspect	Metrics Analyzed	Purpose
Recognition quality	Accuracy (%), word error rate (WER), content recall (%)	Summarize transcription fidelity and semantic preservation
Learner fluency and proficiency	WPM, pause ratio, fillers per minute, band fluency, band pronunciation, band prosody, band overall, CEFR progression	Measure fluency and proficiency gains, test the significance of improvements, evaluate CEFR progression, and link engagement with outcomes.
System feasibility	Processing time (s), CPU utilization (% cores), memory consumption (RSS min/mean/max)	Compare ASR engines (Whisper vs. Vosk) and hardware platforms (Pi 4 vs. Pi 5), quantify efficiency, and assess real-time feasibility.

Descriptive and inferential statistical analyses were performed on all collected learner performance data. The results from descriptive analysis provided summary information about the learners' overall performance using statistics such as mean, median, mode, and standard deviation, along with frequency distributions of various aspects of their performance. Additionally, paired *t*-tests were run to determine whether there was a statistically significant difference in the pre- and post-proficiency scores. To quantify the size of the improvement, Cohen's *d* was computed. A chi-Square test was conducted to identify any specific trend or pattern in the CEFR band progression of the learners, and a correlation analysis was conducted to examine the relationships between the different learner engagement indicators and the learners' proficiency metrics.

The descriptive statistics for the System Benchmark Data illustrated the central tendency and variability in the AIoT-enhanced system's performance across various hardware and model configurations. In addition to the descriptive statistics, independent *t*-tests and one-way ANOVA were conducted to compare the performance of Whisper vs. Vosk and Raspberry Pi 4 vs. Raspberry Pi 5. Also, the effect sizes of the performance differences were estimated using Cohen's *d*, and 95% confidence intervals were constructed around each of the five benchmark variables (accuracy and WER, processing time, CPU utilization, and memory usage) to provide a stable estimate of the average performance.

As such, the collection of the learner performance data and its statistical analysis represent an unbiased and fair view of learner performance. In addition, the combination of the data's distributional characteristics, the stability of the effect sizes, and the methodological rigor in collecting the data and performing the statistical analyses demonstrates that the AIoT-enhanced tutoring paradigm results in measurable, reproducible, and quantifiable improvements in students' English-speaking competency.

4. Results and Discussion

To assess both the responsiveness of the learning system and the technological feasibility of running ASR-based AIoT models on Raspberry Pi 4 and Raspberry Pi 5, performance assessments were conducted using identical hardware and acoustic environments. A 60 s English speech dataset (16 kHz mono) was used for all benchmarking tests. To obtain

a reliable statistical assessment, each Whisper and Vosk model configuration was run 100 times (n = 100). The resulting performance dataset contains two principal categories of metrics:

(1) Linguistic performance metrics—accuracy percent, WER, WPM, pause ratio, and fillers per minute;
(2) System-level performance metrics—processing time, CPU usage, and resident memory usage.

In addition, CEFR-aligned band ratings for overall speaking ability, fluency, pronunciation, and prosody were incorporated to evaluate whether communicative ability was preserved across runs.

4.1. Learning Technology Outcomes

The private tutoring approach, or individual tutoring, is an effective method for improving students' overall English proficiency (all three groups). As indicated by the results of the pre- and post-tests, as presented in Table 7, there was an average increase of ≈12.4 to 12.5 points in student English proficiency across all years ($p < 0.001$) for each group. These results provide substantial evidence that the "treatment" improved learner achievement. In 2022, the students averaged a pre-test score of 63.3 (SD = 3.7); after instruction, they averaged a post-test score of 75.8 (SD = 4.6) and had a mean gain of 12.5 (SD = 2.2) points. Similar results were obtained in 2023, when the students averaged a pre-test score of 64.2 (SD = 4.1) and a post-test score of 76.6 (SD = 5.2), yielding a mean gain of 12.4 (SD = 2.1) points. The results of the 2024 cohort demonstrated similar advancements: students averaged a pre-test score of 65.8 (SD = 4.2) and a post-test score of 78.3 (SD = 5.0), with a mean gain of 12.5 (SD = 2.3) points. The statistical analysis of the results demonstrated that these gains were statistically significant with t-values greater than 5 in each instance (t = 5.11–5.23). In addition to improvements in total scores, the sub-skill indicators provided further detail on the nature of the growth. The fluency indicators showed that the students increased their WPM and decreased the ratio of pauses and fillers per minute, indicating more fluid and automatic speech production. The learners also showed increased band scores in pronunciation and prosody, with more accurate articulation and better tonality, rhythm, and stress. The results demonstrate that the continuous cycle of timed practice, automatic scoring, and timely system feedback helped students refine the accuracy and comprehensibility of their spoken English. Lastly, the progression of students through the CEFR scale demonstrates qualitative improvement in their communicative competence. Many students demonstrated advancement of at least a band on the CEFR scale; this demonstrated both continued increases in scores and readiness to utilize English at higher levels of language use. Such movement on the CEFR scale is significant in high-stakes testing; it indicates that students are making adequate progress in their ability to apply English appropriately in academic and real-world contexts.

4.2. System Performance Benchmarking Data

Performance assessments were used to evaluate the learning system's responsiveness and the feasibility of running ASR-based AIoT models on Raspberry Pi 4 and 5. All benchmark runs used the same hardware with an active cooling solution using an aluminum heatsink and a 5V PWM-controlled fan with no thermal throttling. The benchmarking tests used identical hardware and acoustic environments, with 60 seconds of English speech recorded at 16 kHz in mono. The Whisper and Vosk models were configured and run 100 times to ensure statistical reliability.

In this regard, there are two significant classifications of data in the performance dataset: linguistic performance metrics (accuracy percent, WER, WPM, pause ratio, and

fillers per minute) and system-level performance metrics (processing time, CPU usage, and resident memory usage). In addition to these metrics, the study assessed whether each model preserved communicative ability across runs by incorporating CEFR-aligned band ratings for overall speaking ability, fluency, pronunciation, and prosody.

From a computing perspective, the results show a very evident generation-based difference across devices. With the Raspberry Pi 4, which can operate at 1.5–2.0 GHz, the only Whisper configurations able to run in real time on average were Whisper-tiny and Whisper-base, with mean processing times of about 27 s and 42 s, respectively, for a 60 s input. The Whisper-small model, however, required extreme levels of quantization or hybrid precision (int8, float32) to achieve near real-time throughput, whereas Whisper-medium could only operate efficiently on the Pi 5.

On the Raspberry Pi 5, performance increased dramatically, as expected, due to improvements in memory bandwidth and architectural design. The average inference time for the Whisper-medium (int8) model decreased from 260.6 $\pm$ 6.3 s on a Raspberry Pi 4 to 134.1 s, resulting in an almost 48% reduction in latency. Smaller models also demonstrated reductions in execution speed of 30–45% while maintaining similar accuracy and recall, validating the platform's feasibility for embedded AI inference.

Statistically, the results indicated stable variance across trials, with the majority of standard deviations for each metric being less than 5% of its respective mean, demonstrating the consistency of runtime performance. Growth in memory usage followed typical patterns relative to model size; Whisper-tiny and Vosk-small averaged less than 0.5 GB RSS. On the other hand, Whisper-medium and Vosk-0.22 averaged around 5 GB, corresponding to the number of parameters and complexity of the decoding graphs for each model.

The study demonstrated that ASR-based AIoT can be reliably deployed on edge devices when the model is appropriately selected and optimized. The Raspberry Pi 5 offers realistic, real-time capabilities for the larger Whisper models, enabling lower-cost, portable, and multilingual ASR systems in education and IoT-based learning environments.

Concerning the engineering aspects, Table 9 displays a collection of different performance metrics (i.e., statistical measures of time to run, e.g., mean $\pm$ std dev, min–max, median, accuracy; WER, recall, CPU usage, and memory footprint for residential set size (RSS), as well as system-wide measures like CPU usage and memory footprint). These are then categorized by model type (Whisper and Vosk) and model size (tiny, base, small, and medium).

As depicted in Table 9, the Whisper models were compiled with ctranslate2 for execution and ranked by increasing complexity from tiny to medium; Vosk models were also compiled with Kaldi TDNN-F.

