Wiserep AI Platform Technical Specifications

Enterprise Conversational AI Platform
Version 3.0 | September 2025

Table of Contents

1. System Overview

1.1 Platform Description

Wiserep AI is an enterprise-grade conversational AI platform designed for high-volume call center automation and customer engagement. The platform delivers human-like AI assistants capable of handling complex multilingual conversations across voice and text channels with enterprise-scale reliability and security[2][4].

1.2 Core Capabilities

- **Human-like Conversational AI** with 99.7% accuracy in intent recognition
- **Multilingual Support** for 40+ languages including English, French, and Arabic
- **Enterprise Scale** handling 15,000+ concurrent calls with sub-100ms latency
- **Advanced Voice Processing** with neural TTS and ASR technologies
- **Comprehensive Integrations** with 200+ enterprise systems and telephony platforms

1.3 Key Performance Metrics

- **Accuracy**: 99.7% intent recognition across supported languages
- **Scalability**: 15,000+ concurrent calls, 50,000+ monthly interactions
- **Latency**: <100ms response time for real-time conversations
- **Uptime**: 99.99% SLA with enterprise-grade reliability
- **Languages**: 40+ supported languages with cultural context awareness

2. Architecture & Infrastructure

2.1 System Architecture Foundation

Wiserep AI follows a cloud-native, microservices architecture designed for enterprise scalability and reliability[2].

2.1.1 Core Technology Stack

- **Frontend**: React.js 18+ with TypeScript, responsive design
- **Backend**: Node.js 18+ and Python 3.10+ microservices
- **AI Engine**: Large Language Models (13B+ parameters) with GPU acceleration
- **Voice Processing**: Neural TTS/ASR with real-time streaming
- **Database**: MongoDB clusters with Redis caching, PostgreSQL for transactions
- **Infrastructure**: Kubernetes orchestration with Docker containerization
- **Message Queue**: Apache Kafka for event streaming, RabbitMQ for task queuing

2.1.2 Architectural Patterns

- **Event-Driven Architecture** with CQRS implementation
- **Circuit Breaker Pattern** for fault tolerance and resilience
- **Distributed Caching** with Redis Cluster for sub-millisecond responses
- **Load Balancing** with intelligent routing and health monitoring
- **Auto-Scaling** based on demand patterns and resource utilization

2.2 Infrastructure Components

2.2.1 Compute Layer

- **AI Accelerator Cards**: NVIDIA A100/H100 GPUs for LLM inference
- **CPU Clusters**: Intel Xeon or AMD EPYC processors optimized for AI workloads
- **Memory Architecture**: High-bandwidth memory (HBM) for GPU acceleration
- **Container Orchestration**: Kubernetes with Helm chart deployments

2.2.2 Storage Architecture

- **High-Speed Storage**: NVMe SSD arrays with RAID configurations
- **Tiered Storage**: Hot data on NVMe, warm data on SSD, cold data archived
- **Data Replication**: Synchronous replication across availability zones
- **Backup Strategy**: Automated backups with point-in-time recovery

2.2.3 Network Infrastructure

- **Low-Latency Networking**: 10Gbps+ Ethernet with RDMA capabilities
- **Software-Defined Networking**: Virtual networks with micro-segmentation
- **Load Balancers**: Layer 4/7 load balancing with SSL termination
- **CDN Integration**: Global content delivery for optimal performance

3. AI Engine Specifications

3.1 Large Language Model Architecture

The Wiserep AI platform integrates advanced LLMs optimized for conversational AI applications[2][3].

