Cresta AI - Technical Assessment

Executive Summary

Assessment Date: January 2026
Platform Type: Cloud-native, AI-powered contact center intelligence platform
Deployment Model: SaaS (AWS-hosted)
Primary Use Cases: Real-time agent coaching, conversational AI, virtual agents, quality management

Cresta provides enterprise-grade AI solutions for contact centers, combining real-time agent assistance, autonomous AI agents, and conversation intelligence with a focus on security, compliance, and measurable business outcomes.

---

1. Platform Architecture

1.1 High-Level Architecture

Infrastructure Foundation:

- **Cloud Provider:** Amazon Web Services (AWS) - consolidated from previous multi-cloud architecture
- **Deployment Model:** Multi-tenant SaaS with customer-specific isolated databases
- **Architecture Type:** Microservices-based, containerized (Kubernetes/EKS)
- **Geographic Distribution:** Global AWS data centers with regional deployments

Core Technology Stack:

- **Frontend:** React-based web applications (TypeScript)
    - Monorepo structure using Lerna for package management
    - Vite build system for rapid development
    - Electron-based desktop agent application
- **Backend:** Python-based microservices architecture
- **Orchestration:** Kubernetes (Amazon EKS) with Argo Workflows
- **ML Infrastructure:** PyTorch framework on AWS compute infrastructure

1.2 AI/ML Architecture

Ocean-1 Foundational Model:

- Proprietary domain-specific model for contact centers
- Built on Mistral-based architecture with LoRA (Low-Rank Adaptation) fine-tuning
- Leverages foundation models (GPT-4, Claude) with heavy customer-specific fine-tuning
- Multi-model architecture supporting 20+ LLMs across different layers

Model Serving Infrastructure:

- Primary inference provider: Fireworks AI (for low-latency LLM serving)
- LoRA adapters enable domain-specific customizations without full model retraining
- Real-time processing requirements: <100ms latency for agent assistance
- Hybrid architecture: Combines generative AI with structured business logic

Key Technical Capabilities:

- Real-time speech recognition (ASR) with multiple provider support
- Natural Language Understanding (NLU) for intent detection
- Generative AI for content creation (summaries, responses)
- Agentic AI systems capable of reasoning, planning, and multi-step action

---

2. Data Flows & Processing

2.1 Real-Time Voice Processing Pipeline

Audio Capture & Processing:

1. **Ingestion:** Audio streams captured from 20+ CCaaS integrations (Amazon Connect, Five9, Genesys, etc.)
2. **Pre-processing:** `gowalter` service handles incoming audio, PII redaction, and storage
3. **Transcription:** ASR (Automatic Speech Recognition) generates partial and final transcripts
4. **Utterance Formation:** Audio chunks grouped into logical conversation units based on silence/speaker changes
5. **Storage:** Transcripts persisted to PostgreSQL database via `apiserver`

Real-Time Intelligence Flow:

```
Audio Input → gowalter (PII redaction) → ASR → Partial Transcripts →
apiserver → PostgreSQL → ML Services → Inference Graphs →
Real-time Guidance → Agent Desktop Application
```

Recovery Mechanisms:

- WebSocket failure recovery with audio replay capability
- No data loss design with buffering and retry logic
- Partial transcript refinement as additional context becomes available

2.2 Data Processing Layers

Business Logic Layer:

- Customer-specific policies and rules engine
- Configurable workflows without code (Opera interface)
- Real-time decision trees for agent guidance

ML Services Layer:

- Inference graphs orchestrate multiple model workflows
- Customer-specific model customizations via LoRA
- Multi-model ensemble for task-specific optimization

Integration Layer:

- Bidirectional data sync with CRM systems (Salesforce, Zendesk, etc.)
- Real-time metadata enrichment from customer databases
- Post-call summary push to ticketing/case management systems

---

3. Data Storage & Retention

3.1 Storage Architecture

Primary Data Stores:

- **Relational Database:** Amazon Aurora (PostgreSQL)
  - Customer-specific isolated databases (separate DB per customer)
  - Conversation transcripts, metadata, and interaction logs
  - Agent performance metrics and quality scores
- **Object Storage:** Amazon S3
  - Audio recordings (with PII redaction applied)
  - Training datasets (1-100GB per export)
  - Model artifacts and checkpoints
- **Document Storage:** Integrated knowledge bases and FAQs

Data Isolation:

- Strict tenant separation with dedicated databases per customer
- No cross-customer data access or model training
- Separate processing pipelines for each customer environment

3.2 Data Retention & Lifecycle

Retention Policies:

- Retention periods determined by:
    1. Service purpose requirements
    2. Legal/regulatory obligations
    3. Legitimate business interests (e.g., legal proceedings)
- Customer-configurable retention windows
- Automated data lifecycle management

Data Residency:

- Primary storage: Switzerland, EU, or EEA
- Customer-specific geographic constraints supported
- Data processing within committed regions

---

4. Security Controls & Architecture

4.1 Security Framework

Enterprise Security Posture:

- **Certifications:**
    - SOC 2 Type II (no findings)
    - ISO/IEC 27001:2013 (Information Security)
    - ISO/IEC 27701:2019 (Privacy Information Management)
    - ISO/IEC 42001 (AI Management System) - among first certified
    - PCI-DSS Service Provider Level 2
    - HIPAA compliant (Business Associate eligible)

Security Program Components:

- Secure Development Lifecycle (SDLC)
- Static application security testing (SAST)
- Dependency scanning and vulnerability management
- Mandatory secure coding training for all engineers
- Third-party penetration testing (regular cadence)
- CVSS scoring for vulnerability prioritization

- 24-hour security incident triage SLA

4.2 Access Controls

Authentication & Authorization:

- Role-based access control (RBAC)
- Multi-factor authentication (MFA) support
- Principle of least privilege enforcement
- Regular access reviews and audits

Network Security:

- Dedicated customer subdomains for traffic isolation
- Encrypted data transmission (TLS 1.2+)
- Web Application Firewall (WAF)
- DDoS protection via AWS Shield

Monitoring & Detection:

- Continuous security monitoring
- Centralized logging and SIEM integration
- Automated anomaly detection
- Security incident response team

4.3 Data Protection

Encryption:

- **At Rest:** AES-256 encryption for stored data
- **In Transit:** TLS 1.2+ for all data transmission
- **Key Management:** AWS KMS for encryption key lifecycle

PII Protection:

- Industry-leading PII redaction algorithms
- Automatic redaction at ingestion (before storage)
- Payment card data scrubbing (PCI-DSS compliant)
- Configurable PII handling policies per customer

Data Boundaries:

- No sensitive signals used in model training
- Protected attributes excluded (medical, financial metadata)

- Explicit consent required for data usage
- Customer data ownership maintained

---

5. Compliance & Governance

5.1 Regulatory Compliance

Data Privacy Regulations:

- GDPR compliant (EU operations)
- CCPA compliant (California operations)
- HIPAA eligible for healthcare customers
- Industry-specific compliance frameworks supported

AI Governance:

- ISO/IEC 42001 certified (AI Management)
- Responsible AI framework with transparency focus
- Bias detection and mitigation protocols
- Model explainability techniques (Chain-of-Thought, Model Critique)

5.2 Transparency & Explainability

Model Interpretability:

- Chain-of-Thought (CoT) reasoning for output transparency
- Model-based critique systems for quality assurance
- Evidence-based reasoning to support recommendations
- Customer-facing confidence scores and explanations

Monitoring & Validation:

- Post-processing validation layers
- Live performance tracking and drift detection
- Continuous model evaluation against test sets
- Regression testing for production deployments

5.3 Third-Party Processors

Current Sub-Processors:

- Fireworks AI (LLM inference)

- Various CCaaS providers (telephony integration)
- AWS (infrastructure)

Sub-Processor Management:

- Formal DPA (Data Processing Agreements)
- Regular security assessments
- Change notification requirements
- Audit rights maintained

---

6. Integration Architecture

6.1 Supported Integrations

Contact Center Platforms (20+ integrations):

- Amazon Connect
- Genesys Cloud
- Five9
- 8x8
- Twilio
- Nice inContact

CRM & Business Systems:

- Salesforce
- Zendesk
- Microsoft Dynamics
- HubSpot
- Marketo
- ServiceNow

Data & Analytics:

- Databricks
- Amazon SQS
- Custom data warehouses

6.2 Integration Methods

API & SDK Support:

- RESTful APIs with comprehensive documentation
- SDKs available in multiple languages (Ruby, Python, PHP, Node.js, Java, C#)
- WebSocket support for real-time streaming
- Webhook notifications for event-driven integrations

Authentication:

- OAuth 2.0 for API authentication
- API key management with rotation policies
- Service-to-service authentication

Data Exchange:

- Real-time bidirectional sync
- Batch data imports/exports
- Streaming data pipelines
- ETL support for historical data migration

---

7. Performance & Scalability

7.1 Performance Metrics

Latency Requirements:

- Real-time agent assistance: <100ms response time
- ASR transcription: Near real-time with <2s lag
- API response times: <500ms for standard operations
- ML inference: <100ms for production models

Throughput:

- Supports large enterprise scale (10,000+ agents)
- Concurrent conversation processing
- Horizontal scaling via Kubernetes
- Auto-scaling based on load

7.2 Reliability & Availability

High Availability Design:

- Multi-AZ deployment in AWS

- Redundant system components
- Load balancing across availability zones
- Automated failover mechanisms

Disaster Recovery:

- Regular backup procedures
- Point-in-time recovery capabilities
- Cross-region replication (customer-configurable)
- Business continuity planning

SLA Commitments:

- Service availability targets (enterprise tier)
- Performance guarantees for real-time features
- Support response times (5AM-6PM PST)
- Incident escalation procedures

---

8. Operational Controls

8.1 Monitoring & Observability

System Monitoring:

- Application performance monitoring (APM)
- Infrastructure health monitoring
- Custom dashboards per customer environment
- Alerting for anomalies and thresholds

ML Operations (MLOps):

- Model performance tracking
- Data drift detection
- Model versioning and rollback capabilities
- A/B testing framework for model improvements

8.2 Quality Assurance

Testing Framework:

- Automated unit and integration testing
- End-to-end (E2E) testing automation

- Simulation testing for AI agents
- LLM-based evaluation frameworks
- Regression testing for production changes

Quality Metrics:

- 100% conversation capture for analysis
- Automated quality scoring (QA)
- Agent performance benchmarking
- Customer satisfaction tracking (CSAT, NPS)

8.3 Change Management

Deployment Process:

- Continuous integration/continuous deployment (CI/CD)
- Phased rollout strategy for customer deployments
- Feature flags for controlled releases
- Rollback procedures for failed deployments

Customer Onboarding:

- Dedicated implementation support
- Forward-deployed engineers for enterprise customers
- Center of Excellence establishment
- Training and enablement programs

---

9. Risk Assessment

9.1 Technical Risks

Risk	Severity	Mitigation
Model hallucination/inaccuracy	High	Hybrid architecture with business logic guardrails, continuous validation
Integration complexity	Medium	Extensive SDK support, dedicated integration team
Data breach/unauthorized access	High	Multi-layer security controls, SOC 2 compliance, encryption
System downtime	Medium	HA architecture, redundancy, SLA commitments
Scalability limitations	Low	Cloud-native design, auto-scaling, proven at scale

9.2 Compliance Risks

Risk	Severity	Mitigation
GDPR violations	High	ISO 27701 certified, DPA enforcement, data residency controls
HIPAA violations (healthcare)	High	HIPAA compliance, BAA available, PHI protection
Consent management (voice recording)	High	Enhanced disclosure requirements, customer consent workflows
AI bias/discrimination	Medium	Diverse training data, bias monitoring, fairness evaluation
Cross-border data transfer	Medium	Regional deployment options, standard contractual clauses

9.3 Operational Risks

Known Challenges:

- Customer-specific AI model requires training period (weeks to months)
- Complex enterprise deployments require dedicated resources
- Backend integration complexity reported by some users
- Performance edge cases in very high-volume scenarios

Legal Considerations:

- Ongoing class action lawsuit regarding voice recording consent (Galanter v. Cresta)
- Third-party processor liability concerns
- Enhanced disclosure requirements for AI usage

---

10. Cost & Consumption Model

10.1 Pricing Structure

Licensing Model:

- Enterprise subscription pricing (not SMB-focused)
- Per-agent-per-month pricing typical
- Indicative pricing: ~$1.20/unit/month for Agent Assist Chat (unverified)
- Custom pricing for large deployments
- Private offers available via AWS Marketplace

Cost Drivers:

- Number of licensed agents
- Feature set selection (Agent Assist, AI Agent, Conversation Intelligence)
- Integration complexity
- Storage and retention requirements
- Support tier (white-glove enterprise support included)

10.2 Resource Consumption

Infrastructure Costs:

- AWS compute for ML inference (EC2, EKS)
- Database storage (Aurora)
- Object storage (S3) for audio and training data
- Data transfer costs for real-time streaming
- Third-party AI inference (Fireworks AI)

Optimization Features:

- LoRA adapters reduce compute requirements vs. full model fine-tuning
- Spot instances for cost-effective batch training
- Efficient model serving infrastructure
- Caching and optimization layers

---

11. Reporting & Analytics

11.1 Standard Reports

Agent Performance:

- Real-time coaching effectiveness
- Quality scores and trends
- Behavior benchmarking vs. top performers
- Skills gap analysis

Operational Metrics:

- First call resolution (FCR) rates
- Average handle time (AHT)
- Customer satisfaction (CSAT/NPS)
- Transfer and escalation rates

Business Intelligence:

- Conversation insights and trends
- Topic and sentiment analysis
- Churn risk identification
- Revenue opportunities per interaction

11.2 AI Analyst

Natural Language Querying:

- Ask questions in plain English about conversation data
- Automated insight generation
- Anomaly detection and alerting
- Custom report generation

Data Visualization:

- Interactive dashboards
- Trend analysis charts
- Comparative performance views
- Drill-down capabilities to individual conversations

---

12. Recommendations

12.1 Technical Due Diligence

Before Implementation:

1. **Legal Review:** Assess consent management requirements for voice recording, particularly for California operations
2. **Integration Assessment:** Evaluate complexity of existing CCaaS/CRM integrations
3. **Data Residency:** Confirm regional data storage requirements align with capabilities
4. **Performance Testing:** Conduct pilot with representative workload and latency requirements
5. **Security Validation:** Review SOC 2 report and conduct vendor security assessment

12.2 Implementation Best Practices

Success Factors:

1. Allocate dedicated "AI linguist" or model tuning specialist
2. Plan for 2-3 month model training and optimization period

3. Establish Center of Excellence for ongoing refinement
4. Ensure executive sponsorship for organizational change management
5. Define clear KPIs and success metrics upfront

12.3 Ongoing Governance

Operational Oversight:

1. Regular review of model performance and accuracy
2. Periodic compliance and security audits
3. User feedback collection and product improvement cycle
4. Vendor relationship management and SLA monitoring
5. Cost optimization reviews (compute, storage)

---

13. Conclusion

Strengths:

- Enterprise-grade security and compliance posture (SOC 2, ISO 27001/27701/42001, HIPAA)
- Proven at scale with Fortune 500 customers
- Advanced AI capabilities with hybrid guardrail architecture
- Comprehensive integration ecosystem (20+ CCaaS platforms)
- Strong focus on explainability and responsible AI

Considerations:

- Enterprise pricing not suitable for SMB/mid-market
- Complex implementation requiring dedicated resources
- Legal considerations around voice recording consent (ongoing litigation)
- Customer-specific model training requires patience and iteration
- Vendor lock-in considerations with proprietary Ocean-1 model

Overall Assessment:
Cresta represents a mature, enterprise-grade AI platform with robust security controls, comprehensive compliance certifications, and proven results for large-scale contact center operations. The technical architecture is well-designed for scale, reliability, and security. Primary considerations are around implementation complexity, cost structure, and ensuring proper consent management for voice recording applications.

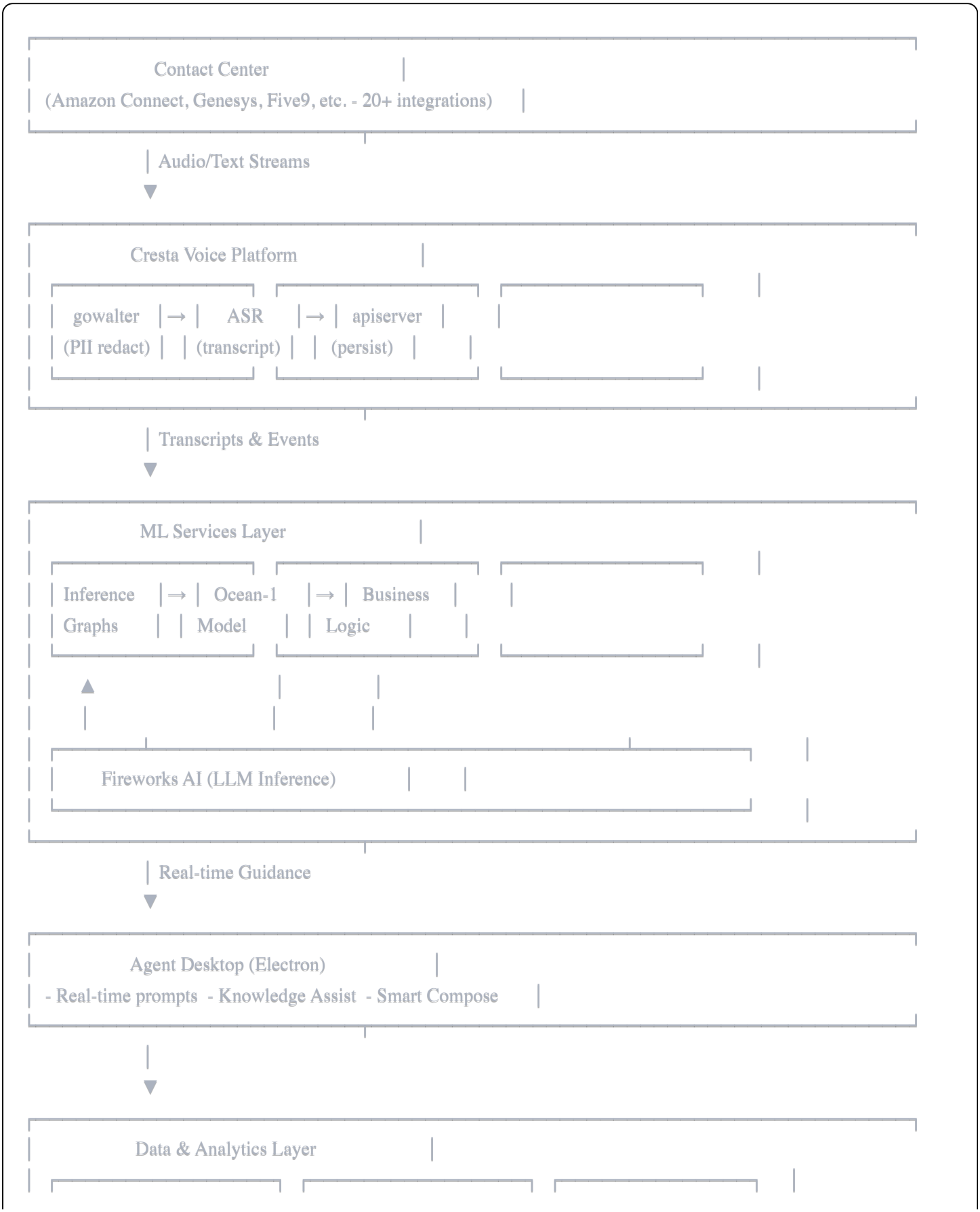
Risk Rating: Medium-Low (technical) / Medium (legal-compliance)
Suitability: Large enterprises with complex contact center operations and security/compliance requirements