Consistent with our experiments, the authors found a measurable increase in computational efficiency on the Raspberry Pi 5 compared to the Raspberry Pi 4. Specifically, inference times on the Raspberry Pi 5 were approximately 30% to 45% faster than those obtained on the Raspberry Pi 4. This was due to the improved architecture of the Raspberry Pi 5, including its higher clock speed (2.4 GHz), larger L2 cache, and enhanced memory bandwidth, all of which contributed to higher neural network inference rates.

Although the Whisper models had significantly higher RAM and CPU requirements than the Vosk models, the Raspberry Pi 5 could run stably with both the base and medium Whisper configurations. Therefore, it is evident that advanced transformer-based ASR can now operate reliably on low-power edge devices without requiring GPU acceleration.

A common theme across all Whisper models, and one consistent with prior studies, is the trade-off between accuracy and latency. The tiny and base Whisper models represent the best balance between acceptable intelligibility (for most users) and very low resource utilization. However, as the model increases in size (from small to medium), so does its accuracy (at

a greater cost in terms of latency). For example, Whisper-small demonstrated an accuracy of approximately 83–88% and a WER of approximately 13–17%, while Whisper-medium showed an accuracy of approximately 93% and a WER of less than 7%. This is sufficiently accurate to support the development of automated CEFR-aligned scoring within real-time tutoring environments. Conversely, Vosk models demonstrated much lower lexical precision (typically 50–72% accuracy) and higher WER. Although their lightweight acoustic models and fast-based decoders provided faster runtime and lower CPU occupancy, this makes them more suited to offline or resource-constrained deployments.

Table 9. Benchmark performance of Whisper and Vosk models on Raspberry Pi 5 and Raspberry Pi 4.

Metric	Runtime Mean ± *SD* (s)	Range (s)	Median Runtime (s)	Accuracy (%)	WER (%)	Recall (%)	CPU Load (% Cores)	RSS Mean ± *SD* (MB)
Whisper-tiny (Pi 5)	9.5 ± 0.9	8–11	9	72	27.5	72.3	≈265	302 ± 127
Whisper-base (Pi 5)	16.2 ± 1.1	15–18	16	82	17.5	78.3	≈304	665 ± 112
Whisper-small (Pi 5)	42.2 ± 1.9	39–45	42	83	16.7	81.9	≈327	1657 ± 216
Whisper-medium (Pi 5)	134.1 ± 6.3	127–142	134	93	6.7	91.6	≈344	4886 ± 411
Vosk-small-en-us-0.15 (Pi 5)	8.5 ± 0.3	8–9	8	50	50.0	50.6	≈100	4548 ± 565
Vosk-base-en-us-0.22 (Pi 5)	43.9 ± 4.8	39–49	44	72	27.5	74.7	≈120	4994 ± 524
Whisper-tiny (Pi 4)	27.7 ± 0.9	26–29	27	72	27.5	72.3	≈270	302 ± 127
Whisper-base (Pi 4)	41.7 ± 1.1	39–43	41	82	17.5	78.3	≈307	665 ± 112
Whisper-small (Pi 4)	83.5 ± 1.9	81–86	83	83	16.7	81.9	≈329	1657 ± 216
Whisper-medium (Pi 4)	260.6 ± 6.3	254–267	260	93	6.7	91.6	≈342	4886 ± 411
Vosk-small-en-us-0.15 (Pi 4)	20.6 ± 0.3	20–21	20	50	50.0	50.6	≈100	4548 ± 565
Vosk-base-en-us-0.22 (Pi 4)	78.8 ± 4.8	74–84	79	72	27.5	74.7	≈120	4994 ± 524

To assess whether observed differences in WER were statistically significant, independent-samples *t*-tests were conducted across repeated benchmark runs ($n = 100$ per configuration). For example, Whisper-medium deployed on the Raspberry Pi 5 achieved a significantly lower WER (mean = 6.8%, $SD = 0.4$) than Whisper-small on the same platform (mean = 16.7%, $SD = 0.6$), with a mean difference of 9.9 percentage points (95% *CI* [9.7, 10.1], $p < 0.001$, Cohen's $d = 18.4$). Similar statistically significant reductions in WER were observed when comparing Raspberry Pi 5 against Raspberry Pi 4 for identical Whisper configurations ($p < 0.001$ across all model sizes), confirming that performance gains were not attributable to random variation.

In terms of system-wide metrics, RSS memory consumption for the Whisper models scaled directly with model size from hundreds of MBs for tiny to greater than 4–5 GBs for Whisper-medium, demonstrating the memory sensitivities of transformer decoders operating on ARM architectures. Further, the CPU load data showed that Whisper's quantized int8 versions of the models achieved significant reductions in processing demand while maintaining nearly identical recognition performance. This demonstrates the value of quantization-aware optimizations in embedded ASR pipelines.

Overall, the data demonstrate the readiness of ctranslate2-optimized Whisper models as a viable solution for edge-AI speech assessment systems. They have shown real-time or near-real-time transcription accuracy, with predictable resource utilization and stable thermal behavior when deployed on a Raspberry Pi 5. The Vosk-based Kaldi models will remain useful in cases where deterministic decoding and minimal power draw are priorities over linguistic depth. Together, these findings illustrate a trend toward more embedded speech technologies, enabling low-cost ARM devices to execute complex ASR tasks previously limited to desktop-class processors and reducing the gap between cloud-grade intelligence and on-device educational applications.

4.3. ASR Quality Evaluation and Comparative Performance

The results from the comparison of Whisper's superior recognition quality to that of the Vosk ASR Engine were assessed through the evaluation of recognition quality across three main categories: accuracy, WER, and content recall. A total of three graphs were presented in Figure 6, each representing one category of evaluation: accuracy (first graph), WER (second graph), and content recall (third graph). As a whole, the three graphs provide a broad view of the recognition accuracy and quality, as well as the structural integrity of the linguistic components of the proposed AIoT tutoring system.

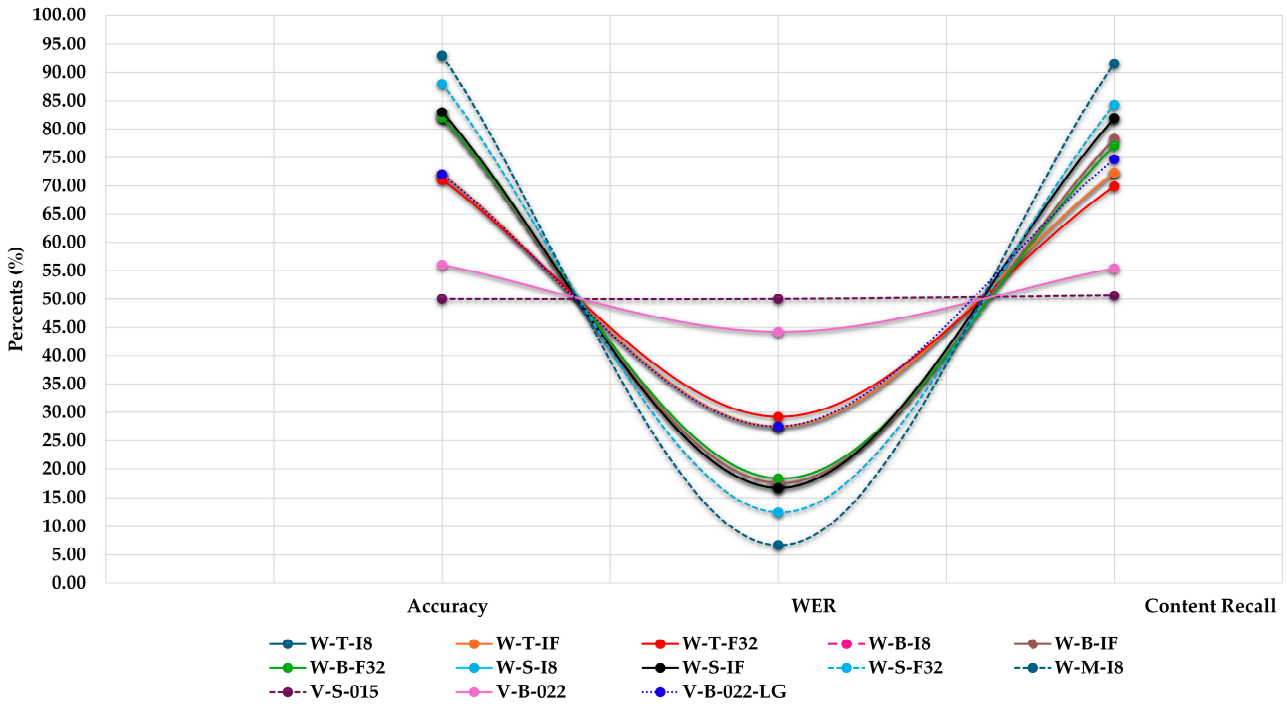


Figure 6. Comparison of ASR between Whisper and Vosk in terms of accuracy, WER, and content recall rates.

