Implementation Roadmap

Type	Plan
Status	Not Started
Priority	Urgent
Assigned To	Thinh Hung Ho Minh Hoàng
Due Date	@September 5, 2025
Dependencies	None
Description	Strategic plan outlining the different types of customer interactions the system will handle, conversation flows, and escalation procedures.

Digital Human Restaurant Assistant - 4-Week Sprint Plan (2 People × 2h/day)

Team Structure & Specialization

Person A: Senior AI Engineer (Backend/Media Specialist)

- **Primary**: Backend architecture, orchestrator, voice processing, avatar generation
- **Tools**: vLLM, Whisper, Zipvoice TTS, FastAPI
- **Responsibility**: Complex system integration, performance optimization

Person B: Junior AI Engineer (Agent/RAG Specialist)

- **Primary**: LangGraph agents, RAG system, conversation logic
- **Tools**: LangGraph, LangChain, Qdrant, sentence-transformers
- **Responsibility**: Agent workflows, knowledge base, conversation flows

🎯 4-Week Ultra-Focused Sprint Plan (56 Total Hours)

Week 1: Foundation & Core Systems (14h each)

Person A - Infrastructure & Voice Pipeline (14h)

```
Day 1 (2h): Project Setup & vLLM Integration
├── Docker compose with vLLM server
├── FastAPI backend skeleton
├── PostgreSQL + Redis setup
└── Test vLLM API endpoints

Day 2 (2h): Audio Processing Pipeline
├── OpenAI Whisper integration (local whisper-cpp or faster-whisper)
├── Basic VAD using py-webrtcvad or silero-vad
├── Audio streaming FastAPI endpoints
└── Test audio → text pipeline

Day 3 (2h): TTS & Avatar Foundation
├── Coqui TTS setup (YourTTS or XTTS-v2)
├── Basic Three.js scene setup
├── Simple avatar mesh (VRM or GLB model)
└── Audio playback in browser

Day 4 (2h): Backend Services Architecture
├── FastAPI session management
├── WebSocket server for real-time communication
├── Basic message queue (Redis Streams)
└── Database models (SQLAlchemy)
```

```
Day 5 (2h): WebRTC Foundation
├── Simple WebRTC audio capture in frontend
├── MediaRecorder API for audio streaming
├── Socket.io for real-time communication
└── Basic HTML/JS interface

Day 6 (2h): Integration & Testing
├── End-to-end voice pipeline test
├── vLLM → TTS → Avatar basic flow
├── Performance baseline measurement
└── Fix critical integration issues

Day 7 (2h): Orchestrator Foundation
├── Basic orchestrator service structure
├── Session routing logic
├── Event handling framework
└── System health monitoring
```

Person B - Agent & RAG System (14h)

```
Day 1 (2h): LangGraph Setup & Agent Architecture
├── LangGraph environment setup
├── Design agent state schemas
├── Create basic dialogue agent structure
└── Test simple conversation flow

Day 2 (2h): RAG System Foundation
├── ChromaDB or Qdrant setup
├── sentence-transformers embedding model (all-MiniLM-L6-v2)
├── Basic document ingestion pipeline
└── Restaurant knowledge base preparation

Day 3 (2h): Dialogue Agent Core Logic
├── Intent classification logic
├── Basic conversation context management
├── Simple response generation
└── Integration with vLLM backend

Day 4 (2h): Reservation Agent Development
├── Table management logic
├── Availability checking algorithms
├── Booking confirmation workflows
└── Basic validation and error handling

Day 5 (2h): RAG Integration
├── Knowledge retrieval implementation
├── Context injection into prompts
├── Relevance scoring and filtering
└── Test FAQ and menu queries

Day 6 (2h): Agent Tools Development
├── LangChain tool implementations
├── Database query tools
├── Notification tools (email/console)
└── Tool calling integration

Day 7 (2h): LangGraph Workflow Integration
├── Connect dialogue and reservation agents
├── State management between agents
├── Error handling and retry logic
└── End-to-end agent testing
```

Week 1 Target: Working voice-to-agent-to-voice pipeline with basic reservation capability

Week 2: Integration & Avatar Enhancement (14h each)

