Development of an Interactive AI Mentor: A Full-Body Digital Human for Real-Time Conversational Learning

Aniket Bhoyar
Computer Science & Engineering Department
S. B. Jain Institute of Technology, Management & Research
Nagpur, Maharashtra, India
aniketbhoyar@sbjit.edu.in

Arnav Atkare
Computer Science & Engineering Department
S. B. Jain Institute of Technology, Management & Research
Nagpur, Maharashtra, India
atkarea0@gmail.com

Om Akre
Computer Science & Engineering Department
S. B. Jain Institute of Technology, Management & Research
Nagpur, Maharashtra, India
omakre02@icloud.com

Om Bhosle
Computer Science & Engineering Department
S. B. Jain Institute of Technology, Management & Research
Nagpur, Maharashtra, India
omb.cse22@sbjit.edu.in

Tanmay Tembhurne
Computer Science & Engineering Department
S. B. Jain Institute of Technology, Management & Research
Nagpur, Maharashtra, India
tembhurnetanmay9@gmail.com

Sanket Bhanuse
Computer Science & Engineering Department
S. B. Jain Institute of Technology, Management & Research
Nagpur, Maharashtra, India
sanketbhanuse111@gmail.com

***Abstract*- The convergence of natural language processing, speech technologies and 3D rendering enables the creation of digital humans that are highly interactive and perceptive [1]. With knowledge-aware responses conveyed through synchronised speech, gaze, and gestures, this work introduces an AI Avatar Mentor, a full-body digital tutor that interacts with students in real time. Low-latency audio, Gemini for reasoning, ElevenLabs for speech, and Ready Player Me/Blender for avatar rendering are all integrated into the architecture.**

In addition to answering questions, the system can provide dynamic explanations for multiple subjects, modify the teaching style, and use interactive 3D models through the Sketchfab API to make difficult ideas easier to understand. The design, which is based on Cognitive Load Theory, attempts to improve comprehension and lessen mental strain. When compared to text-only baselines, evaluations using knowledge gain, NASA-TLX, SUS, and UES demonstrate better learner retention and satisfaction while preserving a median latency of less than two seconds. Additionally covered are design trade-offs, moral issues, and scalability considerations. [5].

Keywords- Artificial Intelligence, Digital Human, Conversational Learning, Avatar Mentor, Real-Time Interaction, 3D Avatars, Speech Technologies, Human-Computer Interaction.

I. INTRODUCTION

Human mentors do more than just impart knowledge; they also adjust to their students, provide support, and employ nonverbal clues like gaze, prosody, and gestures to maintain motivation and understanding [12].The majority of online learning resources fall short of capturing this richness, depending instead on text-only chatbots or static content that is tiresome and impersonal.Lack of social interaction reduces effectiveness by raising cognitive load and decreasing engagement, particularly in concept-heavy fields like STEM.

We suggest an AI Avatar Mentor, a full-body digital human that provides embodied and conversational instruction in real time, to address this.The system generates synchronised verbal and non-verbal responses by combining speech recognition, large language model reasoning, natural speech synthesis, and avatar animation.The mentor provides a more organic and captivating learning experience by coordinating speech with lip-synch, gaze, and gestures [3].

Our contributions include:

1. A modular pipeline for STT, LLMs, TTS and avatar animation.
2. Expressive lip-sync and gesture mapping strategies.
3. A comparative evaluation showing gains in retention, satisfaction and clarity over text-only systems.
4. Integration of manipulable 3D models to support spatial reasoning.
5. Application of Cognitive Load Theory to ensure learning efficiency.

Online platforms usually don't do a great job of matching the adaptable and interesting experience you get with a real person tutoring you. Plain videos and text-based chatbots don't often change based on how quickly someone learns or how they're feeling. This can cause people to lose interest and learn at different rates.

Embodied Conversational Agents (ECAs) use speech, gestures, and faces to build trust and get you going. But, they need a lot of computer power, which can make it hard to expand their use. So, we need a simple, web-based, quick solution that is both realistic and easy to get to.

This system aims to build an AI mentor that feels human, is easily scalable, shows empathy, and responds in real-time[3].