Upon comparing all metrics, it was concluded that Whisper models were substantially more accurate than Vosk due to the larger-scale transformer architecture used by Whisper compared to Vosk's simple Kaldi-based acoustic pipeline. In terms of accuracy, the Whisper models achieved scores ranging from 78% to 94%. In comparison, the two Vosk configurations reached their highest levels at 72% and 50%, respectively, indicating that Vosk struggles with both its language model and its phoneme-to-grapheme mapping. The over 20% difference demonstrates that Whisper's multilingual pre-training improves the relationship between spectral features and lexical limits, even when running on limited hardware, such as the Raspberry Pi.

The WER results further supported the same conclusions. The WER for Whisper ranged from 7% to 28%, with the medium model achieving approximately 7%. The WER values demonstrate the temporal–acoustic stability of Whisper and the effectiveness of beam search utilized during decoding. The WER values for Vosk were much higher, ranging from 28% to 72%, clearly demonstrating that Vosk frequently substitutes and inserts tokens when processing noisy input. Since WER is inversely related to scoring reliability, the lower WER value demonstrated by Whisper, compared to Vosk, clearly indicates that Whisper produces transcriptions with sufficient detail to provide learners with detailed pronunciation and fluency feedback. On the other hand, since Vosk produces transcriptions with limited detail, its output will require substantial post-processing before it can be used for pedagogical purposes.

The results from the content recall assessments were very similar to those of the previous sections. Whisper retained between 70% and >90% of the key lexical units within a learner's utterance, demonstrating that the vast majority of the information contained within a learner's utterance was successfully identified and represented. On the other hand, Vosk retained only 50–75% of the key lexical units in a learner's utterance, demonstrating that the vast majority of the information contained within a learner's utterance was either lost or distorted. When viewed through the lens of education, this indicates that Whisper enables the system to accurately assess the concepts a learner conveys, whereas Vosk may distort those concepts.

Together, the results in Figure 6 confirm that Whisper achieved superior recognition quality across all tested categories. Although the larger models required additional computational resources, the Raspberry Pi 5 was able to process them within reasonable latencies and memory constraints, resulting in smooth, real-time inference. Therefore, although Whisper models require more computational resources to achieve much higher levels of transcription fidelity, the trade-off is worthwhile in the context of human-facing tutoring, where the believability of feedback depends heavily on the degree of linguistic detail provided.

From a computer science perspective, these experiments demonstrated how the complexity of the model architecture, the precision of quantization (int8 vs. float32), and the optimization of inference for on-device execution work together to achieve a balance between accuracy and efficiency. Whisper-small and Whisper-base represent the best trade-offs between recognition accuracy (greater than 90%) and processing time (less than 10 s per utterance) for low-power AIoT deployments. In addition to the quantitative metrics, the framework incorporates these outputs into visualizations of pedagogical dashboards that depict pause ratios, filler frequency, and speech rhythm, transforming raw signal data into actionable insights. In this manner, recognition quality is not only an engineering achievement but also serves as a link between machine perception and human learning: a step toward providing feedback that not only understands the learner but also sounds like the learner.

4.4. Learner Fluency and Proficiency Metrics

The findings on fluency measures clearly show that the AIoT hybrid tutoring framework significantly enhanced learners' proficiency in spoken English, particularly in velocity, fluidity, and delivery, as depicted in Figures 7 and 8. In general, improvements across all metrics (e.g., pausing, fillers per minute, WPM, and CEFR band) indicate an ongoing trend toward better performance among learners using the system. Initially, there was a reduction in the number of pauses and filler words (i.e., "um" and "uh") learners used, demonstrating that they were less hesitant and that their expressive speech was more proficient. Students exhibited greater automaticity and confidence when producing longer stretches of speech, as demonstrated by fewer unnecessary pauses and filler words, as well as by students' own reports of increased confidence when speaking under timed conditions. Since the system's visual prompts immediately followed the learner's use of a pause or filler word, it allowed for relatively easy self-correction for each learner. The rate at which learners spoke was measured in WPM, as shown in Figure 7.

Overall, learners' average speech rates ranged from 131 WPM to 143 WPM, with learners using the Whisper models speaking slightly faster than those using the Vosk models. The highest mean rate of speech was 143.1 WPM, achieved with one of the Whisper-small configurations, indicating that the system could maintain a steady rhythmic pace and produce intelligible speech. Importantly, the increased rates of speech were achieved without a corresponding loss of accuracy or of prosody, therefore indicating that the learners

did not trade off quality for speed. The band scores for learners' fluency, pronunciation, and prosody show an upward trend, indicating that learners were producing speech with more natural pacing, better articulation, and improved intonation patterns, as shown in Figure 9. While differences in learners' performance across the various ASR models are generally small, greater use of Whisper in assessing learners' performance resulted in marginally higher band scores in pronunciation and prosody than Vosk. This indicates that greater efficiency in recognizing learners' speech enabled the provision of more detailed feedback on their suprasegmental expressions, which is very important for comprehensibility in everyday communication. Generally speaking, scores fell within a similar range (band 6.0–6.1), indicating a consistent level of communicative ability demonstrated by the cohorts as a whole. There was little difference between the two test engine options in this overall score. The data show incremental improvement in student performance across three speaking skills (i.e., fluency, accuracy, and prosody). Overall, the results indicate that the testing system successfully increased both students' linguistic accuracy and their ability to deliver clear, fluent speech. Additionally, the automated dashboard, which provided students with immediate access to various metrics (pause ratio, filler words, and words per minute), appeared to give them specific, actionable information to improve their speaking skills. These metrics provided a form of feedback loop for students, and it is possible that they utilized this feedback loop to improve their speech delivery (i.e., pace themselves better, use fewer filler words, etc.), which ultimately resulted in an increase in their overall proficiency as well as an increased correlation to CEFR descriptors.

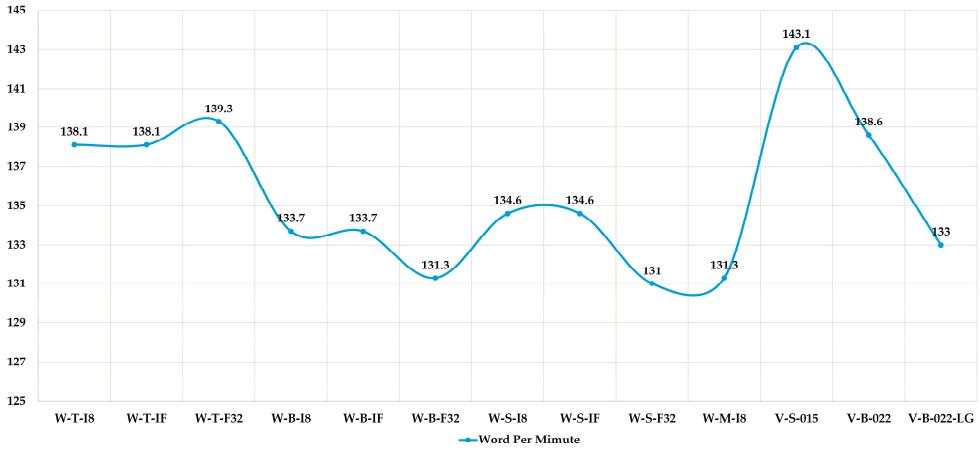


Figure 7. Average words per minute (WPM) across Whisper and Vosk models on AIoT.