Person A - Advanced Avatar & Voice (14h)

```
Day 1 (2h): Avatar Animation System
├── Facial animation rigging
├── Viseme mapping for lip-sync
├── Basic gesture animations
└── Animation state machine

Day 2 (2h): Lip Sync Implementation
├── Phoneme extraction from TTS
├── Real-time mouth movement
├── Smooth animation blending
└── Timing synchronization

Day 3 (2h): Voice Quality Enhancement
├── Noise reduction post-processing
├── Audio normalization
├── Multiple TTS voice options
└── Voice emotion parameters

Day 4 (2h): Real-time Streaming Optimization
├── Audio chunk processing
├── Streaming TTS implementation
├── Buffer management
└── Latency optimization

Day 5 (2h): Avatar Personality System
├── Different animation styles
├── Emotion-based expressions
├── Gesture variety implementation
└── Context-aware animations

Day 6 (2h): Performance Optimization
├── Model quantization (ONNX/TensorRT)
├── Caching strategies
├── Memory management
└── GPU utilization optimization
```

```
Day 7 (2h): Frontend Polish
├── Responsive UI design
├── Loading states and error handling
├── Audio controls and settings
└── Visual feedback improvements
```

Person B - Advanced Agents & Conversation (14h)

```
Day 1 (2h): Advanced Dialogue Patterns
├── Multi-turn conversation handling
├── Context switching logic
├── Conversation memory enhancement
└── Interruption handling

Day 2 (2h): Reservation Logic Enhancement
├── Complex booking scenarios
├── Alternative suggestion engine
├── Waitlist management
└── Booking modification handling

Day 3 (2h): RAG System Enhancement
├── Advanced retrieval strategies
├── Multi-document reasoning
├── Contextual re-ranking
└── Answer synthesis improvement

Day 4 (2h): Agent Coordination
├── Inter-agent communication
├── Shared state management
├── Conflict resolution
└── Priority handling

Day 5 (2h): Vietnamese Language Optimization
├── Vietnamese-specific prompts
├── Cultural context integration
├── Local restaurant terminology
└── Code-switching handling

Day 6 (2h): Error Handling & Recovery
├── Graceful failure modes
├── Automatic retry mechanisms
├── Fallback conversation strategies
└── User experience during errors

Day 7 (2h): Conversation Analytics
├── Conversation logging
├── Success metrics tracking
├── Performance monitoring
└── A/B testing preparation
```

Week 2 Target: Natural conversation with animated avatar, sophisticated reservation handling

---

Week 3: Polish & Restaurant Features (14h each)

Person A - Production Features (14h)

```
Day 1 (2h): Multi-Language Support
├── Language detection
├── TTS voice switching
├── Audio quality per language
└── Mixed language handling

Day 2 (2h): Advanced Audio Processing
├── Speaker separation (basic)
├── Background noise filtering
├── Echo cancellation
└── Audio quality monitoring
```

```
Day 3 (2h): Avatar Customization
├── Multiple avatar models
├── Clothing/appearance options
├── Restaurant branding integration
└── Dynamic avatar switching

Day 4 (2h): System Monitoring
├── Performance metrics collection
├── Real-time system health
├── Error tracking and alerts
└── Resource usage monitoring

Day 5 (2h): Deployment Preparation
├── Docker containerization
├── Environment configuration
├── Dependency management
└── Startup scripts

Day 6 (2h): API Documentation
├── OpenAPI specifications
├── Usage examples
├── Integration guides
└── Troubleshooting docs

Day 7 (2h): Final Integration Testing
├── End-to-end testing
├── Load testing basics
├── Bug fixes and optimization
└── Demo preparation
```

Person B - Business Logic & Knowledge (14h)

```
Day 1 (2h): Restaurant Domain Knowledge
├── Comprehensive menu RAG data
├── Policy and procedure docs
├── Common customer scenarios
└── Vietnamese restaurant customs

Day 2 (2h): Advanced Conversation Scenarios
├── Complaint handling workflows
├── Special dietary requirements
├── Group booking logic
└── Upselling strategies

Day 3 (2h): Reservation Intelligence
├── Optimal table assignment
├── Dynamic pricing awareness
├── Seasonal/event considerations
└── Customer history integration

Day 4 (2h): Staff Integration Features
├── Staff notification systems
├── Kitchen integration prep
├── Manager override capabilities
└── Reporting and analytics

Day 5 (2h): Customer Experience Enhancement
├── Personalization features
├── Repeat customer recognition
├── Preference learning
└── Loyalty program basics

Day 6 (2h): Quality Assurance
├── Conversation quality metrics
├── Response appropriateness checking
├── Cultural sensitivity validation
└── Edge case handling

Day 7 (2h): Documentation & Training
├── Agent behavior documentation
├── Conversation flow diagrams
├── Training data preparation
└── Performance tuning guides
```

Week 3 Target: Restaurant-ready system with professional features

Week 4: Integration & Deployment (14h each)