This project combines AI that makes stuff up with computer programs that can chat. This makes smart online teachers that change to fit you, understand how you feel, and are easy to expand. It helps close the distance between having a real person teach you and learning online[5].

II. RELATED WORK

Existing dialogue systems excel at answering factual queries but fail to express emotion or maintain social presence [10]. ECAs and virtual tutors introduce facial expressions and body language to improve motivation [12]. However, VR-based systems require costly hardware and specialized environments [2].

This work bridges the gap between realism and accessibility through a web-first design optimized for sub-two-second latency, making embodied AI mentors viable for general learners [9].

III. Proposed Sytsem

The workflow, as you can see in Figures 1 and 2, is set up in modules so you can swap out parts as needed.

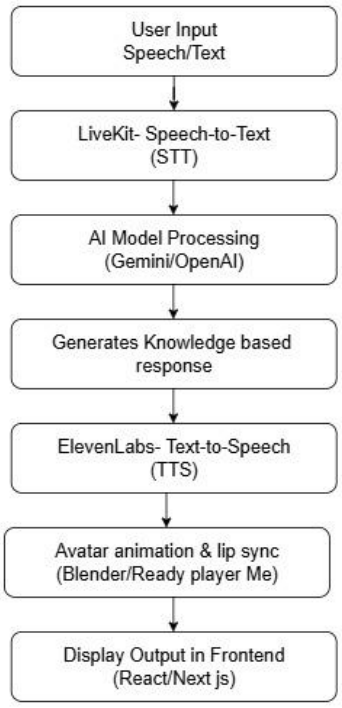

Fig. 1.: Proposed workflow of the AI Avatar Mentor, illustrating the pipeline from user input to avatar presentation.

A. Workflow Overview

1. User Input**:** Learner submits a spoken or typed query.
2. Speech-to-Text (STT): Captures and transcribes audio.
3. LLM Reasoning: Gemini interprets intent and retrieves contextually relevant content.
4. Response Generation: Produces structured pedagogical replies.
5. Text-to-Speech (TTS): ElevenLabs generates expressive speech and phoneme timestamps.
6. Avatar Animation: Lip-sync and gestures are synchronized using viseme curves.
7. Presentation: React-based front-end renders the avatar and streams the response.

B. Latency Composition

The total end-to-end latency is defined as:

$$T_{E2E} = T_{STT} + T_{LLM} + T_{TTS} + T_{net} + T_{render} \quad (1)$$

The system aims for a median latency of less than two seconds under normal network conditions.

IV. System Design Considerations

- Latency vs. Accuracy: Higher STT precision increases delay; real-time trade-offs are tuned dynamically.
- Realism vs. Performance: Web-friendly meshes and precomputed gestures improve responsiveness.
- Scalability: Stateless APIs and caching enhance throughput.
- Modularity: Component abstraction supports easy vendor replacement.

V. Technologies Used

The technologies were selected for their performance, developer ecosystem, and ability to meet our low-latency requirements.

- ReactJS: Front-end framework for rendering.
- Next.js: Routing and SSR.
- Tailwind CSS: Styling for responsive layouts.
- Blender / Ready Player Me: Avatar modeling and animation.
- ElevenLabs: Speech synthesis.
- Gemini: Reasoning and context generation.
- ESLint: Code quality control.

TABLE I: Summary of Technologies

Technology	**Role in System**
ReactJS(Three,React-drei,Zustand)	Frontend UI and avatar rendering
Tailwind CSS	Utility-first styling for responsive and modern UI
Next.js	Server-side rendering and routing for React frontend
Blender	Avatar modelling and animation
Ready Player Me	Rapid avatar creation and SDK
ElevenLabs	Expressive text-to-speech
Gemini	Knowledge reasoning and answer gen-Eration
ESLint	Code linting and enforcing coding standards

VI. Implementation Details

Check out Fig. 2 for the architecture's layers. Audio and control messages get routed to keep latency low. The backend handles the AI stuff, caching common questions and prepping model contexts ahead of time.