3.1.1 Model Specifications

- **Parameter Scale**: 13B to 70B+ parameters depending on deployment requirements
- **Architecture**: Transformer-based models with attention mechanisms
- **Training Data**: Domain-specific datasets with multilingual coverage
- **Fine-tuning**: LoRA and QLoRA techniques for efficient customization
- **Context Window**: Up to 32K tokens for extended conversations

3.1.2 Natural Language Processing Engine

- **Intent Recognition**: 99.7% accuracy with multi-intent support
- **Named Entity Recognition**: Custom domain entities with 98.5% precision
- **Context Management**: Multi-turn conversation memory up to 50+ exchanges
- **Sentiment Analysis**: Real-time emotional intelligence with cultural awareness
- **Language Detection**: Automatic detection and switching between 40+ languages

3.1.3 Conversation Management

- **Dialog State Tracking**: Multi-domain state management
- **Response Generation**: Context-aware, compliant responses
- **Escalation Logic**: Intelligent routing to human agents
- **Personalization**: User behavior-based response adaptation

3.2 Model Training & Optimization

3.2.1 Training Infrastructure[2]

- **GPU Requirements**: NVIDIA A100 (80GB HBM2e) or equivalent
- **Training Methods**: Supervised, unsupervised, and reinforcement learning
- **Distributed Training**: Multi-node, multi-GPU configurations
- **Model Compression**: Quantization and pruning for deployment optimization

3.2.2 Performance Optimization

- **Inference Acceleration**: TensorRT and ONNX Runtime optimization
- **Model Caching**: Intelligent caching strategies for frequent queries
- **Batch Processing**: Dynamic batching for throughput optimization
- **Edge Deployment**: Model quantization for edge computing scenarios

4. Voice & Audio Processing

4.1 Speech Recognition (ASR)

Advanced automatic speech recognition with enterprise-grade accuracy[4].

4.1.1 ASR Engine Specifications

- **Accuracy**: 95%+ word error rate in noisy environments
- **Languages**: 40+ languages with dialect support
- **Real-time Processing**: <150ms transcription latency
- **Noise Suppression**: Deep learning-based noise cancellation
- **Speaker Diarization**: Multi-speaker identification and separation

4.1.2 Audio Processing Features

- **Audio Codecs**: Support for G.711, G.722, Opus, AAC formats
- **Sample Rates**: 8kHz to 48kHz adaptive sampling
- **Acoustic Echo Cancellation**: Real-time echo suppression
- **Voice Activity Detection**: Intelligent silence detection
- **Audio Enhancement**: Bandwidth extension and quality improvement

4.2 Text-to-Speech (TTS)

Neural voice synthesis with human-like naturalness[4].

4.2.1 TTS Engine Specifications

- **Voice Quality**: 4.8/5.0 Mean Opinion Score (MOS)
- **Latency**: <300ms for real-time synthesis
- **Languages**: 40+ languages with native pronunciation
- **Voice Variety**: 100+ premium voices with emotional range
- **Custom Voices**: Voice cloning with 30+ hours of training data

4.2.2 Advanced Voice Features

- **Prosody Control**: Pitch, rate, and volume modulation
- **SSML Support**: Speech Synthesis Markup Language compliance
- **Emotional Synthesis**: Context-aware emotional expression
- **Voice Biometrics**: Speaker verification and authentication
- **Real-time Streaming**: Low-latency streaming synthesis

4.3 Voice Processing Pipeline

- **Audio Capture**: Multi-channel audio input processing
- **Preprocessing**: Noise reduction and signal enhancement
- **Feature Extraction**: Mel-frequency cepstral coefficients (MFCC)
- **Neural Processing**: Deep learning model inference
- **Post-processing**: Output refinement and optimization

5. Integration & APIs

5.1 API Architecture

Comprehensive RESTful APIs and real-time communication protocols[3].

5.1.1 RESTful API Specifications

- **API Version**: OpenAPI 3.0 specification
- **Authentication**: OAuth 2.0, JWT tokens, API keys
- **Rate Limiting**: Configurable limits per client/endpoint
- **Documentation**: Interactive Swagger/OpenAPI documentation
- **Versioning**: Semantic versioning with backward compatibility

5.1.2 Real-time Communication

- **WebSocket Support**: Bi-directional real-time messaging
- **Server-Sent Events**: One-way streaming for live updates
- **GraphQL Endpoint**: Flexible data querying and mutations
- **Webhook Infrastructure**: Event-driven integrations with retry logic

5.2 Enterprise System Integrations

5.2.1 Telephony Systems[3][4]

- **Twilio Integration**: Programmable voice and messaging APIs
- **Teams Integration**: Microsoft Teams calling and collaboration
- **SIP Protocol**: Industry-standard Session Initiation Protocol
- **WebRTC Support**: Browser-based real-time communication
- **PBX Connectivity**: Integration with major PBX systems (Avaya, Cisco, etc.)