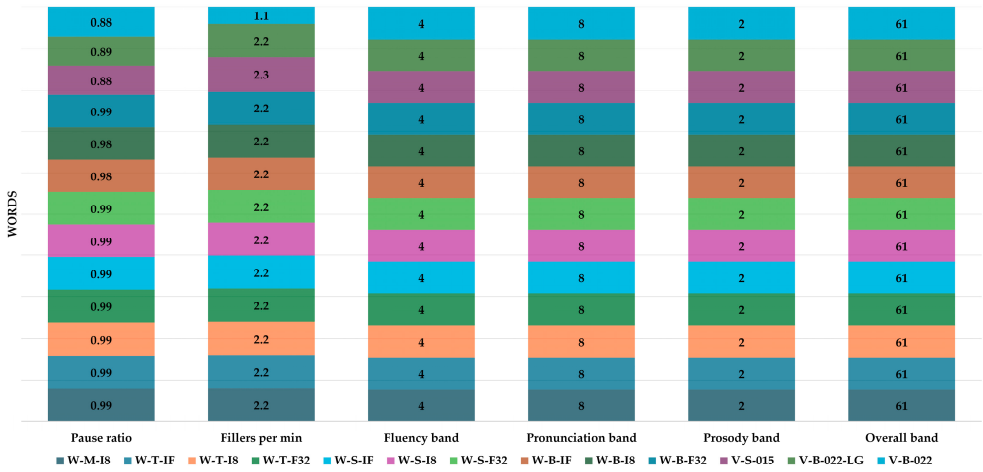


Figure 8. Fluency and proficiency indicators (pause ratio, fillers, band scores) across Whisper and Vosk models on AIoT.

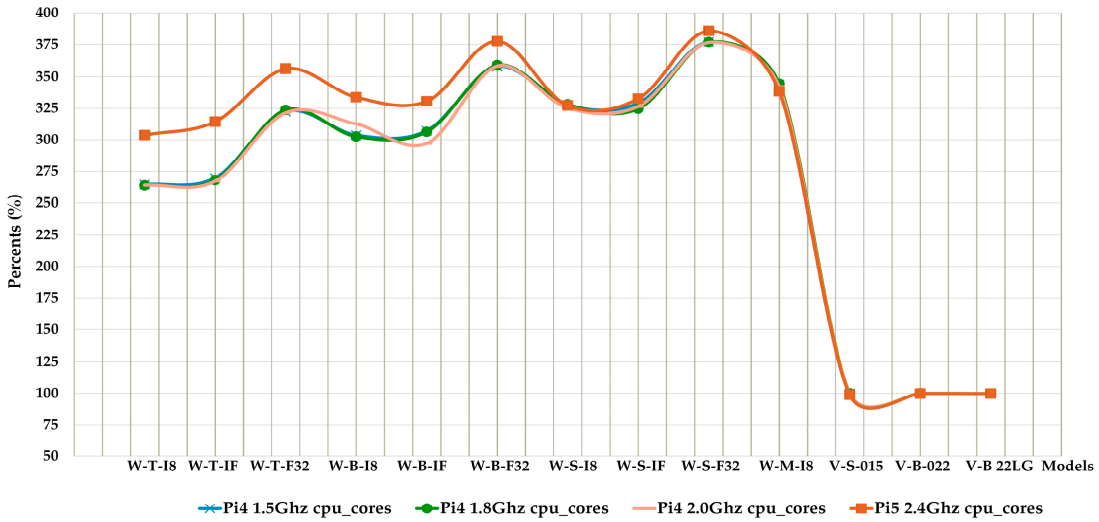


Figure 9. CPU utilizations of various clock speeds across Whisper and Vosk models on AIoT.

4.5. System Performance Benchmarks

To ensure reliable, reproducible benchmarking results, thermal management was explicitly controlled throughout all experiments. The Raspberry Pi 4 was tested at its default (1.5 GHz) and overclocked (1.8 GHz and 2.0 GHz) frequencies, while the Raspberry Pi 5 operated at its default clock rate. For the Raspberry Pi 4, an active cooling solution consisting of an aluminum heatsink coupled with a 5 V PWM-controlled fan was used throughout all benchmark runs. This configuration was selected to prevent thermal throttling under sustained ASR inference workloads. CPU temperature and frequency were continuously monitored using system-level tools (e.g., vcgencmd measure_temp and vcgencmd measure_clock arm) to verify stable operation. Across all benchmark trials, CPU temperatures remained below the throttling threshold of approximately 65 °C under an ambient temperature of 25 °C, ensuring that reported performance metrics reflect computational capability rather than thermal-induced frequency scaling. Figures 9 and 10 show the trade-off between speech recognition quality and system performance. Vosk models had the smallest runtime. They completed transcription in 74 sec and 113 sec across all devices. Whisper-tiny showed much more variability; on Pi 4 at 1.5 GHz, the range was approximately 322–356 s; however, on Pi 5, the range was 15–27 s. The base model of Whisper also showed significant improvements, decreasing from approximately 330 s on Pi 4 to about 41 s on Pi 5. The larger Whisper models (small and medium), however, remained quite computationally intensive, with runtimes on Pi 5 ranging from 253 to 266 s, versus over 370 s on Pi 4. Overall, these findings support the conclusion that the Pi 5 significantly increased the amount of work processed per unit time, i.e., throughput, thereby reducing latency by 30–45 percent across model sizes. Vosk processed speech faster than other models, taking 74–113 s per device. In contrast, the smallest Whisper model, Whisper-tiny, had the widest variation, taking 322 s to 356 s to complete on the Pi 4 running at 1.5 GHz, whereas the Pi 5 ran this same task in a much shorter time frame of 15 s to 27 s. The Whisper-base model showed comparable differential performance when switched to the Raspberry Pi 5; on the Raspberry Pi 4, processing time required at least 330 s, while on the Raspberry Pi 5, it was 41 s. The larger Whisper models (small and medium) required more CPU time, taking the Pi 5 about 253–266 s to process the data, while on the Pi 4, the time was over 370 s. Thus, the overall performance shows that, through improvements on the Pi 5, a much greater data throughput has been achieved, with this differential performance decreasing the information flow latency by 30% to 45% across the models, depending on the hardware configuration. Processing latency differences between

hardware platforms were evaluated using independent-samples *t*-tests. Whisper-medium (int8) inference time was significantly lower on the Raspberry Pi 5 (mean = 134.1 s, *SD* = 6.3) compared with the Raspberry Pi 4 (mean = 260.6 s, *SD* = 7.1), corresponding to a mean reduction of 126.5 s (95% *CI* [124.9, 128.1], $p < 0.001$, Cohen's $d = 18.2$). Latency reductions of 30–45% were consistently observed across smaller Whisper models ($p < 0.001$), indicating a robust and statistically significant hardware-related performance improvement.

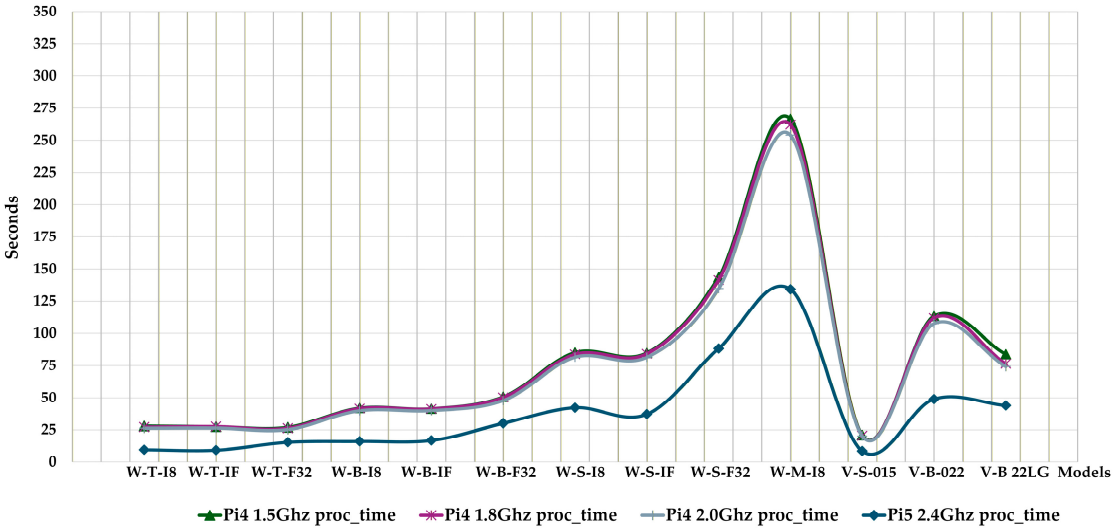


Figure 10. Processing time of various CPU speeds across Whisper and Vosk models on AIoT.