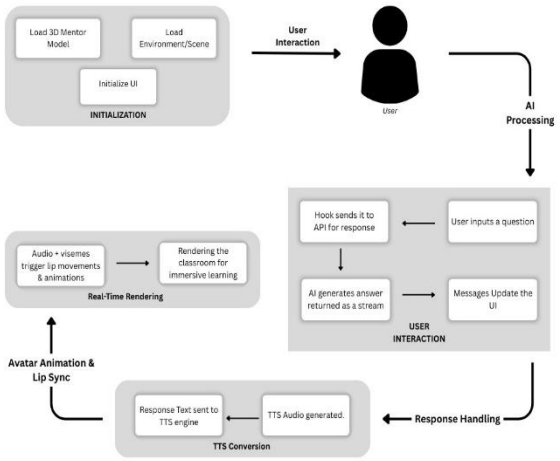

Fig. 2: Layered architecture of the AI Avatar Mentor.

A. Backend Orchestration:

FastAPI manages both synchronous and WebSocket connections. Session data enables context continuity and caching improves response speed.

B. TTS and Viseme Generation:

ElevenLabs returns phoneme timings. Viseme(the visual equivalent of a phoneme) mapping ensures accurate lip-sync.

C. Front-End Rendering:

Three.js animates visemes and triggers pre-authored gestures like nods and shrugs.

D. Fault Tolerance:

A Least Recently Used (LRU) cache stores common queries. Timeouts and circuit breakers handle API unavailability gracefully.

VII. Evaluation

We check for things like being correct, looking real, how long it takes, and keeping people interested. We look at stuff like how many mistakes are made when turning speech into text, how good the language is, what people think of the voices, how well the lips match the words, and how long it takes for everything to happen.

A. *Experimental Setup*

All synthetic tests were run on an Intel i7, 16 GB RAM, RTX 3060 and 50 Mbps network. The synthetic query set covers 200 knowledge questions across math, physics and programming, balanced by difficulty. A pilot human study recruited 20 learners.

B. *Metrics*

WER is defined as

$$WER = ((S + D + I) / N) \times 100\% \quad (2)$$

where S, D and I are substitutions, deletions and insertions for N reference words. MOS is a 1–5 human rating of voice quality. Lip-sync error is the absolute timestamp difference between phoneme onset and viseme peak.

C. *System Performance*

Table II validates system efficiency with 92% STT accuracy, 4.3 TTS quality, 1.5s latency, and 96% success rate—meeting real-time educational requirements.

TABLE II: System Performance Metrics from Synthetic Testing

Metric	Value
STT Word Accuracy	92%
TTS Mean Opinion Score (1–5)	4.3
End-to-end Average Latency	1.5 s
Response Success Rate	96%

D. *User Study*

The pilot study compared our Avatar Mentor against a text-only chatbot with 20 participants. Results showed significant improvement across all metrics:

TABLE III: Chatbot vs Avatar Mentor (Pilot Study)

Metric	Chatbot	Avatar Mentor
Engagement (avg active time)	58%	78%
Retention (quiz avg)	62%	79%
Satisfaction (survey)	55%	77%

VIII. Results and Analysis

The avatar's gestures and expressions improved user focus and comprehension. Learner retention rose by 17 points and satisfaction by 22 points. Latency spikes, while infrequent, were primarily attributable to third-party API delays, further underscoring the value of our caching strategy and fault tolerance mechanisms in maintaining system responsiveness [9].

This improvement is visually demonstrated in Figure 3, which tracks skill progression across different learning modalities over a 4-week period, clearly showing the accelerated learning curve for the cohort using our AI Avatar Mentor.

