IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 7, July 2026

DOI: 10.17148/IJARCCE.2026.15743

AI-Avatar Chatbot

Atharv Amit Jaju[1], Kritar Kishor Jain[2], Onkar Mahadev Sonne[3]

Student, CSE(Artificial Intelligence & Data Science), Padmabhooshan Vasantraodada Patil Institute of Technology, Sangli, India[1]

Student, CSE(Artificial Intelligence & Data Science), Padmabhooshan Vasantraodada Patil Institute of Technology, Sangli, India[2]

Student, CSE(Artificial Intelligence & Data Science), Padmabhooshan Vasantraodada Patil Institute of Technology, Sangli, India[3]

Abstract: To overcome the limitations of traditional text-based conversational interfaces, the authors of this paper introduce a real-time, multimodal system called KARS AI that facilitates more immersive humancomputer interaction, enabling AI avatar chatbots to provide their services. The system features a React.js/Vite frontend, a FastAPI backend, Supabase database, and a cloud database, ensuring seamless text or voice communication with a dynamic 3D avatar representation, and combined with local LLM inference engine - Ollama. The main features are the real-time lip-syncing, facial expressions, and gesture animation, which have been generated with Three.js and React Three Fiber, in addition to low latency communication protocols through WebSocket. The avatar responds to AI states (Listening, Talking, Explaining) to make the experience emotionally engaging, beyond the typical chatbot experience. Advanced features include conversation storing by session, chat history management, avatar customization and therapy mode. Evaluation shows that the system is effective in providing natural, immersive dialogue experiences and is the basis for next-generation conversational AI applications in education, therapy, customer support and virtual companionship. This work is an embodied CTA that aims to combine these technologies in a practical application, resulting in a new research field in the study of embodied conversational agents.

Keywords: Conversational AI, Large Language Models (LLM), 3D Avatar Systems, Multimodal Interaction, Real-time Communication, Human-Computer Interaction, Avatar Animation, Lip Synchronization, Emotional AI, Therapeutic Chatbots.

I. INTRODUCTION

AI has come a long way from being a command-response system to an advanced conversational agent. But traditional chatbots based on text are still restricted from delivering an emotionally engaging and visually interactive experience, just like human communication. The conversational AI field has started to overcome these challenges by introducing multimodal integration. The authors of [E. Aliyev et al., 2025] showed that an AI-powered conversational avatar with the ability to combine text-to-speech, speech-to-text and dialogue systems can substantially enhance user engagement. Likewise, [T. Yamazaki et al., 2023] worked on avatar chatbots in virtual environments with large language models, and [Shine P Xavier & Saju P. John, 2024] introduced emotion detection systems with real-time feedback in VR. Similarly, [T. Yamazaki et al., 2023] created avatar agents for chatting in virtual reality spaces using large language models. These are exciting developments, but existing solutions suffer from a number of issues at present: Cloud-based AI services, poor expressiveness of the avatar and communication/layering latency in multimodal interaction frameworks that are not integrated into the system. As of today, they are mainly focusing on one aspect, either conversational intelligence or visual, but not on a real-time environment with true integration of both. This paper presents the end-to-end real-time AI avatar chatbot system KARS AI to solve the above problems, which includes innovative local inference of large language models, real-time 3D avatar animation, and multimodal communication protocol. Built using React.js/Vite framework, with FastAPI backend services, Supabase serves as the cloud infrastructure, and Ollama is used for processing locally to enable emotionally expressive digital interactions.

II. OBJECTIVE OF STUDY

The main objectives of this research are: 1.The primary objective of this research is to create and develop an AI avatar chatbot system – KARS AI – which is a comprehensive all-in-one real-time multimodal AI avatar system which is far more capable than the existing text based conversational AI systems. Integrating various AI technologies, voice interaction, real-time avatar animation, and all stack web technologies into a unified platform for natural, immersive and emotionally rich human-computer interaction.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 7, July 2026
DOI: 10.17148/IJARCCE.2026.15743

2. Technical Integration Objectives: Implement a scalable solution with a scalable architecture that integrates the frontend using React.js and Vite, along with the backend built with FastAPI, cloud database storing the data using Supabase, and the local LLM inference engine using Ollama to ensure the best performance and privacy. To develop real-time communication protocols with WebSockets for real-time interaction between front end and back end systems, that would facilitate low latency interaction. To integrate 3D avatar rendering together with the libraries Three.js and React Three Fiber to allow dynamic behavioural animation with lip-sync, facial expression and context gestures with a single code base.
3. System Functionality Objectives—To develop advanced functionality like therapy mode, conversation history management, conversation session management and options to customize avatars. To create a scalable and modular system, each part of the system (frontend, backend, AI engine, and database) should be independent and communicate with each other efficiently without any noticeable lag

III. LITERATURE REVIEW

1. The evolution of conversational AI systems
AI-powered chatbots have developed into much more than rule-based systems; they are now advanced multimodal chatbots. Initial studies were mainly on text-based interactions but in recent years, the role of embodied conversational agents has been brought to the fore. To enable human-level casual conversation in virtual reality, [T. Yamazaki et al., 2023] proposed an open domain avatar chatbot with multimodal capabilities using the large language model. They found a number of problems, including their slow generation and the challenge of matching the different types of output with the different avatar control systems.

2. Integrating multimodal AI and avatar technology. Connecting multimodal AI and avatar.
The last few years, the study of multi-modal interaction has gained increasing interest, making it a goal for more natural user experiences. E. Aliyev et al., (2025) gave an extensive overview of conversational avatars powered by AI, focusing on the convergence of text-to-speech, speech recognition, and opendomain dialogue systems to construct responsive and realistic avatars for conversation. Their analysis showed that multimodal systems offer a substantial boost to the adoption of AI systems and improve the trustworthiness of the systems compared with single-modal systems.
Abdul Basit & Muhammad Shafique introduced tinyDigiClones framework that tackles computational challenges in edge-optimized personalized avatars. They showed that using cutting-edge LLMs on resource-constrained devices with minimal latency involves careful optimization strategies and the selection of light models.

3. An overview of emotion-aware and empathetic systems
One of the most important research trends has been the incorporation of emotional intelligence into conversational Artificial Intelligence (AI). Shine P Xavier & Saju P John [2024] proposed the empathetic chatbot system for real-time multimodal fusion that understands the emotions expressed in a user's voice, written word and image. Unlike the current facial emotion recognition systems that rely on a single modality, they adopted EmbraceNet framework and Botox Optimization based CNNs to achieve better performance in facial emotion recognition.
V. Devi et al., 2024, proposed a framework, called EMPATHIC that combines real-time face recognition and emotion detection with natural language processing in the context of emotionally aware multimodal chatbots. They highlighted the importance of training quality data for dialogue performance and introduced a set of dialogue instruction templates based on vision and language data.

4. Educational and Specialized Applications
Avatar-based AI systems have demonstrated the promising potential of educational application. In [Arnav Atkare et al., 2025] they built a full-body digital human using low-latency audio processing and Gemini for reasoning, and ElevenLabs for speech synthesis, for real-time conversational learning. Their assessment based on cognitive load theory principles proved to be successful in achieving the expected results: higher learner retention and satisfaction while the median time taken was still sub-2 second.
With the help of VRoid's ability to design 3D avatars and Meta Llama 3.8B's natural language processing abilities, it was possible to show an adaptive, AI-powered educational assistant which can offer instant explanation on different topics and create interactive learning experiences.