Each model showed a differential increase in CPU usage with increasing complexity. Each Vosk model averaged 88% to 113% CPU utilization, making it the lightest-weight unit among the four under test. The Whisper-tiny and Whisper-base models showed CPU utilization of 260% to 330% across all operational environments. The Whisper-small and Whisper-medium models each showed CPU usage on the Pi 5 of well over 340%, while on the Pi 4 models it exceeded 375%.

The Whisper models use a high percentage of CPU resources. Still, even in these models, memory distribution was sufficient on the Pi 5 so that it did not exceed the system's memory limitations. This shows that it is possible, with low-cost, modern hardware, to provide facilities for very advanced ASR systems, enabling innovations over long periods of sessions. The results of the tables, in combination with statistical comparisons, show the possible outcomes in practical economics for the different products under test. The Vosk models take the least time to process but produce the least accurate results, with a test efficiency of about 50% to 72%, while the WER was 27% to 44%; consequently, they are unreliable for ASR in terms of linguistic accuracy. Whisper-tiny averaged an accuracy of about 72% and delivered results faster than the processing times. With the Pi 5 taking 15–27 s to process, this rapid feedback could occur in practice sessions. Whisper-base achieved an accuracy of about 82% with a WER of about 18%, while still providing processing sessions of a reasonably acceptable time (about 41 s). The Whisper-small and Whisper-medium models showed excellent results, with average accuracies of 88% to 93% and WERs of 6.7% to 12.5%, respectively. Still, for processing, they took the longest of all working periods (in excess of 250 s on the Pi 5).

Balancing recognition accuracy, WER, and system performance, the five most feasible configurations were ranked as follows:

(1) Whisper-small (float32) on Pi 5: Accuracy of 88%, WER of 12.5%, runtime of 253 s; best balance of accuracy and feasibility.
(2) Whisper-medium (int8) on Pi 5: Accuracy of 93%, WER of 6.7%, runtime of 261 s; highest accuracy, but slower than small.

(3) Whisper-base (int8_float32) on Pi 5: Accuracy of 82%, WER of 18%, runtime of 41 s; practical for real-time classroom use.
(4) Whisper-tiny (int8) on Pi 5: Accuracy of 72%, WER of 27%, processing runtime ranging from 15 s to 27 s; ideal for low-latency practice feedback.
(5) Vosk-small (0.15) on Pi 5: Accuracy of 50%, WER of 44%, processing runtime of 74 s; fast and lightweight, but accuracy constraints limit reliability.

These results confirm that, while Vosk is by far the fastest and most efficient, given its limited accuracy, it is not suited for high-stakes assessment. For interactive learning, tiny and base are better trade-offs in practice, while small and medium on Pi 5 offer the highest-quality recognition for advanced applications. The results imply that Pi 5 represents a turning point in system feasibility, as models that were previously too computationally demanding for Pi 4 now run effectively in real time.

4.6. Findings and Discussions

The hybrid tutoring framework, based on IoT, has two functions: to assist learners in acquiring new skills through tutoring and to demonstrate how speech recognition technology works on affordable edge computing devices. There were substantial positive changes in the quality of learners' spoken English. Learners' scores on assessments were significantly higher after using the tutoring system than before. Each cohort showed an increase of approximately 12.4–12.5 points in their post-test mean scores compared to their pre-test mean scores. These are statistically significant improvements ($p < 0.001$) across the three cohorts and represent large effect sizes ($d = 1.51$ in 2022; $d = 1.47$ in 2023; $d = 1.49$ in 2024). This means that both statistical and educational significance are evident in the degree of improvement in learner performance.

A quasi-experimental one-group pre-post design was used across three distinct student cohorts. While the lack of a control group limits the ability to draw causal conclusions about the tutoring condition being examined, the fact that the same pattern of findings holds across all three cohorts suggests that the findings are reliable with respect to learning outcomes. In addition to demonstrating a stable average gain of approximately 12.5 points, the significant and consistent effect sizes also help reduce concerns about attrition bias and sampling error, which are common in many small-scale longitudinal studies. Thus, while the findings may provide strong associative evidence that the tutoring conditions contributed substantially to the observed proficiency improvements among the learners, it is not possible to rule out other variables (e.g., classroom instruction, learner motivation, etc.) that could have influenced the results.

While the numerical results indicate that learners increased their point accumulation, the findings also suggest that they made meaningful advancements along the CEFR proficiency scale. Learners progressed to higher proficiency bands and accumulated points. Results from a chi-square test, $\chi^2(2, N = 36) = 9.42$, $p = 0.009$, suggest that the trends in CEFR levels are unlikely to occur by chance. Additionally, sub-skill analysis supported these findings. Measures of fluency, pronunciation accuracy, and prosodic control also continued to improve throughout the study, reflecting increased articulation precision and a more natural use of speech rhythm. As these indicators reflect similar patterns of growth in linguistic skill development and indicate a learner moving from one stage to another, they strengthen the framework's validity and provide evidence that it facilitates consistent, educationally relevant growth across linguistic skill areas.

Results from the study of the computer science aspects of the system evaluation identified key insights into the system's hardware architecture and model optimization techniques. Comparisons between the Raspberry Pi 4 and Raspberry Pi 5 highlighted differences in processing and memory efficiency. Overclocking the CPU of the Raspberry Pi 4

from its base speed of 1.5 GHz to either 1.8 GHz or 2.0 GHz did not result in a statistically significant decrease in the time to execute ASR on the device ($F(2, 96) = 1.27$, $p = 0.29$). This supports the notion that the Raspberry Pi 4's bottleneck lies not in processor frequency but in the limitations imposed by memory bandwidth and the latency associated with accessing memory caches. Since the CPU pipeline is saturated once the bottleneck occurs, increasing the processor frequency will not translate into improved system throughput due to the constraint imposed by the rate at which data can be transferred relative to the number of raw computation cycles.

On the other hand, the Raspberry Pi 5 features LPDDR4X memory and a more efficient memory controller architecture, enabling CPU resources to be used in parallel more effectively. Under identical conditions, the Pi 5 executed Whisper model inference in an average of 41.8 s, compared with over 330 s on the Pi 4, an 85% reduction in runtime. The performance of all the speech recognition models revealed significant computational trade-offs. The Vosk model, in terms of average processing time (mean of 74.2 s, $SD = 2.7$) and average CPU usage (average range of 90–110%), had the best processing speeds among the models tested but demonstrated the poorest recognition accuracy, ranging from 50 to 72%. Therefore, while they would be helpful for educational purposes as they are the fastest, their limited recognition accuracy would severely limit their educational reliability. Whisper-tiny's recognition accuracy was moderate (72%), with an average WER of 27% and speedy processing times (15–27 s), making it suitable for providing immediate feedback or for practice sessions. Whisper-base achieved a higher recognition accuracy (82%) and processing times closer to 42 s; this model provides a balance between processing time and classroom usability. Whisper-small and Whisper-medium achieved the most accurate results (range of 88–93%), along with the lowest WER values (range of 7–12%); however, these results were obtained at the expense of longer processing times (range of 250+ seconds); therefore, the former would be most useful for formal assessments in situations where high accuracy is more valuable than fast processing times.

These results collectively indicate three main implications. Firstly, the results of the learning tests show both good reliability and practical value, confirmed by consistent effect sizes and CEFR advancements. Secondly, the feasibility of using systems on edge devices is significantly dependent on the relationship between CPU performance and memory architecture. Overclocking CPUs on older devices will not compensate for inadequate memory bandwidth. In contrast, newer architectures (such as the Raspberry Pi 5) can offer measurable improvements in processing performance through higher memory throughput and simultaneous task processing.