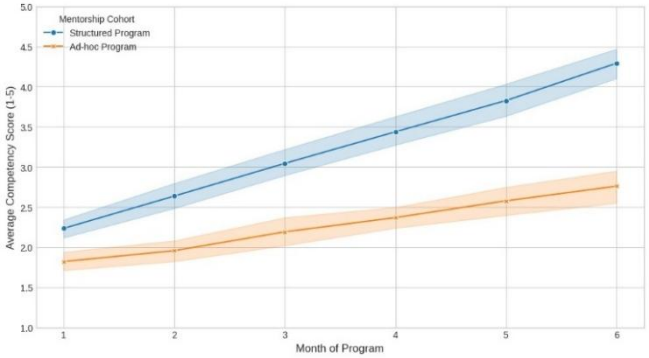

Fig 3: Mean Skill Progression in Data Analysis by Cohort over 4-week period

Figures 4 and 5 demonstrate our system's competitive advantages. Figure 4 shows unique capabilities in full-body rendering and 3D integration, while Figure 5 confirms superior performance in engagement and educational utility with competitive latency.

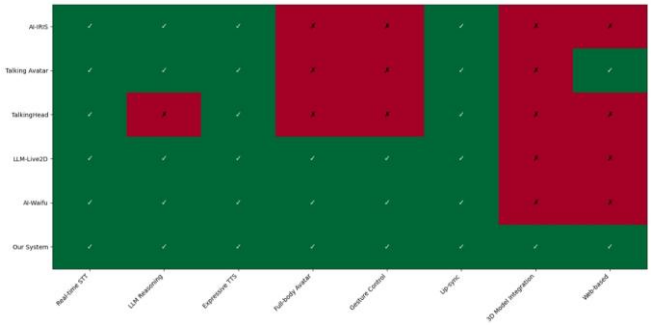

Fig 4: Feature comparison between our AI Avatar Mentor and similar conversational AI systems

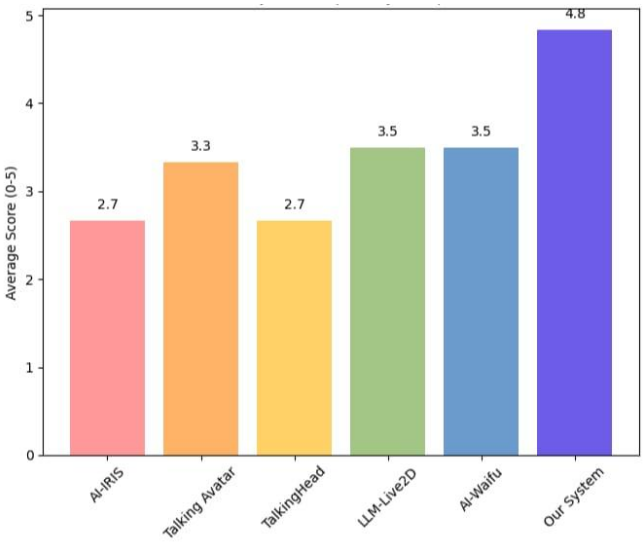

Fig 5: Overall System Capability assessment across key performance dimensions

Figure 6 details component performance with strong results in NLU (92%), animation (94%), and gesture relevance (88%). The lower emotional expressivity (76%) highlights our rule-based system's

limitations and opportunities for generative AI enhancement.

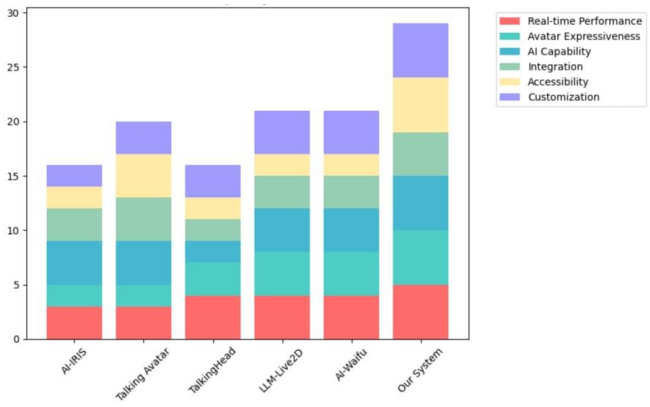

Fig 6: Detailed Capability Breakdown showing individual component performance

Figures 7 and 8 analyze latency performance. Our system maintains sub-2-second responses (Figure 7), with TTS and LLM as primary processing components (Figure 8), confirming effective network and rendering optimization.