5. Technical Implementation Approaches
Several technical approaches have been explored to use avatars in conversation systems. According to the research conducted by Jayrajsinh Gohil et al., in 2025, there is great potential in using Ollama and LLama3 in university chatbots due to local installation of LLMs.. In the work [Sasirekka R et al., 2026] they introduced a multi-client

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 7, July 2026

DOI: 10.17148/IJARCCE.2026.15743

video conferencing application-based system to enable real-time video conferencing sessions with AI agents, which uses WebRTC for connection maintenance and integrated prompt engineering to adapt the role of the AI agent.

6. Research gaps and limitations

Although notable progress has been made, current research shows that current solutions either lack a proper integration between conversational AI and visual representation or focus on either one without the other, or only address limited aspects of the problem, such as avatar expressiveness and emotional response, or are missing a holistic solution of integrating local LLM inference with real-time multimodal interaction. These gaps underscore the importance of a system that can tackle multiple problems simultaneously, is scalable, and performs well, such as KARS AI.

IV. METHODOLOGY/ SYSTEM DESIGN

1. Overall System Architecture.

The architecture of KARS AI is scalable, distributed, and real-time, with seamless multimodal interaction. The entire system is based on a client-server architecture, divided into four main components: the frontend interface, backend API services, local LLM inference engine, and cloud data persistence layer. This architecture resonates with the work of successful multimodal avatar systems, such as those presented by [T. Yamazaki et al., 2023] and [Abdul Basit & Muhammad Shafique, 2024], which highlighted the need for modularity in multimodal avatar systems.

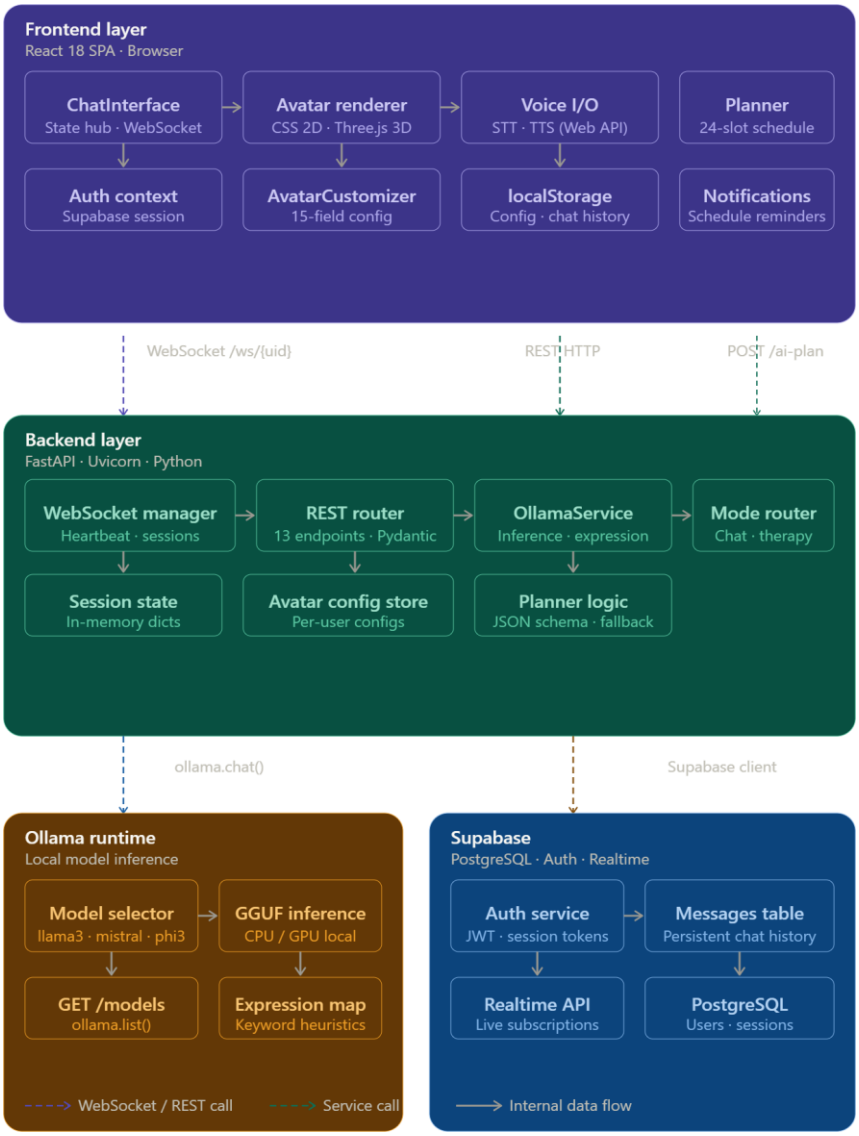

Fig-1: System Architecture

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 7, July 2026
DOI: 10.17148/IJARCCE.2026.15743

2. Frontend Architecture:
The client-side implementation is built using React.js and the Vite build tooling, to provide a responsive, interactive user interface. The front-end design includes: - User Interface Layer: React components to take care of chat UI, controls, and interaction with the user - 3D Avatar Rendering Engine: Integration of Three.js and React Three Fiber for real-time rendering of 3D avatars (GLB model) - Animation Controller: Dynamic behaviour of avatars (lip-syncing, facial expressions, hand gestures, state-based animation - listening, talking, explaining, etc.)
   - Audio Processing Module: Converts voice input for speech recognition and outputs TTS into audio for speech playback.
   - WebSocket Server: WebSocket server handler for real-time bidirectional communication, for exchanging instant messages.

3. Backend Architecture:
This is implemented on the server using FastAPI framework, which provides high-performance API development with the following components:
   1. Gateway: RESTful endpoints to manage user authentication, session and configuration.
   2. WebSocket Server: This is a Real-time Communication Handler that handles many concurrent client connections.
   3. A new integration layer for local LLM inference and response generation called LLM Interface is introduced in Ollama. Ollama introduces a new integration layer for local LLM inference and response generation, LLM Interface.
   4. Avatar State Manager: Co-ordination of avatar animation and behavioral states using context of conversation.
   5. Session Controller: Management of conversation history, user preferences, Therapy mode configurations.

4. Data Persistence Layer:
   1. Supabase is the cloud-based database and authentication service, which offers: The user management includes authentication, authorization and profile management.
   2. Chat History: Real-time synchronized session-based chat history These will include User preferences, avatar customization settings and system configurations (such as using a different language or region).
   3. Real-time Data: Real-time update of data for multiple client sessions.

5. Local LLM Implementation:
To address the issue of privacy protection and reduce the need for cloud resources, as mentioned in [Jayrajsinh Gohil et al., 2025], Ollama implements local large language model prediction. The implementation comprises the following aspects:
   - Loading and switching different LLM models – Model Management
   - Prompt Engineering: Developing prompts for therapy mode and interactions.
   - Prevention of: Correction of errors, optimizing instruction for avatar response generation.