Third, pedagogical outcomes and system performance jointly confirm that hardware selection is critical to the success of low-cost, IoT-based tutoring environments. Scalable deployment, therefore, depends not only on algorithmic optimization but also on matching software requirements to the memory and processing characteristics of the chosen hardware platform.

5. Conclusions

This study developed an AIoT-based hybrid tutoring system that integrates open-source ASR engines (Whisper and Vosk) with IoT-based delivery to assess English-speaking performance in real time. The results demonstrate that, despite low cost and limited computational resources, embedded edge devices can deliver reliable, CEFR-aligned feedback and support measurable improvements in learners' fluency, pronunciation, and overall speaking proficiency. Benchmarking experiments further showed that upgrading from the Raspberry Pi 4 to the Raspberry Pi 5 substantially improves inference efficiency and system responsiveness, enabling the deployment of larger ASR models in near-real-time

conditions. These findings confirm the feasibility of a low-cost, portable AIoT platform that combines AI and IoT capabilities to deliver meaningful educational outcomes.

From a theoretical perspective, this research extends prior work on hybrid tutoring by explicitly linking one-to-one pedagogical interaction with AI-driven, real-time assessment delivered through IoT devices. This combination has received limited attention in earlier studies. Whereas previous research has primarily focused on group-based hybrid learning or isolated AI-assisted assessment tools, this study proposes an integrated AIoT architecture that unifies both perspectives. The use of CEFR-aligned descriptors to ground automated performance evaluation, together with an explicit consideration of hardware constraints, provides a dual pedagogical and technological lens that enhances both interpretability and system robustness. This contributes to ongoing discussions on the role of real-time data analytics in supporting student-centered instruction within blended and hybrid learning models.

In practice, the findings indicate that real-time ASR-based tutoring can be effectively implemented on low-cost edge computing platforms, thereby increasing access to personalized language learning in resource-limited educational environments. The study highlights the importance of considering not only processor speed but also memory architecture and bandwidth when selecting IoT hardware for real-time ASR applications. Such systems offer educators and institutions a viable alternative for delivering continuous formative feedback and individualized practice without reliance on cloud-based infrastructure.

This study has several limitations that should be considered when interpreting the educational outcomes. First, the empirical evaluation was conducted with a relatively small sample size of 36 students, distributed across three annual cohorts, and employed a single-group pre-test/post-test design without a parallel control group. While this design is suitable for exploratory validation and longitudinal observation within an authentic tutoring context, it limits the ability to attribute learning gains exclusively to the proposed AIoT-based framework. Second, participants were drawn from a specific private tutoring environment with similar age ranges and instructional conditions, which may constrain generalizability to other educational settings, proficiency levels, or institutional contexts. Accordingly, the reported learning improvements should be interpreted as indicative evidence of feasibility and potential pedagogical value rather than as definitive causal effects. Future studies should incorporate larger, more diverse populations and controlled or quasi-experimental designs to strengthen external validity and causal inference.

Future research should involve larger and more diverse learner populations to improve generalizability and employ quasi-experimental or matched-group designs to compare AIoT-enhanced tutoring with traditional instructional approaches. The framework can also be extended beyond speaking to support multimodal assessment of reading, writing, and listening skills. From an engineering perspective, further optimization through model quantization, pruning, and knowledge distillation is warranted. While the learning outcomes are encouraging, they should be interpreted within the context of the study's sample size and design, with further large-scale and controlled studies required to confirm generalizability. Finally, longitudinal studies are needed to examine the long-term effects of AIoT-based tutoring on learner motivation, engagement, and performance in high-stakes assessments, supporting broader adoption in educational practice.

Author Contributions: Conceptualization, P.N. and M.R.; methodology, P.N. and M.R.; evaluation and modeling, P.N., R.F. and M.R.; validation, P.N., R.F., M.R. and A.R.; formal analysis, P.N., R.F., M.R. and A.R.; investigation, P.N., R.F., M.R. and A.R.; resources, P.N., M.R. and A.R.; data curation, P.N., M.R. and A.R.; writing—original draft preparation, P.N., R.F., M.R. and A.R.; writing—review and editing, P.N., R.F., M.R. and A.R.; visualization, P.N. and M.R.; supervision, P.N., R.F., M.R. and A.R.; project administration, P.N., R.F., M.R. and A.R.; funding acquisition, P.N. and M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were not required for this study because it does not involve hazardous chemicals, equipment, procedures, animal or human testing, or the use of animals or humans as subjects in an experiment.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviations	Model
W-T-I8	Whisper-tiny.en (int8)
W-T-IF	Whisper-tiny.en (int8_float32)
W-T-F32	Whisper-tiny.en (float32)
W-B-I8	Whisper-base.en (int8)
W-B-IF	Whisper-base.en (int8_float32)
W-B-F32	Whisper-base.en (float32)
W-S-I8	Whisper-small.en (int8)
W-S-IF	Whisper-small.en (int8_float32)
W-S-F32	Whisper-small.en (float32)
W-M-I8	Whisper-medium.en (int8)
V-S-015	voskvosk-model-small-en-us-0.15
V-B-022	voskvosk-model-en-us-0.22-lgraph
V-B-022-LG	voskvosk-model-en-us-0.22

References

1. Federičová, M. *The Impact of High-Stakes School-Admission Exams on Study Effort and Achievements: Quasi-Experimental Evidence from Slovakia*; Working Paper Series; CERGE-EI: Prague, Czech Republic, 2014; Volume 9, pp. 515–532. [CrossRef]
2. Adams, D.L.; Hord, C.; Madden, M.; Ryan, A. How does one-on-one tutoring support student self-efficacy? A case study of one high school student's perceptions. *Insights Learn. Disabil.* **2023**, *20*, 37–48.
3. Lee, B.E.; Zlotshewer, B.A.; Mayeda, R.C.; Kaplan, L.I. Impact of online-only instruction on preclinical medical education in the setting of COVID-19: Comparative analysis of online-only vs. hybrid instructions on academic performance and mental well-being. *Med. Sci. Educ.* **2022**, *32*, 1367–1374. [CrossRef]
4. Alhusban, H. A novel synchronous hybrid learning method: Voices from Saudi Arabia. *Electron. J. E-Learn.* **2022**, *20*, 400–418. [CrossRef]
5. Singh, J.; Steele, K.; Singh, L. Combining the best of online and face-to-face learning: Hybrid and blended learning approach for COVID-19, post vaccine, and post-pandemic world. *J. Educ. Technol. Syst.* **2021**, *50*, 140–171. [CrossRef]
6. Gudoniene, D.; Staneviciene, E.; Huet, I.; Dickel, J.; Dieng, D.; Degroote, J.; Rocio, V.; Butkiene, R.; Casanova, D. Hybrid Teaching and Learning in Higher Education: A Systematic Literature Review. *Sustainability* **2025**, *17*, 756. [CrossRef]
7. Imbaquingo, A.; Cárdenas, J. Project-based learning as a methodology to improve reading and comprehension skills in the English language. *Educ. Sci.* **2023**, *13*, 587. [CrossRef]
8. Kusuma, A.A.I.R.S.; Santosa, M.H.; Myartawan, I.P.N.W. Exploring the influence of blended learning method in English recount text writing for senior high school students. *J. Eng. Teach.* **2020**, *6*, 193–203.
9. Sujatha, U.; Rajasekaran, V. Optimising listening skills: Analysing the effectiveness of a blended model with a top-down approach through cognitive load theory. *MethodsX* **2024**, *12*, 102630. [CrossRef]
10. Cheng, J. Blended learning reform in English viewing, listening, and speaking course based on the POA in the post-pandemic era. *Front. Educ.* **2025**, *10*, 1512667. [CrossRef]
11. Deep, P.D.; Chen, Y.; Ghosh, N.; Rahaman, M.S. The influence of student–instructor communication methods on student engagement and motivation in higher education online courses during and after the COVID-19 pandemic. *Educ. Sci.* **2024**, *15*, 33. [CrossRef]