5.2.2 CRM Platforms[3]

- **Salesforce**: Lightning components, Apex triggers, Einstein AI
- **Microsoft Dynamics**: Power Platform integration, Azure AD sync
- **HubSpot**: Custom objects, workflows, reporting dashboards
- **SAP Customer Experience**: Business rule engine, process automation
- **Custom CRM**: Flexible API adapters for proprietary systems

5.2.3 Business Intelligence

- **Tableau**: Real-time data connections, custom visualizations
- **Power BI**: DirectQuery, composite models, report embedding
- **Looker**: LookML modeling, custom dimensions and measures
- **Qlik Sense**: Associative modeling, self-service analytics

5.3 Middleware & Integration Layer

- **API Gateway**: Kong or AWS API Gateway with security policies
- **Message Broker**: Apache Kafka for high-throughput event streaming
- **ETL Capabilities**: Data transformation and synchronization
- **Service Mesh**: Istio or Linkerd for microservice communication

6. Performance & Reliability

6.1 Performance Benchmarks

6.1.1 Response Time Metrics

- **Voice Response**: <300ms for 95th percentile
- **Text Processing**: <100ms for complex queries
- **API Calls**: <50ms for cached data retrieval
- **Database Queries**: <10ms for indexed operations
- **End-to-End Latency**: <500ms for complete conversation flow

6.1.2 Throughput Specifications

- **Concurrent Conversations**: 15,000+ simultaneous sessions
- **Message Processing**: 10,000+ messages per second
- **API Requests**: 100,000+ requests per minute
- **Voice Calls**: Real-time processing without quality degradation
- **Data Ingestion**: 1TB+ per day sustained throughput

6.2 Scalability & High Availability

6.2.1 Auto-Scaling Capabilities[2]

- **Horizontal Scaling**: Dynamic pod scaling based on CPU/memory utilization
- **Vertical Scaling**: Resource allocation optimization
- **Predictive Scaling**: ML-based demand forecasting
- **Geographic Distribution**: Multi-region deployment support

6.2.2 High Availability Architecture

- **Uptime SLA**: 99.99% availability guarantee
- **Disaster Recovery**: RTO <15 minutes, RPO <5 minutes
- **Fault Tolerance**: Automatic failover and recovery
- **Load Distribution**: Active-active deployment across zones

6.3 Monitoring & Observability

6.3.1 Performance Monitoring

- **Real-time Metrics**: CPU, memory, network, and application metrics
- **Custom Dashboards**: Grafana-based visualization
- **Alerting System**: Prometheus with automated incident response
- **Distributed Tracing**: Jaeger for end-to-end request tracking

6.3.2 Business Intelligence

- **Conversation Analytics**: Intent analysis, sentiment tracking
- **Performance KPIs**: Resolution rates, escalation metrics
- **User Behavior**: Interaction patterns and journey analysis
- **Quality Metrics**: Satisfaction scores, accuracy measurements

7. Security & Compliance

7.1 Security Architecture

7.1.1 Multi-Layered Security Framework

- **Zero-Trust Architecture**: Identity-based access control
- **Network Segmentation**: Micro-segmentation with firewalls
- **DDoS Protection**: Traffic analysis and rate limiting
- **Web Application Firewall**: OWASP Top 10 protection
- **Intrusion Detection**: ML-based anomaly detection

7.1.2 Data Protection

- **Encryption at Rest**: AES-256 encryption for stored data
- **Encryption in Transit**: TLS 1.3 for all communications
- **Key Management**: Hardware Security Module (HSM) integration
- **Data Masking**: PII anonymization in non-production environments
- **Secure Deletion**: Cryptographic erasure for data disposal