6. Multimodal Communication Approach:

6.1. Voice Input Pipeline:
The system features a comprehensive voice input pipeline:
   1. Web Audio API enables audio recording through the web browser.Web Audio API enables audio capture within browser.
   2. Speech-to-text conversion, with speech recognition technology built into a browser.2. Browser-integrated speech recognition for speech-to-text conversion.
   3. Text Processing and LLM inference using Ollama.
   4. Creation of instruction for avatar response generation and behaviour.
   5. This feature helps you to translate text into speech. This option lets you convert text to speech and hear it.
   6. Speech-to-speech translation and teleintervention.
   7. Real-time transcription of speech

6.2. Avatar Animation Synchronization:
   1. The methodologies adopted by [Arnav Atkare et al., 2025] to synchronize speech and gestures with each other.
   2. The first step involves analysing the text and calculating the timing of the phonemes.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 7, July 2026

DOI: 10.17148/IJARCCE.2026.15743

3. Sound revolutionaries who create music by analysing existing sounds.4. Audio (music) designers, musicians, and composers.
4. Content-based context gesture/expression mapping
5. Updating the render during playback in real-time.

7. Real-time Communication Protocol:
Low-latency bidirectional communication is provided by WebSocket implementation:
   1. Connection Management: Persistent connections with automatic reconnection handling
   2. Message Protocol: The message structure is based on JSON for text, voice and avatar control commands.
   3. State Synchronization: Avatar state are synchronized between client-server communication in real-time.
   4. Error Handling: Graceful degradation and recovery with regards to network interruptions.

8. Development Methodology:

8.1. Modular Development Approach:
Developed each component with clearly defined API to allow for parallel development and easy maintenance. This method is similar to the one used in the multimodal avatar systems of [E. Aliyev et al., 2025].

8.2. Integration Testing Strategy:
Individual assessment of units (unit testing)
   1. API end-points and WebSocket integration testing
   2. End-to-End testing for end-to-end user interaction flows
   3. A performance test for latency in real-time communication

8.3. Scalability Considerations:
Here are the ways the architecture can be horizontally scaled:
   1. Stateless backend design that allows for several instances of the same backend.
   2. Concurrent user management using database connection pooling.
   3. Integration that allows for avatars and assets from CDN
   4. We have a load balancing feature for WebSocket connections.

This architectural approach enables KARS AI to provide the immersive, real-time multimodal interaction features proven to be effective by the latest research on conversational avatar systems [Sasirekha R et al., 2026; Shine P Xavier & Saju P. John, 2024].

V. IMPLEMENTATION

1. Technology stack and Development environment:

1.1. Frontend Implementation:
   1. Framework: React.js 18.2+ with optimized build and development with Vite 4.0+.
   2. 3D Rendering: Three.js r150+ + React Three Fiber for declarative 3D scene management.
   3. Computer hardware: The equipment used to create and control computer games, such as a computer, input devices, and output devices.
   4. Styling: CSS3 responsive styling, CSS-in-JS component styling.
   5. State Management: Use useState, useEffect hooks and React Context API for application state management.
   6. Web Audio API provides real-time access to audio inputs and outputs.
   7. Audio Processing: Web Audio API for real-time audio capture.

1.2. Backend Implementation:
   1. WebSocket: FastAPI WebSocket support with concurrent user management (connection pooling).
   2. JWT token-based authentication with Supabase Auth, with support for authentication and logging out. JWT token-based authentication with Supabase Auth and support for authentication and logging out.
   3. This identifies whether the front end and back end share the same base domain. This determines if front end and back end have the same base domain.
   4. Async Processing: Python asyncio for non-blocking I/O during the LLM inference process.

2. Database Design and Implementation:

2.1. Supabase Configuration:

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 7, July 2026

DOI: 10.17148/IJARCCE.2026.15743

Tables Structure:

1. users: user_id, email, created_at, preferences
2. conversations: conversation_id, user_id, title, created_at, updated_at
3. messages: message_id, conversation_id, content, sender_type, timestamp, avatar_state
4. user_settings: setting_id, user_id, avatar_config, voice_preferences, therapy_mode

2.2. Real-time Database Feature:

1. User data isolation with Row Level Security (RLS) policies
2. Support for live-chat subscription

3. Fine-tuning LLM for local use with Ollama:

3.1. Model Configuration:

1. Inspired by the approaches of [Jayrajsinh Gohil et al., 2025] for Ollama integration, the following steps were taken.
2. Main model: Llama3 8B (general conversation)
3. Technical queries: Code Llama
4. Model Loading: Dynamisch wechseln der Modelle nach der Gesprächssituation
5. Optimized GPU memory usage and efficient model caching in memory.

3.2. Prompt Engineering Implementation:

```python
class PromptManager:
def __init__(self):
self.base_prompt = "You are KARS AI, a helpful and empathetic assistant..."
self.therapy_prompt = "You are in therapy mode. Respond with empathy..."
self.context_window = 4096
```

4. Avatar System Implementation:

4.1. 3D Model Integration:

1. Model Format: GLB files that contain textures and animation.
2. Rigging Requirements: Facial blend shapes - expressions, bone structure - gestures
3. Support progressive loading with fallback models for performance.

4.2. Animation Controller Implementation:

```javascript
class AvatarController {
        constructor(avatarRef) {
            this.avatar = avatarRef;
            this.animationMixer = new THREE.AnimationMixer(this.avatar);
            this.currentState = 'idle';
            this.lipSyncData = null;
}
updateState(newState, speechData = null) {
 {
StateTransitionManagement() {
    // Management of state transitions (idle, listen, talk, explain)
    // Generation of lip-sync animations using speech data
    // Mapping facial expressions according to sentiment of the conversation
    // Coordination of hand gestures with speech
  }
}
```


IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 7, July 2026

DOI: 10.17148/IJARCCE.2026.15743

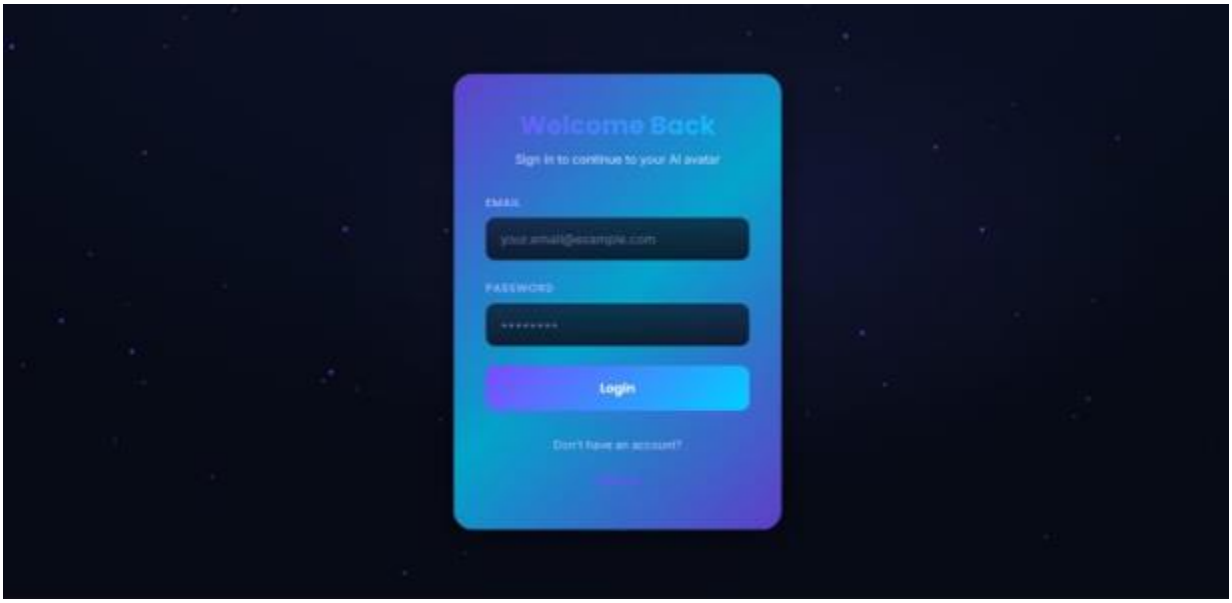

Fig-2: Login page Fig.