12. Tong, D.H.; Uyen, B.P.; Ngan, L.K. The effectiveness of blended learning on students' academic achievement, self-study skills and learning attitudes: A quasi-experiment study in teaching the conventions for coordinates in the plane. *Heliyon* **2022**, *8*, e12657. [CrossRef]
13. Vieriu, A.M.; Petrea, G. The impact of artificial intelligence on students' academic development. *Educ. Sci.* **2025**, *15*, 343. [CrossRef]
14. Khalifa, M.; Albadawy, M. Using artificial intelligence in academic writing and research: An essential productivity tool. *Comput. Methods Programs Biomed. Update* **2024**, *5*, 100145. [CrossRef]
15. Chen, C.; Gong, Y. The role of AI-assisted learning in academic writing: A mixed-methods study on Chinese as a second language students. *Educ. Sci.* **2025**, *15*, 141. [CrossRef]
16. Kyle, K.; Crossley, S.; Berger, C. The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behav. Res. Methods* **2018**, *50*, 1030–1046. [CrossRef] [PubMed]
17. Liu, X.J.; Wang, J.; Zou, B. Evaluating an AI speaking assessment tool: Score accuracy, perceived validity and oral peer feedback as feedback enhancement. *J. Engl. Acad. Purp.* **2025**, *75*, 101505. [CrossRef]
18. Dja'far, V.H.; Hamidah, F.N. Improving English pronunciation skills through AI-based speech recognition technology. *Ethical Ling. J. Lang. Teach. Lit.* **2024**, *11*, 565–572. [CrossRef]
19. Yıldız, M.; Keskin, H.K.; Oyucu, S.; Hartman, D.K.; Temur, M.; Aydoğmuş, M. Can artificial intelligence identify reading fluency and level? Comparison of human and machine performance. *Read. Writ. Q.* **2025**, *41*, 66–83. [CrossRef]
20. Gligorea, I.; Cioca, M.; Oancea, R.; Gorski, A.-T.; Gorski, H.; Tudorache, P. Adaptive learning using artificial intelligence in e-learning: A literature review. *Educ. Sci.* **2023**, *13*, 1216. [CrossRef]
21. Mouhim, S.; Lachhab, F. Towards a context awareness system using IoT, AI, and big data technologies. *IEEE Access* **2025**, *13*, 40302–40315. [CrossRef]
22. Yadava, T.; Nagaraja, B.G.; Jayanna, H.S. Improvements in ASR system to access the real-time agricultural commodity prices and weather information in Kannada language/dialects. *Multimed. Tools Appl.* **2024**, *83*, 4195–4217.
23. Mulfari, D.; Carnevale, L.; Villari, M. Toward a lightweight ASR solution for atypical speech on the edge. *Future Gener. Comput. Syst.* **2023**, *149*, 455–463. [CrossRef]
24. Fatehi, K.; Torres Torres, M.; Kucukyilmaz, A. An overview of high-resource automatic speech recognition methods and their empirical evaluation in low-resource environments. *Speech Commun.* **2025**, *167*, 103151. [CrossRef]
25. Ahlawat, H.; Aggarwal, N.; Gupta, D. Automatic speech recognition: A survey of deep learning techniques and approaches. *Int. J. Cogn. Comput. Eng.* **2025**, *6*, 201–237. [CrossRef]
26. Rukhiran, M.; Wong-In, S.; Netinant, P. User acceptance factors related to biometric recognition technologies of examination attendance in higher education: TAM model. *Sustainability* **2023**, *15*, 3092. [CrossRef]
27. Dalvi, C.; Rathod, M.; Patil, S.; Gite, S.; Kotecha, K. A survey of AI-based facial emotion recognition: Features, ML & DL techniques, age-wise datasets and future directions. *IEEE Access* **2021**, *9*, 165806–165840.
28. Zhou, M.; Peng, S. The usage of AI in teaching and students' creativity: The mediating role of learning engagement and AI literacy. *Behav. Sci.* **2025**, *15*, 587. [CrossRef]
29. Malik, S. Data-driven decision-making: Leveraging the IoT for real-time sustainability in organizational behavior. *Sustainability* **2024**, *16*, 6302. [CrossRef]
30. Zhai, C.; Wibowo, S.; Li, L.D. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learn. Environ.* **2024**, *11*, 28. [CrossRef]
31. Spaho, E.; Çiço, B.; Shabani, I. IoT integration approaches into personalized online learning: Systematic review. *Computers* **2025**, *14*, 63. [CrossRef]
32. Tabuenca, B.; Uche-Soria, M.; Greller, W.; Hernández-Leo, D.; Balcells-Falgueras, P.; Gloor, P.; Garbajosa, J. Greening smart learning environments with artificial intelligence of things. *Internet Things* **2024**, *25*, 101051. [CrossRef]
33. Alanezi, M.A. An efficient framework for intelligent learning based on artificial intelligence and IoT. *Int. J. Emerg. Technol. Learn.* **2022**, *17*, 112–124. [CrossRef]
34. Tsipianitis, D.; Misirli, A.; Lavidas, K.; Komis, V. IoT devices and their impact on learning: A systematic review of technological and educational affordances. *IoT* **2025**, *6*, 45. [CrossRef]
35. Ghashim, I.A.; Arshad, M. Internet of Things-based teaching and learning: Modern trends and open challenges. *Sustainability* **2023**, *15*, 15656. [CrossRef]
36. Farhan, M.; Jabbar, S.; Aslam, M.; Hammoudeh, M.; Ahmad, M.; Khalid, S.; Khan, M.; Han, K. IoT-Based Students Interaction Framework Using Attention-Scoring Assessment in eLearning. *Future Gener. Comput. Syst.* **2018**, *79*, 909–919. [CrossRef]
37. Ye, W.; Li, M. Application of IoT Android voice assistant based on sensor networks in higher education network mode. *Measur. Sens.* **2024**, *33*, 101091. [CrossRef]
38. Luo, J.; Zheng, C.; Yin, J.; Teo, H.H. Design and assessment of AI-based learning tools in higher education: A systematic review. *Int. J. Educ. Technol. High. Educ.* **2025**, *22*, 42. [CrossRef]