7.1.3 Access Control & Authentication

- **Multi-Factor Authentication**: TOTP/HOTP with biometric options
- **Single Sign-On**: SAML 2.0 and OAuth 2.0 support
- **Role-Based Access Control**: Granular permission management
- **Privileged Access Management**: Session recording and monitoring
- **API Security**: JWT tokens with refresh rotation

7.2 Compliance Certifications

7.2.1 International Standards[2][3]

- **SOC 2 Type II**: Service Organization Control certification
- **ISO 27001**: Information Security Management certification
- **PCI DSS Level 1**: Payment card data protection compliance
- **FedRAMP Moderate**: Government cloud service authorization
- **NIST Cybersecurity Framework**: Implementation compliance

7.2.2 Regional Compliance[2][3]

- **GDPR**: European data protection regulation compliance
- **CCPA**: California Consumer Privacy Act compliance
- **LGPD**: Brazilian data protection law compliance
- **PIPEDA**: Canadian privacy legislation compliance
- **Data Localization**: Regional data residency requirements

7.2.3 Industry-Specific Compliance

- **HIPAA**: Healthcare data protection (with BAA)
- **Financial Services**: Banking regulation compliance
- **Government**: Security clearance and classification support
- **Telecommunications**: Regulatory compliance frameworks

8. Deployment Options

8.1 Cloud Deployment Models

8.1.1 Public Cloud

- **AWS Deployment**: EC2, EKS, RDS, and managed services
- **Azure Integration**: Virtual machines, AKS, and cognitive services
- **Google Cloud**: GKE, Cloud SQL, and AI Platform
- **Multi-Cloud Support**: Vendor-agnostic deployment strategies

8.1.2 Private Cloud

- **On-Premises**: Dedicated infrastructure deployment
- **VMware Integration**: vSphere and vCenter compatibility
- **OpenStack**: Open-source cloud platform support
- **Hybrid Connectivity**: Secure VPN and dedicated connections

8.1.3 Hybrid Deployment

- **Edge Computing**: Local processing with cloud coordination
- **Data Residency**: Selective data placement strategies
- **Workload Distribution**: Optimal resource utilization
- **Disaster Recovery**: Cross-environment backup and recovery

8.2 Container Orchestration

8.2.1 Kubernetes Specifications

- **Version Support**: Kubernetes 1.25+ compatibility
- **Helm Charts**: Standardized deployment packages
- **Ingress Controllers**: NGINX, Istio, or cloud-native options
- **Storage Classes**: Persistent volume management
- **Service Mesh**: Advanced networking and security

8.2.2 Container Management

- **Docker Images**: Optimized container images
- **Registry Integration**: Private and public registry support
- **Image Scanning**: Security vulnerability assessment
- **Resource Limits**: CPU and memory constraints

- **Health Checks**: Liveness and readiness probes

9. Hardware Requirements

9.1 Server Specifications

9.1.1 AI/LLM Servers (Recommended Configuration)[3][4]

```
Component: GPU LLM Servers (x2)
CPU: 16 vCPU (Intel Xeon or AMD EPYC)
Memory: 128 GB DDR4 ECC
Storage: 1 TB NVMe SSD
GPU: 4x NVIDIA A100 (80GB) or L40 GPUs
Network: 25Gbps Ethernet with redundancy
```

9.1.2 Application Servers

```
Component: API Gateway & Web Interface (x3)
CPU: 8 vCPU
Memory: 64 GB DDR4
Storage: 250 GB NVMe SSD
Network: 10Gbps Ethernet
Load Balancing: Layer 4/7 with health checks
```

9.1.3 Database Servers

```
Component: SQL Server Cluster (x2)
CPU: 16 vCPU
Memory: 256 GB DDR4 ECC
Storage: 1 TB NVMe SSD (SAN-backed)
Configuration: Always On HA with synchronous replication
```