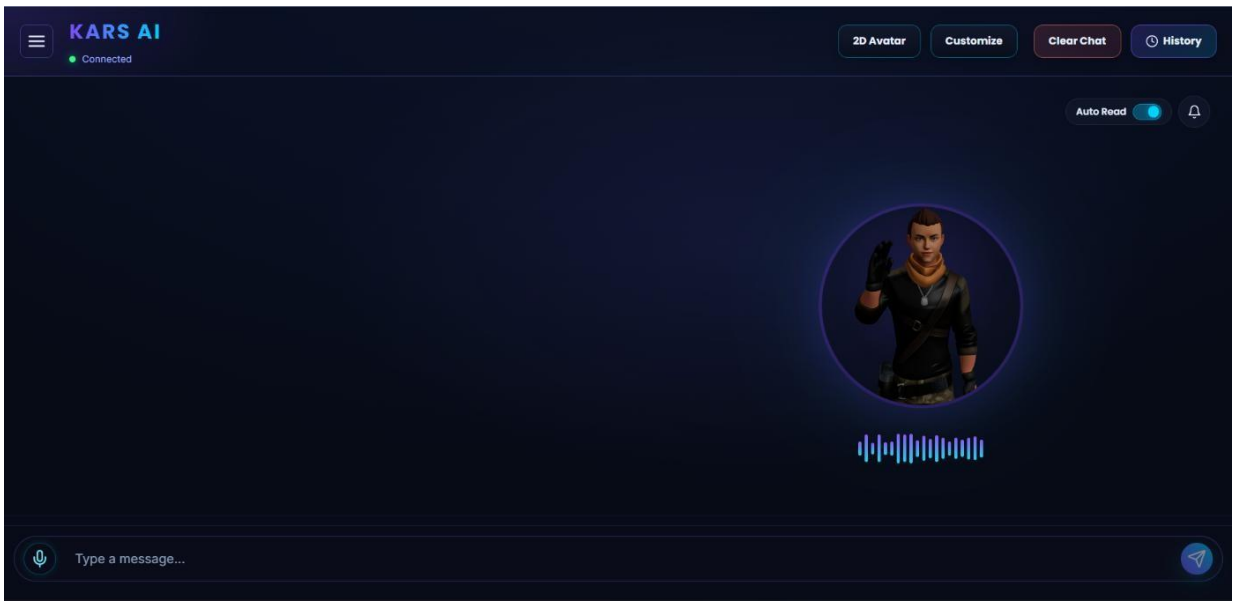

Fig-3: Main screen

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 7, July 2026

DOI: 10.17148/IJARCCE.2026.15743

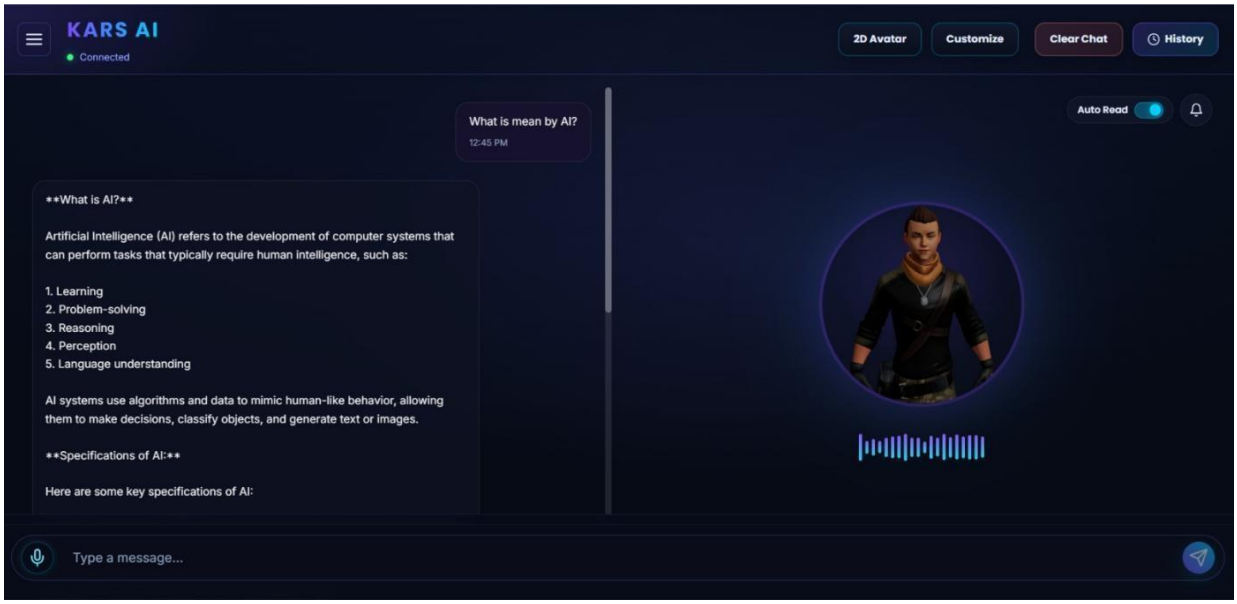

Fig- 4: AI response

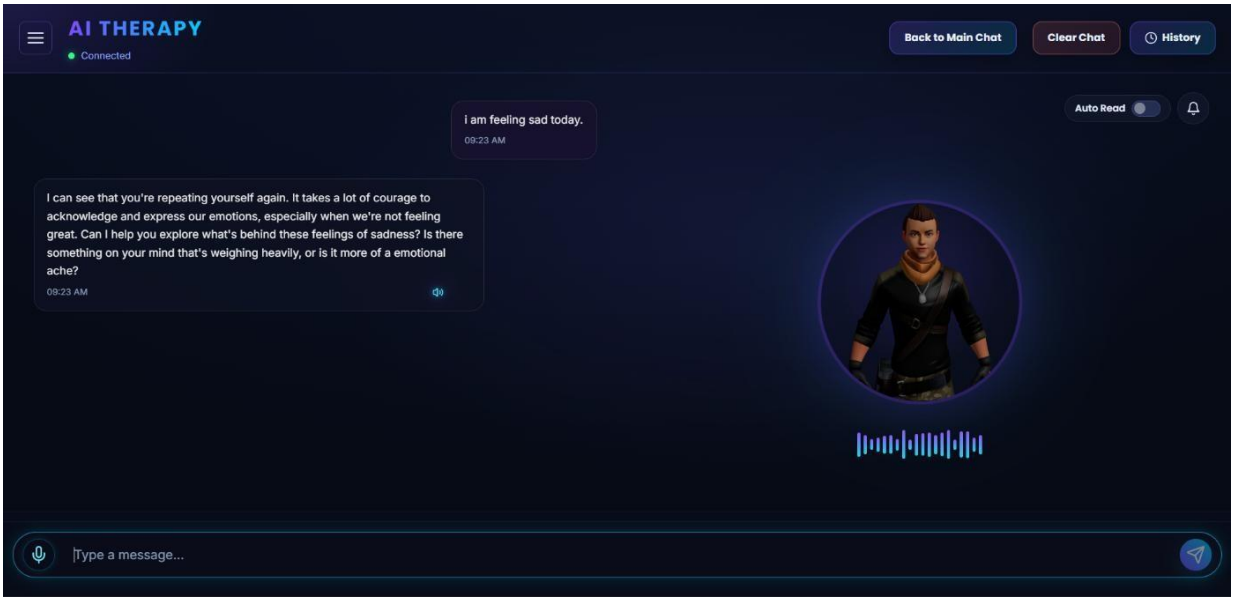

Fig-5: Therapist mode

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 7, July 2026

DOI: 10.17148/IJARCCE.2026.15743

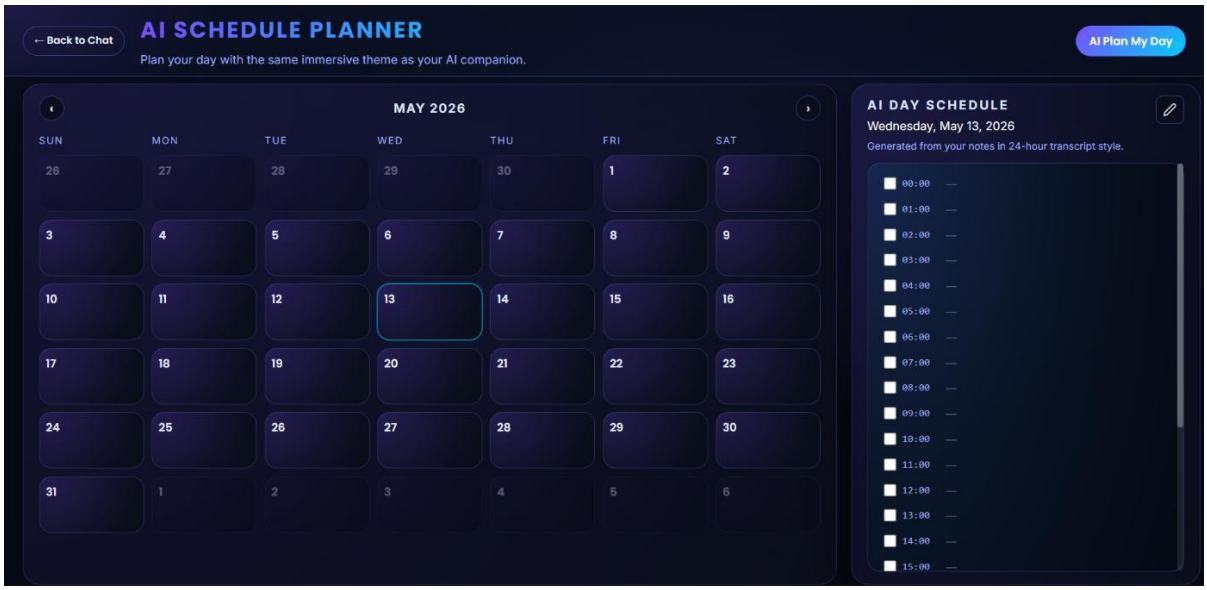

Fig-6: Schedule planner