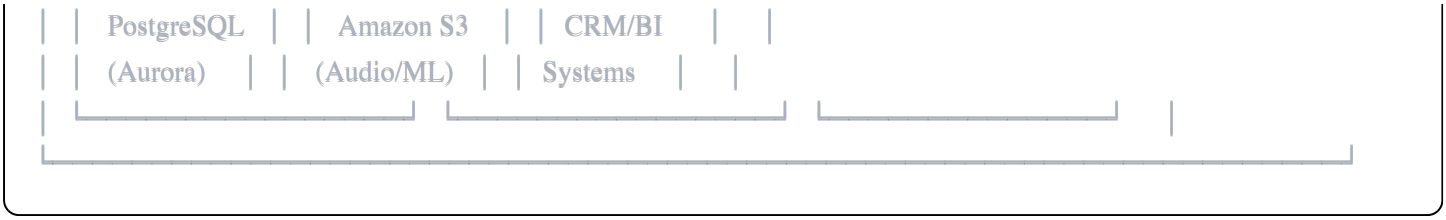
Recommendation: Proceed with pilot, ensure legal review of consent management, validate integration complexity

---

Appendix

A. Reference Architecture Diagram





B. Key Technical Specifications

Component	Technology	Purpose
Frontend	React, TypeScript, Vite	Web applications, agent UI
Desktop App	Electron	Agent assist desktop application
Backend Services	Python, microservices	Business logic, orchestration
ML Framework	PyTorch	Model training and inference
Container Orchestration	Kubernetes (EKS)	Service deployment, scaling
Workflow Engine	Argo Workflows	ML pipeline orchestration
Database	PostgreSQL (Aurora)	Transactional data, transcripts
Object Storage	Amazon S3	Audio recordings, ML artifacts
ASR	Multiple providers	Speech-to-text conversion
LLM Inference	Fireworks AI, Ocean-1	Real-time model serving
Monitoring	Custom dashboards, APM	Observability, alerting

C. Compliance Certifications Summary

Certification	Scope	Status	Last Updated
SOC 2 Type II	Security, Availability, Confidentiality	Active	December 2025
ISO 27001	Information Security Management	Active	2025
ISO 27701	Privacy Information Management	Active	2025
ISO 42001	AI Management System	Active	2025
PCI-DSS	Payment Card Industry	Level 2 SP	Current
HIPAA	Health Information Privacy	BAA Available	2023+

D. Contact Information

Cresta Security Team: security@cresta.ai
Enterprise Sales: partners@cresta.ai
Support Hours: Monday-Friday, 5AM-6PM PST
Vulnerability Disclosure: security@cresta.ai (PGP available)
Trust Center: https://trust.cresta.com

Document Version: 1.0
Assessment Date: January 19, 2026
Prepared for: Technical Due Diligence Review