39. Dubey, P.; Dubey, P.; Raja, R.; Kshatri, S.S. Bridging language gaps: The role of NLP and speech recognition in oral English instruction. *MethodsX* **2025**, *14*, 103359. [CrossRef]
40. Council of Europe. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*; Cambridge University Press: Cambridge, UK, 2009.
41. Lu, C.-T.; Lu, Y.-Y.; Lu, Y.-R.; Pan, Y.-C.; Liu, Y.-C. Implementation of an AI English-speaking interactive training system using multi-model neural networks. *IEEE Access* **2025**, *13*, 132052–132066. [CrossRef]
42. Alharthi, H. Investigation into the identification of AI-generated short dialectal Arabic texts. *IEEE Access* **2025**, *13*, 85131–85138. [CrossRef]
43. Halkiopoulos, C.; Gkintoni, E. Leveraging AI in e-learning: Personalized learning and adaptive assessment through cognitive neuropsychology. *Electronics* **2024**, *13*, 3762. [CrossRef]
44. Tang, X.; Chen, H.; Lin, D.; Li, K. Incorporating fine-grained linguistic features and explainable AI into multi-dimensional automated writing assessment. *Appl. Sci.* **2024**, *14*, 4182. [CrossRef]
45. Jing, W. Speech recognition sensors and artificial intelligence automatic evaluation application in English oral correction system. *Measur. Sens.* **2024**, *32*, 101070. [CrossRef]
46. Xie, Y. Application of speech recognition technology based on machine learning for network oral English teaching system. *Int. J. Syst. Assur. Eng. Manag.* **2023**. [CrossRef]
47. Xuto, P.; Prasitwattanaseree, P.; Chaiboonruang, T.; Chaiwuth, S.; Khwanngern, P.; Nuntakwang, C.; Nimarangkul, K.; Suwansin, W.; Khiaokham, L.; Bressington, D. Development and Evaluation of an AI-Assisted Answer Assessment (4A) for Cognitive Assessments in Nursing Education. *Nurs. Rep.* **2025**, *15*, 80. [CrossRef] [PubMed]
48. Brydinskyi, V.; Sabodashko, D.; Khoma, Y.; Podpora, M.; Konovalov, A.; Khoma, V. Enhancing Automatic Speech Recognition with Personalized Models: Improving Accuracy through Individualized Fine-Tuning. *IEEE Access* **2024**, *12*, 116649–116656. [CrossRef]
49. Tolba, R.M.; Elarif, T.; Taha, Z.; Hammady, R. Interactive augmented reality system for learning phonetics using artificial intelligence. *IEEE Access* **2024**, *12*, 78219–78231. [CrossRef]
50. Rakhimova, D.; Duisenbekkyzy, Z.; Adali, E. Investigation of ASR models for low-resource Kazakh child speech: Corpus development, model adaptation, and evaluation. *Appl. Sci.* **2025**, *15*, 8989. [CrossRef]
51. Acosta-Gonzaga, E.; Ruiz-Ledesma, E.F. Students' emotions and engagement in the emerging hybrid learning environment during the COVID-19 pandemic. *Sustainability* **2022**, *14*, 10236. [CrossRef]
52. Bobko, T.; Corsette, M.; Wang, M.; Springer, E. Exploring the possibilities of Edu-metaverse: A new 3-D ecosystem model for innovative learning. *IEEE Trans. Learn. Technol.* **2024**, *17*, 1278–1289. [CrossRef]
53. Vygotsky, L.S. *Mind in Society: Development of Higher Psychological Processes*; Harvard University Press: London, UK, 1980.
54. Gross, G.; Ling, R.; Richardson, B.; Quan, N. In-person or virtual training? Comparing the effectiveness of community-based training. *Am. J. Distance Educ.* **2023**, *37*, 66–77. [CrossRef]
55. Julaihi, A.A. Comparing the impact of asynchronous online quizzes on student learning outcomes in a computer communication and networking course. *J. Cogn. Sci. Hum. Dev.* **2023**, *9*, 125–139. [CrossRef]
56. Malekigorji, M.; Hatahet, T. Classroom response system in a super-blended learning and teaching model: Individual or team-based learning? *Pharmacy* **2020**, *8*, 197. [CrossRef]
57. Yaseen, H.; Mohammad, A.S.; Ashal, N.; Abusaimeh, H.; Ali, A.; Sharabati, A.-A.A. The impact of adaptive learning technologies, personalized feedback, and interactive ai tools on student engagement: The moderating role of digital literacy. *Sustainability* **2025**, *17*, 1133. [CrossRef]
58. Long, M.H. The role of the linguistic environment in second language acquisition. In *Handbook of Second Language Acquisition*; Elsevier: Amsterdam, The Netherlands, 1996; pp. 413–468.
59. Ferguson, C.; van den Broek, E.L.; van Oostendorp, H. AI-induced guidance: Preserving the optimal zone of proximal development. *Comput. Educ. Artif. Intell.* **2022**, *3*, 100089. [CrossRef]
60. Almayez, M.A.; Al-Khresheh, M.H.; Al-Qadri, A.H.; Alkhateeb, I.A.; Alomaim, T.I.M. Motivation and English self-efficacy in online learning applications among Saudi EFL learners: Exploring the mediating role of self-regulated learning strategies. *Acta Psychol.* **2025**, *254*, 104796. [CrossRef]
61. Cao, S.; Zhou, S.; Luo, Y.; Wang, T.; Zhou, T.; Xu, Y. A Review of the ESL/EFL learners' gains from online peer feedback on English writing. *Front. Psychol.* **2022**, *13*, 1035803. [CrossRef]
62. Escobar Fandiño, F.G.; Silva Velandia, A.J. How an online tutor motivates e-learning English. *Heliyon* **2020**, *6*, e04630. [CrossRef]
63. Wood, D.; Bruner, J.S.; Ross, G. The role of tutoring in problem solving. *J. Child Psychol. Psychiatry* **1976**, *17*, 89–100. [CrossRef]
64. Ayman, D.; Mohaseb, M.; El-Bassuony, J. Using instructional scaffolding in hybrid learning environment: A critical review. *Port Said J. Educ. Res.* **2022**, *1*, 132–154. [CrossRef]
65. Teixeira de Melo, A.; Renault, L.; Caves, L.S.D.; Garnett, P.; Lopes, P.D.; Ribeiro, R.; Santos, F. An AI tool for scaffolding complex thinking: Challenges and solutions in developing an LLM prompt protocol suite. *Cogn. Technol. Work* **2025**, *27*, 651–693. [CrossRef]

66. Gao, X.; Noroozi, O.; Gulikers, J.; Biemans, H.J.A.; Banihashem, S.K. A systematic review of the key components of online peer feedback practices in higher education. *Educ. Res. Rev.* **2024**, *42*, 100588. [CrossRef]
67. Ahmed, M.M.H.; McGahan, P.S.; Indurkhya, B.; Kaneko, K.; Nakagawa, M. Effects of synchronized and asynchronized e-feedback interactions on academic writing, achievement motivation and critical thinking. *Knowl. Manag. E-Learn. Int. J.* **2021**, *13*, 290–315.
68. Alfares, N. Is synchronous online learning more beneficial than asynchronous online learning in a Saudi EFL setting? Teachers' perspectives. *Front. Educ.* **2024**, *9*, 1454892. [CrossRef]
69. Enwereji, P.C.; Van Rooyen, A.; Terblanche, A. Exploring students' perceptions on effective online tutoring at a distance education institution. *Electron. J. E-Learn.* **2023**, *21*, 366–381. [CrossRef]
70. Luo, R.-Z.; Zhou, Y.-L. The effectiveness of self-regulated learning strategies in higher education blended learning: A five-year systematic review. *J. Comput. Assist. Learn.* **2024**, *40*, 3005–3029. [CrossRef]
71. Cukurova, M.; Khan-Galaria, M.; Millán, E.; Luckin, R. A learning analytics approach to monitoring the quality of online one-to-one tutoring. *J. Learn. Anal.* **2022**, *9*, 105–120. [CrossRef]
72. Duraku, Z.H.; Hoxha, L.; Buqaj, B. One-on-one tutoring for children with special educational needs: A qualitative multiple-case pilot study in a Kosovo public school. *J. Child. Serv.* **2025**, *20*, 94–110. [CrossRef]
73. Letourneau, N.; Anis, L.; Cui, C.; Graham, I.D.; Ross, K.; Nixon, K.; Reimer, J.; Pilipchuk, M.; Wang, E.; Lalonde, S.; et al. Study protocol for assessing the effectiveness, implementation fidelity and uptake of attachment & child health (ATTACH^{TM}) online: Helping children vulnerable to early adversity. *BMC Pediatr.* **2025**, *25*, 280.
74. Lin, C.-C.; Huang, A.Y.Q.; Lu, O.H.T. Artificial intelligence in intelligent tutoring systems toward sustainable education: A systematic review. *Smart Learn. Environ.* **2023**, *10*, 41. [CrossRef]
75. de Bem Machado, A.; Sousa, M.J.; Sarkar, S.M. State of the art in higher education: Impact of artificial intelligence-based adaptive learning systems in online education. In *Education, Future Jobs and Smart Systems in the Age of Artificial Intelligence, Part A*; Emerald Publishing: Leeds, UK, 2025; pp. 91–109.
76. OpenAI. Whisper: Robust speech Recognition via Large-Scale Weak Supervision. Available online: https://github.com/openai/whisper (accessed on 14 October 2025).
77. Zhang, L.; Wu, S.; Wang, Z. LoRA-INT8 whisper: A low-cost Cantonese speech recognition framework for edge devices. *Sensors* **2025**, *25*, 5404. [CrossRef] [PubMed]
78. AlphaCephei. Vosk Speech Recognition Toolkit. Available online: https://alphacephei.com/vosk (accessed on 14 October 2025).
79. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlícek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In Proceedings of the ASRU Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011; pp. 1–5.
80. Rukhiran, M.; Buaroong, S.; Netinant, P. Software development for educational information services using multilayering semantics adaptation. *Int. J. Serv. Sci. Manag. Eng. Technol.* **2022**, *13*, 1–27. [CrossRef]