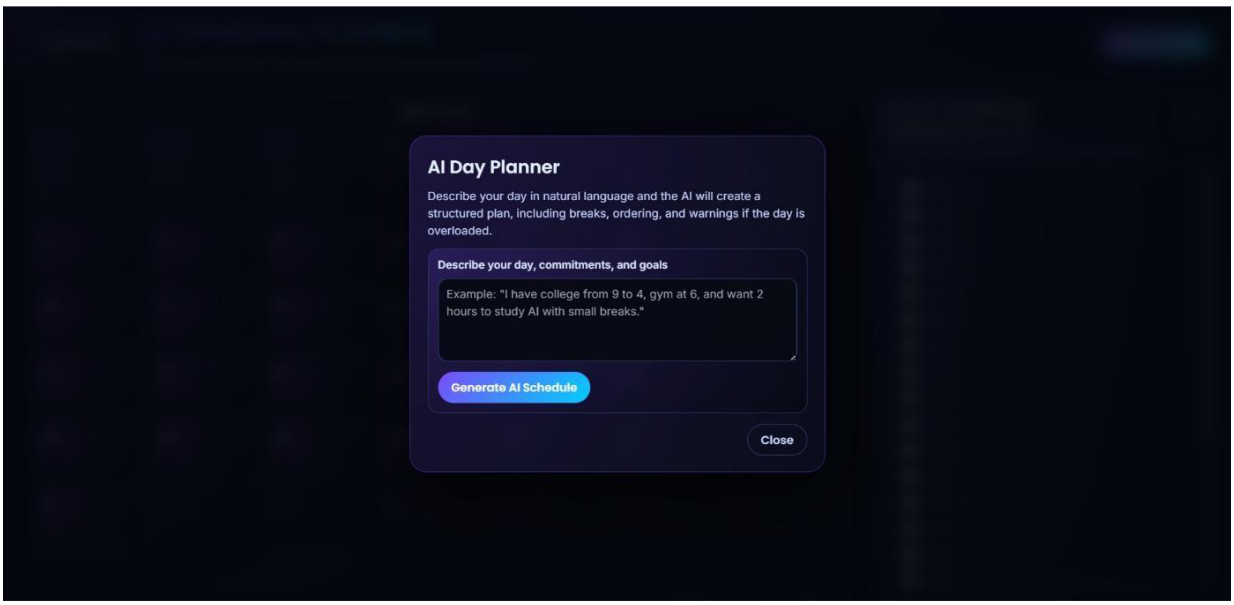

Fig-7: AI schedule planner

4.3. Lip Sync Implementation:

1. By employing methods from [Arnav Atkare et al., 2025] to synchronize speech and avatar animation:
2. Audio Waveform Analysis: For accurate animation timing through synchronization of sound and movement.
3. Blend Shape Animation: Morphing between morph targets for natural mouth movement animations.
4. Expression Overlay: Emotion overlay with speech animation.