Person A - Final System Integration (14h)

```
Day 1 (2h): System Architecture Finalization
├── Component integration testing
├── Performance bottleneck identification
├── Memory leak detection
└── Resource optimization

Day 2 (2h): Production Deployment Setup
├── Docker compose production config
├── Environment variable management
├── SSL/TLS configuration
└── Reverse proxy setup (Nginx)

Day 3 (2h): Monitoring & Logging
├── Centralized logging setup
├── Metrics collection (Prometheus)
├── Basic dashboard (Grafana)
└── Alert configuration

Day 4 (2h): Security & Hardening
├── API rate limiting
├── Input validation
├── CORS configuration
└── Security headers

Day 5 (2h): Performance Optimization
├── Model loading optimization
├── Caching strategy implementation
├── Database query optimization
└── Memory usage optimization

Day 6 (2h): Demo Preparation
├── Demo scenario preparation
├── Demo data setup
├── Presentation materials
└── Issue troubleshooting

Day 7 (2h): Final Polish & Handover
├── Code cleanup and documentation
├── Deployment instructions
├── Troubleshooting guide
└── Future development roadmap
```

Person B - Agent Optimization & Testing (14h)

```
Day 1 (2h): Agent Performance Tuning
├── Response time optimization
├── Context window management
├── Memory efficiency improvement
└── Conversation quality enhancement

Day 2 (2h): Comprehensive Testing
├── Edge case conversation testing
├── Error scenario validation
├── Multi-language conversation testing
└── Long conversation handling

Day 3 (2h): Knowledge Base Optimization
├── RAG retrieval accuracy testing
├── Knowledge coverage analysis
├── Response relevance improvement
└── Embedding model fine-tuning
```

```
Day 4 (2h): Conversation Flow Refinement
├── Natural conversation patterns
├── Interruption handling improvement
├── Context switching smoothness
└── User experience optimization

Day 5 (2h): Analytics & Insights
├── Conversation success metrics
├── User satisfaction indicators
├── System performance analytics
└── Business intelligence preparation

Day 6 (2h): Documentation & Knowledge Transfer
├── Agent architecture documentation
├── Conversation design patterns
├── RAG system documentation
└── Maintenance procedures

Day 7 (2h): Demo Support & Final Testing
├── Demo scenario support
├── Real-time debugging capability
├── Performance monitoring during demo
└── Post-demo improvement planning
```

Week 4 Target: Production-ready deployment with comprehensive documentation

---

🛠️ Technology Stack (Open Source Focus)

Core Infrastructure

```
LLM Server: vLLM + Llama-3.1-8B-Instruct or Qwen2.5-7B-Instruct
Speech-to-Text: faster-whisper or whisper-cpp
Text-to-Speech: Coqui TTS (XTTS-v2)
Voice Activity Detection: silero-vad
```

```
Avatar: Three.js + VRM/GLB models
Database: PostgreSQL + Redis
Message Queue: Redis Streams
Embeddings: sentence-transformers
Vector DB: ChromaDB or Qdrant
Agent Framework: LangGraph + LangChain
Backend: FastAPI + WebSocket
Frontend: HTML/JS/Three.js (no React for speed)
```

Vietnamese Language Support

```
LLM: Qwen2.5-7B-Instruct (excellent Vietnamese support)
TTS: Coqui XTTS-v2 (multi-language including Vietnamese)
STT: faster-whisper (Vietnamese model available)
Embeddings: multilingual-e5-large (Vietnamese support)
```

Hardware Requirements

```
Minimum: 16GB RAM, RTX 3060 12GB or RTX 4060 16GB
Recommended: 32GB RAM, RTX 4080 16GB or RTX 4090 24GB
CPU: 8+ cores for audio processing
Storage: 100GB+ SSD for models
```

🎯 Success Criteria (4 Weeks)

Technical Achievements

- ✅ <2 second response time (voice → voice)
- ✅ Natural lip-sync and facial expressions
- ✅ 90%+ Vietnamese speech recognition accuracy
- ✅ Successful table reservations end-to-end
- ✅ Multi-turn conversation memory

- ✅ Restaurant knowledge base with 100+ Q&As

Business Value

- ✅ Can handle 5+ common restaurant scenarios
- ✅ Professional appearance suitable for customer-facing use
- ✅ Staff can manage tables and view conversations
- ✅ Scalable architecture for future enhancements

This aggressive 4-week plan maximizes the use of open-source tools and focuses on core functionality that provides immediate business value for restaurants. The specialization allows each person to become expert in their domain while building towards a cohesive system.