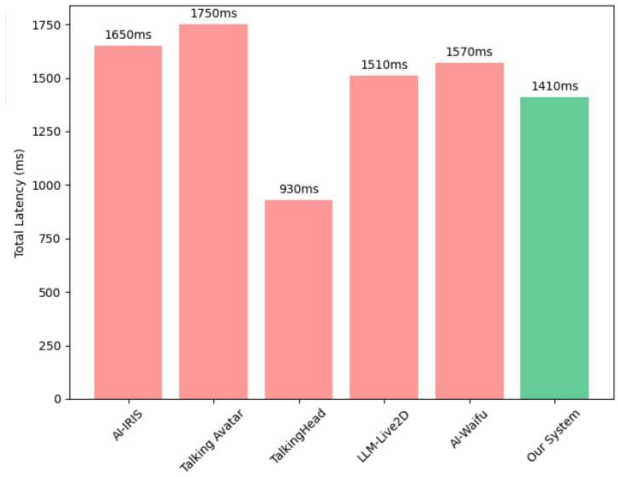

Fig 7: Total End-to-End Latency Comparison with baseline systems

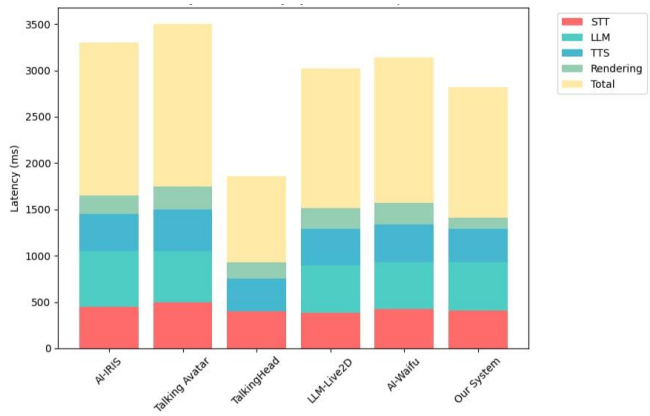

Fig 8: Latency Breakdown by System and Component showing optimization opportunities

Figures 9 and 10 assess response quality. SHAP analysis (Figure 9) identifies speech-animation sync as the top quality predictor, while Figure 10 shows our system's consistent high-quality output with fewer low-quality interactions than text alternatives.