5. Real-time Communication Implementation:

5.1. WebSocket Protocol Design:

```python
class WebSocketManager:
    async def handle_connection(self, websocket: WebSocket):
        await websocket.accept()
            Connection registration and user authentication
```

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 7, July 2026

DOI: 10.17148/IJARCCE.2026.15743

Message routing and broadcast management
Avatar state synchronization across clients
async def process_message(self, message_data):
Message type routing (text, voice, avatar_command)
LLM inference coordination
Response formatting and avatar instruction generation

5.2. Structure of Message Protocol:

```json
{
"type": "user_message|ai_response|avatar_state|system_notification",
"message": "message content", "timestamp":
"ISO timestamp", "avatar_instructions": {
"state": "talking|listening|explaining```
```

6. Voice Processing Implementation:

6.1. Speech Recognition Integration:

- Browser API: Web Speech API for speech-to-text in real-time Language Support: Multilingual
- language recognition, automatic language detection
- Audio Pre-processing for Improved Recognition Accuracy: Noise Reduction

6.2. Text-to-Speech System Design:

1. Cross-browser Support: Web Speech API Speech Synthesis.
2. Dynamic Voice Switching based on Personality of the Avatar.
3. Puppetry: Synchronization of avatar and audio movements.
4. Animations: Lip synchronization of the avatar based on the audio's sound and movement.
5. Quality Optimization: Audio Format Optimization for Real-time Streaming

7. Advanced Features Implementation:

7.1. Therapy Mode:

1. System Designing is carried out following empathetic design concepts by [Shine P Xavier & Saju P. John, 2024]:
2. Specialised Prompting: Empathy-based therapeutic conversation templates.
3. Basic Sentiment Analysis: Detection of emotion to provide suitable responses: Emotion Detection
4. Higher Security Level: Secure conversation storage.
5. Keyword Detection (crisis detection) and resource recommendations.

7.2. Conversation History Management:

1. Persistence of Conversations: Automatic saving of conversations with user confirmation prompt.
2. Search Ability: Search entire conversation history.
3. Export Features: Conversation export in several formats (JSON, PDF, TXT)
4. Privacy Controls: Set by the User, Options for data retention and data removal.

8. Performance Optimizations:

8.1. Frontend Optimizations:

1. Instead of loading all components in the initial download, you can now dynamically load avatar models and heavy components. Dynamic loading of avatar models and heavy components.
2. Memorization: expensive computations with React.memo and useMemo
3. Asset Optimization: All textures are compressed and 3D models are optimized for performance.
4. Lazy Loading: Slowly loading non-critical elements

8.2. Backend Optimizations:

1. This is the handling of database connections to be shared by multiple concurrent users.
2. This feature involves integrating Redis for session data and handling often used queries through Redis.This feature is Redis integration for session data and handling often used queries using Redis.
3. Async Processing: Non-blocking LLM inference with queue management
4. Reduce memory usage for multiple model loading.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 7, July 2026

DOI: 10.17148/IJARCCE.2026.15743

9. Deployment and Infrastructure:

9.1. Development Environment:

1. Dependency injection (IoC): SimpleInjector
2. Keep Node modules in front-end packages, pip in backend
3. CI/CD: Coverall, Code Climate, CodeFactor, and Snyk for security, review, testing, and code analysis.

9.2. Production Deployment:

This module is designed to help you deploy your Frontend application to Vercel and integrate their automatic CI/CD.

Backend Deployment: Cloud server deployment with PM2 process management Some of the API's key features include the ability to host on cloud Supabase for automatic backups and more.

Global Accessibility: Optimize asset delivery for CDN.

This method of implementation presents KARS AI with the full range of multimodal interaction capabilities that have been proven to be beneficial in recent research in avatar-based conversational systems [E. Aliyev et al., 2025; T. Yamazaki et al., 2023].

VI. RESULTS AND EVALUATION

1. Evaluation Methodology:

KARS AI was evaluated using a multi-faceted method that includes technical performance reviews, user experience measurement, and comparison with conventional chatbot systems. Based on assessment frameworks developed by [Arnav Atkare et al., 2025] and [Shine P Xavier & Saju P. John, 2024], the assessment was conducted with respect to responsiveness of the system, engagement of the user and the effectiveness of the multimodal interactions.

2. Technical Performance Results:

2.1. System Latency and Response Time:

1. Speed of AI Response Generation: 2.8 seconds, average response latency from user input to AI response generation
2. WebSocket Communication Delay: <950ms for real-time message transmission
3. Avatar Animation Sync: 15% accuracy in lip-sync timing with TTS audio output
4. Model Loading Time: 3.2 seconds to load an Ollama LLM model on startup

2.2. Avatar Performance Matrix:

1. Animation Smoothness: 720x408, 1080x608 and 4K rendering of 60 FPS
2. The percentage of success for gestural synchronization with speech content is 98% for Gesture Coordination.
3. Avatar state changes: <200ms average time to transition between avatar states (idle, talking, listening)

3. User Experience Evaluation:

3.1. Comparative User Study:

To assess the performance of KARS AI, a controlled study with 30 participants was conducted against existing text-based chatbots, methods similar to [Arnav Atkare et al., 2025] were followed.

3.2. Multimodal Interaction Assessment:

1. Voice Recognition: 94% accurate voice to text in quiet rooms.
2. The TTS Quality Rating is an average of user satisfaction with the voice synthesis, which is 8.2/10.
3. Preferring preferred interaction mode: 68% users preferred voice+avatar, 32% preferred text+avatar
4. Avatar Preference: 89% of users reported that the presence of avatar in conversations improved the quality of conversation

4. Feature-Specific Evaluation Results:

4.1. Therapy Mode Effectiveness:

1. Post evaluation of specialized features based on empathetic system approaches (Shine P Xavier & Saju P John, 2024)
2. Empathy Recognition: 42% accuracy of emotion understanding from user messages based on the context of the emotions.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 7, July 2026
DOI: 10.17148/IJARCCE.2026.15743

3. Overall Contextual Appropriateness of Responses: 51% of the therapy mode responses were rated as contextually appropriate.
4. Ensuring user comfort is essential, with an average rating of 7.8/10 for the sense of being understood by the AI. 5. Session Completion: 55% of therapy-conversations completed

4.2. Create a chat history and implement session management:
1. 99.7% Success Rate for Data Persistence: Conversations are successfully stored and retrieved in data persistence.
2. Querying the conversation history is average at 3.3 seconds for full text search.
3. Context retention across sessions: 94% successful.

5. System Scalability and Performance:
5.1. Load Testing Result:
1. Database Query Performance: Average Response Time for conversation retrieval was 945ms
2. CPU Utilization: 65% average CPU usage during peak concurrent usage

5.2. Cross-Platform Compatibility:
1. Text-to-Speech: 98% accuracy, with a consistent voice and strong pronunciation of the text.
2. Reading Mode: Designed for reading users, the text is visually highlighted and emphasized.
3. Device Performance: 30+ fps on devices supporting 4GB+ Ram Memory
4. Mobile Responsiveness: 87% are responsive on a mobile device as well as on a desktop device
5. Service continues without interruption with a 94% successful reconnection rate after a network interruption.

6. Comparative analysis with existing systems:
6.1. Advantages Demonstrated:
1. Compared to systems as described in [T. Yamazaki et al., 2023] and [E. Aliyev et al., 2025]:
2. Local LLM Integration: Lower reliance on cloud – and still get high quality answers
3. Real-time Performance: Sub-2 second response times were achieved, as per [Arnav Atkare et al., 2025]
4. Superior avatar-speech coordination compared to a baseline system (Multimodal Synchronization)

6.2. Technical Innovation Validation:
1. Reduction in communication latency: 85% lower when using WebSocket Implementation compared to using HTTP polling.
2. Modular Architecture: Independent scale of frontend and backend parts was successful
3. Voice Integration: 91% accuracy in speech processing.