9.1.4 Analytics & Monitoring

```
Component: Analytics Server
CPU: 4 vCPU
Memory: 64 GB DDR4
Storage: 200 GB NVMe SSD
Purpose: Real-time dashboards and visualization
```

9.2 Storage Requirements

9.2.1 Primary Storage

- **Performance Tier**: NVMe SSD arrays for active workloads
- **Capacity Tier**: High-capacity SATA SSDs for warm data
- **Archive Tier**: Object storage for long-term retention
- **RAID Configuration**: RAID 10 for performance and redundancy

9.2.2 Backup & Recovery

- **Local Backup**: High-speed disk arrays for rapid recovery
- **Remote Replication**: Geographic distribution for disaster recovery
- **Cloud Integration**: Hybrid cloud backup strategies
- **Retention Policies**: Automated lifecycle management

9.3 Network Infrastructure

9.3.1 Bandwidth Requirements

- **Internal Network**: 25Gbps+ for high-performance computing
- **External Connectivity**: 10Gbps+ for client communications
- **Backup Network**: Dedicated 10Gbps for data protection
- **Management Network**: Isolated 1Gbps for administration

9.3.2 Network Security

- **Firewalls**: Next-generation firewalls with DPI
- **VPN Concentrators**: Site-to-site and client VPN support
- **Load Balancers**: Application delivery controllers
- **Monitoring**: Network performance and security monitoring

10. Monitoring & Analytics

10.1 Real-time Monitoring

10.1.1 System Metrics

- **Infrastructure Monitoring**: CPU, memory, disk, and network utilization
- **Application Performance**: Response times, throughput, error rates
- **AI Model Metrics**: Inference latency, accuracy, and resource consumption
- **Voice Quality**: Audio quality metrics and speech recognition accuracy

10.1.2 Dashboard & Visualization[4]

- **Executive Dashboards**: High-level KPIs and business metrics
- **Operations Center**: Real-time system health and performance
- **AI Performance**: Model accuracy, response quality, and usage patterns
- **Custom Views**: Role-based dashboards for different stakeholders

10.2 Business Intelligence

10.2.1 Conversation Analytics

- **Intent Analysis**: Most common customer intents and topics
- **Sentiment Tracking**: Customer satisfaction and emotional trends
- **Resolution Metrics**: First-call resolution and escalation rates
- **Agent Performance**: Human agent handoff analysis

10.2.2 Operational Analytics

- **Call Volume Patterns**: Peak hours, seasonal trends, forecasting
- **System Performance**: Response times, availability, capacity utilization
- **Cost Analytics**: Resource utilization and operational efficiency
- **ROI Metrics**: Business impact and return on investment

10.3 Reporting & Insights

10.3.1 Automated Reporting

- **Scheduled Reports**: Daily, weekly, and monthly automated reports
- **Exception Reports**: Automated alerts for anomalies and issues
- **Compliance Reports**: Regulatory and audit compliance documentation
- **Performance Reviews**: Trend analysis and recommendations

10.3.2 Advanced Analytics

- **Predictive Analytics**: Demand forecasting and capacity planning
- **Machine Learning Insights**: Pattern recognition and optimization
- **A/B Testing**: Conversation flow and response optimization
- **Continuous Improvement**: Performance enhancement recommendations

Conclusion

The Wiserep AI platform represents a comprehensive enterprise-grade solution for conversational AI deployment. With its advanced technical architecture, robust security framework, and flexible deployment options, Wiserep AI is designed to meet the most demanding enterprise requirements while providing exceptional scalability, reliability, and performance.

The platform's modular design enables organizations to start with core functionality and expand capabilities as needed, ensuring a future-proof investment that can adapt to evolving business requirements and technological advances.

For detailed implementation planning, custom configuration requirements, or specific integration needs, please contact our technical team for comprehensive consultation and deployment planning services.

Document Version: 3.0
Last Updated: September 2025
Classification: Technical Specifications
Prepared by: Wiserep AI Technical Team