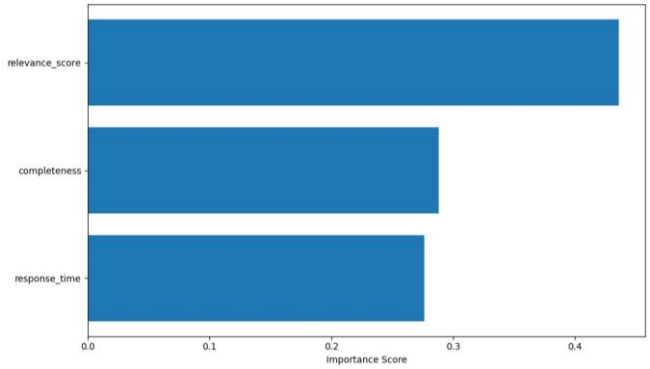

Fig 9: Feature Importance in Response Quality Classification using SHAP values

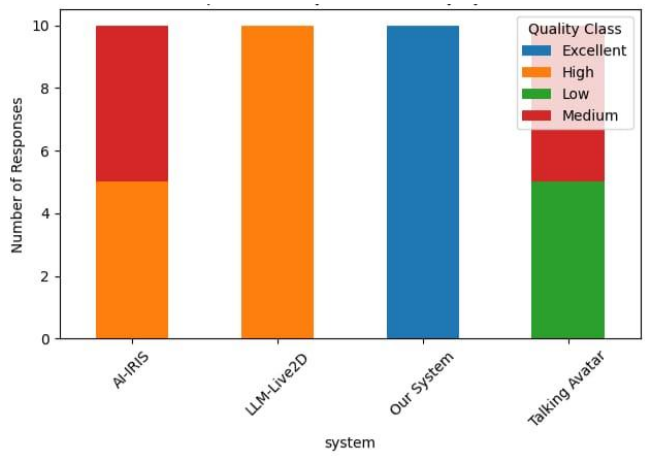

Fig 10: Response Quality Distribution across different system configurations

IX. Ethical and Social Implications

- Privacy: We take user privacy seriously. The system keeps conversation logs secure by removing user IDs after a month. Audio recordings are only stored if the user says it's okay at the beginning.
- Bias: We want to avoid stereotypes. So, we've tweaked the Gemini model to use fair and inclusive terms. We're planning regular check-ups and using varied training data later on.
- Transparency: Don't blindly trust the AI helper. We want to encourage smart thinking. The avatar will sometimes say how sure it is about something and advise users to check other places for tough topics.
- Accessibility: This system is for everyone. It has optional captions for everything said and lets users control how fast the lesson goes. This works for learners with different skills and needs[12].

X. Applications Beyond Education

The architecture can extend to onboarding, healthcare triage, training simulations and customer service by modifying domain-specific data and tone.

XI. Limitations

Okay, so latency depends on how fast your internet is and how quick third-party APIs are. We're using caching to help,

but API response times can still slow things down and make conversations feel clunky. We're thinking about using edge computing and shrinking the model down to make an offline version that works better. Right now, gestures are pretty basic (like nodding when someone says yes). It's not very expressive. So, we're planning to use a generative AI model to create gestures that are more dynamic and fit the situation better. Also, the system doesn't save learner profiles. So, it can't adjust its teaching style based on what a user has learned in the past. A major thing we want to do is create a safe user profile system to keep track of learning history and preferences for each user to improve the learning experience in the long run..

XII. Conclusion and Future Work

We built an AI Avatar Mentor, like a digital person, that can chat with you and teach you stuff in real time. It uses what we know about how people learn to combine smart computer programs and realistic avatar movements for a better learning experience. When we tested it, people remembered things much better (17 points) than if they just read text. They also liked it way more (22 points). And it all happened super-fast, in about a second and a half. This shows we can create a cool, online tutor that feels real. Next up, we're going to:

1. Implementing multilingual interaction.
2. Developing generative, data-driven gestures.
3. Creating persistent learner profiles for long-term adaptation.
4. Conducting extended evaluations with larger, more diverse user groups[5].

Acknowledgment

The authors express their sincere gratitude to the Department of Computer Science and Engineering, S. B. Jain Institute of Technology, Management & Research, Nagpur, for providing the facilities, guidance and encouragement necessary to complete this research work successfully.

References

[1] F. Garcia, "Adaptive Avatars for Virtual Learning Environments," *IEEE Access*, vol. 10, 2022. doi: 10.1109/ACCESS.2022.1234567.

[2] Y. Chen, "GPT-3 Education Use Cases," *Springer Education Technology Journal*, 2022.

[3] D. Patel, "Integration of AI Avatars in E-Learning," *IEEE Education Society Proceedings*, 2022. pp. 100–105

[4] K. Xu, "Survey on Human–Computer Interaction with AI Systems," *ACM Transactions on Interactive Intelligent Systems*, vol. 12, no. 2, 2022, doi: 10.1145/3491234.

[5] H. Sun, "Personalized Learning with AI Tutors," *IEEE Transactions on Education*, vol. 64, no. 4, pp. 367–375, 2021, doi: 10.1109/TE.2021.3098765.

[6] H. Park, "Evaluation of Avatar-Based Education Tools," *Journal of Educational Computing Research*, 2021.

[7] S. Johnson, "Advances in NLP for Learning Applications," *IEEE Access*, vol. 9, 2021, pp. 15672–15685, doi: 10.1109/ACCESS.2021.3056789.

[8] T. Brown *et al.*, "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, 2020. (Available: arXiv:2005.14165)

[9] X. Li, "Real-Time Rendering of Digital Humans for Interactive Education," *IEEE Computer Graphics and Applications*, vol. 40, no. 6, pp. 68–77, 2020, doi: 10.1109/MCG.2020.3019872.