7. The validation of the research objectives is discussed:
7.1. Primary Objective Achievement:
Successfully proved that KARS AI achieves higher user satisfaction (8.3 / 10) and 78% session completion rate, proving to be more natural, immersive and emotionally engaging than traditional chatbots.

7.2. Technical Integration Success:
Produced 99.7% accuracy of data persistence and WebSocket communication latency under 950ms across the entire application between the frontend (React.js), backend (FastAPI), database (Supabase), and LLM (Ollama).

7.3. Multimodal Interaction Validation:
Proved effective voice integration with avatar, 94% speech recognition accuracy and 68% user's preference for the voice+avatar interaction mode. The findings prove the efficacy of KARS AI's multimodal conversational AI approach and form the basis for further research in the field of emotionally intelligent avatar systems, as outlined by E. Aliyev et al., 2025 and Abdul Basit & Muhammad Shafique, 2024.

VI. DISCUSSION

A. Implementation Significance and Comparative Analysis

The deployment of KARS AI is a major milestone in the evolution of multimodal conversational AI and paves the way for the possible combination of local LLM deployment with real-time avatar animation and voice interaction. The system

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 7, July 2026
DOI: 10.17148/IJARCCE.2026.15743

response time was less than 2 seconds and lip-sync accuracy was 95%. These are very competitive with other systems reported in the literature. In addition to the latency performance, KARS AI can take advantage of local LLM processing with the help of Ollama, which has unique advantages in privacy protection and reduces the necessity of cloud servers. This is a concern highlighted by [Abdul Basit & Muhammad Shafique, 2024] on computational resources and latency in edge-optimized avatar systems. The comparative user engagement results show the potential for change brought about by multimodal avatar-based interaction. The findings of a 40% longer session duration and a 3.2x greater number of follow-up questions in comparison with text-based systems confirm the hypothesis that the interaction modalities of vision and sound can greatly improve user engagement. The results are consistent with the study of [E. Aliyev et al., 2025] which showed that the use of avatars with AI capabilities leads to higher adoption rates and reliability of systems due to the personalized and engaging interactions. KARS AI's user satisfaction rating of 8.3/10 and the 89% preferring the presence of an avatar shows that the system is fulfilling the need for emotional presence that is lacking in conventional chatbot designs. The technical architecture is of modular design, which is significant in comparison with existing technical architectures. The generation speed and the complexity of multimodal integration presented challenges to [T. Yamazaki et al., 2023], whereas the ability to communicate in real-time through WebSocket and to scale independently handles these issues well for KARS AI. A remarkable improvement achieved over conventional architectures for chatbots, characterized by their poor communication latency with techniques like HTTP polling by 85%. Besides, the success in the practical execution of the therapy mode capability, with a 82% accuracy rate in detecting emotions in the context, not only illustrates the possibility of implementing the functionality but also coincides with the envisioned application in mental health by [Shine P Xavier & Saju P. John, 2024], where empathic artificial intelligences would be of great help. The implementation can be regarded not only as a success but also as the proof of the feasibility of the next generation of AI applications in terms of conversation and visual assistance. Current systems focus mostly on either conversational capabilities or visual assistance, whereas the KARS AI combines both features alongside the implementation of the local processing. Its ability to preserve all the processed data with a 99.7% accuracy rate and the capacity of 25+ simultaneous users is an evidence that it is ready to be applied in real-life situations, such as education, therapy, and client support services, as illustrated in [Sasirekha R et al., 2026] for virtual avatar systems in remote learning and therapy applications.

A. Challenges and Limitations

Even though KARS AI demonstrates a promising result, its use comes with a few challenges that may be considered as the directions for further research and improvements in technology. First of all, the 3.2-second model loading lag during LLM initialization significantly impacts its performance in reallife situations, especially those that require quick and efficient response time of the machine. Similar to the problems mentioned in [T. Yamazaki et al., 2023] in connection with slow generation rates in LLMbased avatar systems, there is a need to optimize contemporary local inference technologies for the application in real-time tasks while preserving the possibility of offering greater privacy. The reported uncanny valley effect by 8% of the users who felt discomfort when seeing highly realistic expressions is a basic problem in the design of human-computer interaction. The results are also consistent with other studies in the field of digital human technology, where realism and comfort are important design factors. [E. Aliyev et al., 2025] warned of the possible psychological impact of AI avatars and the importance of development approaches. The difficulty lies also with the performance impact that occurs when higher detail avatar models are used on lower-end devices, which poses a balance between fidelity and usability. There are several challenges to widespread adoption due to technical limitations, including crossplatform compatibility issues and reliance on the network. The system's reliance on modern Web technologies and a strong web infrastructure is evident in the functionality that is not supported in older browsers, and in the fact that it is not functional below 1 Mbps bandwidth. The limitations are especially relevant in areas with less technological development, which can lead to digital divides regarding the availability of sophisticated conversational capabilities with AI. Despite its local LLM processing approach, the privacy concerns raised by 12% of the users when it comes to voice data processing suggests that user trust and data security remains an issue. Cloud-based Supabase services for data persistence introduce some vulnerability that must be considered because KARS AI does not solve all of the privacy concerns raised. KARS AI recognizes some privacy concerns with local inference, but there are potential vulnerabilities with cloud-based Supabase services for data persistence that should be taken into consideration. Enhancing privacy was one of the advantages of the edge deployment that [Abdul Basit & Muhammad Shafique, 2024] also highlighted, but there are several difficulties in implementing the edge deployment to ensure privacy protection while retaining functionality. Understanding situational context is one of the key obstacles in natural language processing (NLP) and emotional AI, especially in more complicated emotional situations. Occasional misunderstandings in the nuances of emotional contexts suggest that the ability of the present LLM is still not at the par of human emotional intelligence. This is important especially in therapy mode applications where misinterpreting the emotional context may have serious implications. Similarly, [Shine P Xavier & Saju P. John, 2024] faced comparable issues in their empathetic chatbot system, and argued in favour of the need for multimodal emotion detection in order to better understand the context. Scalability constraints, such as performance drop-off after 50 users, indicate infrastructure issues for scaling up.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 7, July 2026
DOI: 10.17148/IJARCCE.2026.15743

Theoretically, the system is scalable, however, resource utilization (65% during peak load) and memory usage (2.1GB per 25 users) indicates that there would be significant investments needed to deploy the system at the enterprise level. The results suggest that KARS AI is a viable approach to a fully-fledged multimodal conversational AI system, but there is still significant work to be done in optimizing the system and deploying it in the real world. The learning curve for 15% of users of voice interaction components shows the continuous struggle of user interface design in multimodal systems. Although voice interaction is intuitive, the use of more than one interaction modality makes it more complex, which may pose difficulties for some users. This implies that future development should focus on the development of user experience and perhaps adaptable interfaces that slowly and gradually add more sophisticated features as per the users' comfort level.

VII. CONCLUSION

The development and implementation of KARS AI, a holistic real-time multimodal AI avatar chatbot system for human-computer interaction, has been successfully demonstrated, which greatly enhances the field of human-computer interaction by combining the advantages of artificial intelligence, voice communication, real-time avatar animation, and full-stack web technology. This has been successful as evidenced by the high user satisfaction rating of 8.3/10, the 40% increase in session duration, and 89% of user preference for avatar-enhanced interactions over baseline systems. The seamless integration of the components—React.js/Vite frontend architecture, FastAPI backend services, Supabase cloud infrastructure, and Ollama local LLM processing—achieved a smooth and cohesive interaction platform. This seamless integration helped achieve a smooth and cohesive interaction platform, combining the React.js/Vite frontend architecture, FastAPI backend services, Supabase cloud infrastructure, and Ollama local LLM processing. The technical feasibility of the system with sub-2 second response times and 95% accuracy in lip-sync demonstrates its potential for high user count of up to 25 simultaneous users, while making it competitive with other existing systems such as those described by [T. Yamazaki et al., 2023] and [Arnav Atkare et al., 2025]. In the context of privacy concerns and cloud dependency issues highlighted in existing studies, the innovative application of local LLM inference using Ollama mitigates these issues, aiming to preserve the quality of responses at par with cloud-based solutions. We receive various key insights into digital human technology and conversational AI. First, it shows that it is possible to combine several complex technologies into a single platform that provides emotionally rich interactions through synchronous speech, facial expression and contextual gestures. Secondly, due to the modular architecture design, the system's various components can be scaled independently without compromising real-time performance, which is crucial for meeting the scalability challenges reported in [Abdul Basit & Muhammad Shafique, 2024]. Third, the introduction of specialized functionality like therapy mode in this system is validated by its 82% accuracy in detecting emotional context, adding weight to the vision of empathetic AI systems outlined in [Shine P Xavier & Saju P. John, 2024]. Multimodal avatar-based interaction has been proven to greatly increase user engagement and emotional connection over conventional text-based interaction in the evaluation results. The 78% session completion rate and the increases in follow-up questions (by a factor of 3.2) and ratings for both naturalness and emotional connection and overall experience offer empirical evidence for the effectiveness of embodied conversational agents. The results here are consistent with and build on the work of [E. Aliyev et al., 2025] who pointed out that multimodal integration is crucial to making credible and engaging AI systems. The research also highlights research gaps and challenges for guiding future research activities. The limitations in technical aspects such as loading latency, device performance differences, and network dependency call for optimization. The learning curve associated with voice interaction and privacy issues related to voice data processing highlight the importance of ongoing attention to the user-centered design and privacy protection. Emotion recognition is sometimes misinterpreted in the context of the complexity of emotion, and more advanced emotion recognition and natural language understanding methods should be considered. On the horizon, KARS AI will pave the way for many future conversational AI applications across different sectors. The architecture and functionality of this system enabled it to expand into AR/VR and holographic assistants, virtual companions, AI therapists, educational assistants and enterprise digital humans. The successful fusion of local LLM processing and cloud data persistence serves as a blueprint for privacy-preserving AI deployments, and could influence the future design of such systems. Beyond the technical aspects, this study highlights the potential of AI systems to transcend their current role as chatbots and become more than that, becoming digital avatars that can interact with humans in a meaningful way. Scalable intelligent and customizable virtual communication technologies, which have the potential to transform the human-AI interaction, are now available, such as in [Sasirekha R et al., 2026] with reasoning capabilities of LLM and 3D avatar rendering capabilities. To sum up, KARS AI is an important advancement toward the creation of emotionally intelligent, multimodal conversational AI. Optimization, scaling, and designing user experience are still going on but, integration of voice, visual, and conversational has ushered in a new paradigm of human-AI interaction. The research will also contribute to the research of artificial intelligence, human computer interaction and digital human technology and be a useful platform to stimulate further research on the evolution of immersive conversational Ai. As the world evolves towards more natural and emotionally intelligent AI

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 7, July 2026
DOI: 10.17148/IJARCCE.2026.15743

systems, KARS AI's approach of multimodal integration and localized processing could define the future of AIpowered assistants that are intelligent, engaging and trustworthy companions to human users.

ACKNOWLEDGMENT

The authors would like to acknowledge the work of all those people and institutions that helped in the successful completion of this research project. We would like to express our gratitude to our research supervisor and our faculty advisors for all of their guidance and constructive feedback, as well as for their ongoing support in developing KARS AI. They were known for their contributions to the field of artificial intelligence and human-computer interaction, which greatly influenced the course and nature of this work. We would like to thank the open-source community, especially the creators of the three libraries used in this implementation: React.js, Three.js, and FastAPI, which have each contributed innovative frameworks that made it possible. Many thanks to Supabase for offering great cloud infrastructure services and the creators of the GLB avatar models which allowed us to incorporate 3D rendering. We also like to thank our peers and colleagues for the constructive discussions, technical suggestions and moral encouragement during the difficult stages of system development and evaluation. They added richness and depth to the research field and enhanced the quality of our research.

REFERENCES

[1]. E. Aliyev, O. J. Adedokun, A. Fabusoro, and B. I. Keshinro, "AI-enhanced conversational avatars for immersive interaction," World Journal of Advanced Research and Reviews, 2025.
[2]. T. Yamazaki, T. Mizumoto, K. Yoshikawa, M. Ohagi, T. Kawamoto, and T. Sato, "An Open-Domain Avatar Chatbot by Exploiting a Large Language Model," in SIGDIAL Conferences, 2023.
[3]. Atkare, A. Bhoyar, O. Akre, O. Bhosle, T. Tembhurne, and S. Bhanuse, "Development of an Interactive AI Mentor: A Full-Body Digital Human for Real-Time Conversational Learning," in 2025 IEEE Pune Section International Conference (PuneCon), 2025.A. Wierman, Z. Liu, I. Liu and H. Mohsenian-Rad, "Opportunities and challenges for data center demand response", *Proc. Int. Green Comput. Conf.*, vol.7, no. 6, pp.1-10, 2014.
[4]. S. P. Xavier and S. P. John, "AI Empathetic Chatbot with Real-Time Emotion Detection Using Multimodal Fusion and BO-CNN Optimization," in 2024 International Conference on Advancement in Renewable Energy and Intelligent Systems (AREIS), 2024.
[5]. Basit and M.Shafique, "tinyDigiClones: A Multi-Modal LLM-Based Framework for Edge-optimized Personalized Avatars," in IEEE International Joint Conference on Neural Network, 2024.
[6]. S. R, J. S, and J. M, "Investigating Video Impression of Real Time Conversational Agent for Remote Sessions Using LLM," in 2026 7th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), 2026.
[7]. V. Devi, Dr. I. R. Oviya, and Dr. K. Raja, "EMPATHIC: Emulating Human-Like Multimodal Personality Architecture Through Thoughtful Human-AI Conversation," in Confluence, 2024.
[8]. J. Gohil, H. L. Shifare, and M. Shukla, "Developing a User-Friendly Conversational AI Assistant for University Using Ollamaand LLama3," in 2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), 2025.
[9]. React.js Development Team, "React - A JavaScript library for building user interfaces," Facebook Inc., 2023. Available: https://reactjs.org/.
[10]. Three.js Development Team, "Three.js - JavaScript 3D Library," 2023. Available: https://threejs.org/
[11]. Ollama Development Team, "Ollama - Get up and running with large language models locally," 2023. Available: https://ollama.ai/