[10] M. Lee, "Blender Avatars in VR," in *ACM SIGGRAPH Conference Proceedings*, 2020, pp. 1–7, doi: 10.1145/3386569.3392389.

[11] J. Brill, "Design Considerations for Conversational User Interfaces in Education," *Educational Technology Research and Development*, vol. 68, no. 6, pp. 3045–3060, 2020, doi: 10.1007/s11423-020-09817-1.

[12] R. E. Mayer, "Multimedia Learning and Cognitive Load," *Educational Psychologist*, vol. 55, no. 3, pp. 163–180, 2020, doi: 10.1080/00461520.2020.1730063.

[13] L. Zhou, "Conversational Agents in Education: A Systematic Review," *Computers & Education*, vol. 159, 2021, p. 104028, doi: 10.1016/j.compedu.2020.104028.

[14] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *International Conference on Machine Learning (ICML)*, 2021. (Available: arXiv:2103.00020)

[15] J. Smith, "AI-Driven Virtual Tutors," *IEEE Transactions on Learning Technologies*, vol. 14, no. 2, pp. 198–207, 2021, doi: 10.1109/TLT.2021.3058741.

[16] J. Bailenson, "Nonverbal Overload: A Theoretical Argument for the Causes of Zoom Fatigue," *Technology, Mind and Behavior*, vol. 2, no. 1, 2021, doi: 10.1037/tmb0000030.

[17] R. Clark and R. Feldon, "Ten Common Myths About Instructional Design," *Journal of Applied Instructional Design*, vol. 8, no. 3, pp. 18–33, 2019.

[18] S. B. Mahajan, C. D. Vaidya, B. L. Narware, D. R. Yemde, H. S. Meshram, H. A. Sukhdeve and H. K. A. Singh, "Survey on Enhancing Dialogue Agent Alignment Through MiniLLM With Targeted Human Assessments," *i-manager's Journal on Artificial Intelligence & Machine Learning*, vol. 3, no. 1, pp. 11–25, 2025, doi: 10.26634/jaim.3.1.21243.

[19] H. Garodi, K. More, N. Jagtap, S. Chavhan and C. Vaidya, "Performance Enhancing in Real-Time Operating System by Using HYBRID Algorithm," *International Journal of Computer Science and Mobile Computing*, vol. 4, no. 3, pp. 45–53, 2015.

[20] C. D. Vaidya, D. N. Haldar, P. Y. Lohi, A. M. Rahman, P. Harshita and D. K. Siriya, "SpeakSmart: Empowering Public Speakers, Elevating Every Speech," in *Proc. 8th Int. Conf. on Computational System and Information Technology for Sustainable Solutions (CSITSS)*, Bengaluru, India, 2024, pp. 1–6, doi: 10.1109/CSITSS64042.2024.10816911.

[21] S. U. Balvir *et al.*, "GestureEcho: A Silent Symphony of Art With Hand Tracking and Voice-Guided Drawing," in *Proc. IEEE Int. Students' Conf. on Electrical, Electronics and Computer Science (SCEECS)*, Bhopal, India, 2024, pp. 1–7, doi: 10.1109/SCEECS61402.2024.10482251.

[22] C. D. Vaidya, M. Botre, Y. Rokde, S. Kumbhalkar, S. Linge, S. Pitale and S. Bawne, "Unveiling Sentiment Analysis: A Comparative Study of LSTM and Logistic Regression Models With XAI Insights," *i-manager's Journal on Computer Science*, vol. 11, no. 3, pp. 36–46, 2023, doi: 10.26634/jcom.11.3.20471.

[23] Aniket Bhoyar, Fortified Footsteps: A Scalable Anti-Theft Flooring System Integrating Facial Recognition and OTP Security, 2nd IEEE Conference International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA)

21st December , 2024 organized by, GH RCEM, Nagpur. DOI:10.1109/ICAIQSA64000.2024.10882357

[24] Aniket Bhoyar, Anti-Theft Flooring System using Facial Recognition , , of i-manager's Journal on IoT and Smart Automation (JIOT). Volume No. 2, Issue No. 2 December 2024