Chatley AI, Inc.

AI Security Whitepaper

Building Trust in the Age of AI: Chatley's Approach to Secure AI Deployment

Trust Center Document | Public | April 2026
April 2026
chatley.ai/trust-portal

1. Introduction

As conversational AI becomes integrated into business-critical workflows — answering customer calls, booking appointments, qualifying leads — the security of these systems is not a technical footnote. It is the product. This whitepaper describes how Chatley AI, Inc. approaches the unique security challenges of deploying AI voice and SMS agents at scale for enterprise customers.

Chatley AI is a deployment platform, not a model builder. We do not train or own foundational AI models. We use enterprise APIs from leading providers (OpenAI, Anthropic) via certified infrastructure (Voice infrastructure provider) and are responsible for deploying and configuring those models securely on behalf of our customers.

2. The AI Security Landscape

Deploying LLM-powered voice agents introduces security considerations that traditional application security does not fully address:

Prompt Injection	Malicious attempts by callers to override agent instructions, extract system prompt content, or cause th
Knowledge Base Leakage	The risk that an agent reveals proprietary business information (pricing, internal scripts, confidential pro
PII Handling	Voice calls inherently capture spoken personal information. Correct configuration ensures PII is handle
Payment Card Data	Agents that handle payment collection must be configured to prevent spoken card numbers from being
PHI in Healthcare	Healthcare agents may encounter Protected Health Information. HIPAA-compliant configuration is requ
Model Hallucination	LLMs may generate plausible but incorrect information. Knowledge base constraints and scope limits r

3. Chatley's AI Security Framework

3.1 Certified Infrastructure Foundation

All Chatley AI voice processing runs on Voice infrastructure provider's infrastructure, which holds active **SOC 2 Type II, HIPAA, PCI DSS Level 1, GDPR, and CCPA** certifications. This is the first and most important layer of the AI security framework — the underlying platform has been independently audited and certified.

Chatley AI's voice infrastructure runs on a SOC 2 Type II certified platform.

3.2 Knowledge Base Integrity

Each agent operates within a strictly defined knowledge base and system prompt configured by the customer. Controls:

- • Knowledge base content is version-controlled. All changes are tracked with timestamps and user attribution.
- • Access to modify agent knowledge bases and system prompts is RBAC-gated — only authorized users within the customer organization can make changes.
- • Agents are explicitly instructed to decline questions outside their defined scope. The system prompt includes guardrails that take precedence over conversational pressure.
- • Operators can review all agent conversations in real time and retroactively via the Chatley dashboard transcript viewer.

3.3 Prompt Injection Defense

Prompt injection is the primary adversarial threat in LLM-based voice systems — a caller attempting to override the agent's instructions through conversational input. Chatley AI's defense approach:

- • System prompt architecture: customer-defined system prompts are structured to explicitly resist override attempts. Instruction hierarchy is enforced at the prompt level.
- • Anomaly monitoring: unusual conversation patterns — callers repeating override attempts, requests for system information, unexpected topic pivots — are flagged in monitoring.
- • Scope enforcement: agents are trained (via their system prompts) to respond only to in-scope queries. Persistent out-of-scope requests trigger graceful escalation to a human.
- • No training exposure: because we do not train our own models, there is no Chatley-specific model for adversaries to target with data poisoning or model evasion attacks.

3.4 PII and Sensitive Data Handling

Voice calls inherently contain personally identifiable information — names, phone numbers, addresses, health discussions, and potentially payment information. Chatley AI's data handling framework:

- • Data minimization: only data necessary for the service function is collected and stored.
- • Transcript retention: configurable per customer (default 90 days). Zero-Retention Mode available for Enterprise customers on dedicated infrastructure — no transcript stored after call completion.

- • PCI payment data: when PCI compliance mode is active, recording and transcription are disabled during the payment window. Card numbers are collected via DTMF (keypad) or SMS payment link — never spoken aloud or transcribed. Cardholder data never enters any Chatley AI system.
- • PHI in healthcare: when HIPAA mode is active, PHI handling follows HIPAA requirements. A Business Associate Agreement is executed before any PHI is processed.

3.5 AI Provider Data Governance

The AI models powering Chatley AI agents (via Voice infrastructure provider's infrastructure) are operated under enterprise API agreements with the following data governance protections:

- • **Zero-training guarantee — OpenAI, LLC:** Customer data processed via Chatley AI is not used to train OpenAI's foundational models.
- • **Zero-training guarantee — Anthropic, PBC:** Customer data processed via Chatley AI is not used to train Anthropic's foundational models.
- • Voice infrastructure provider enforces equivalent zero-training commitments across all its AI sub-processor agreements.

Your business conversations, customer interactions, and proprietary knowledge base content will never become training data for any AI model.

3.6 Human-in-the-Loop Oversight

Chatley AI is designed for human oversight, not human replacement. Operators retain full visibility and control:

- • Real-time monitoring: operators can view active conversations in the Chatley dashboard as they happen.
- • Manual takeover: any conversation can be taken over with a manual reply. The AI pauses immediately and the operator is in direct control.
- • Escalation triggers: configurable keywords (emergency, urgent, human, etc.) automatically flag conversations for human review.
- • Full audit trail: every message in every conversation is logged with timestamps, sender attribution (AI vs. human), and outcome classification.

4. Our Commitment to Transparency

We believe customers deploying AI in their business deserve to understand exactly how it works and what protections are in place. Our commitments:

- • Complete Subprocessor Registry publicly available at chatley.ai/subprocessors
- • Security Attestation Letter available upon request for enterprise deals
- • Enterprise Security Questionnaire responses (CAIQ-lite) available upon request
- • Infrastructure SOC 2 Type II report available to enterprise customers under NDA

- • Penetration test executive summaries available to enterprise customers under NDA
- • Architecture diagrams, data flow maps, and compliance framework documentation available through the Trust Portal

5. Conclusion

Chatley AI's AI security framework rests on three pillars: certified infrastructure (Voice infrastructure provider's SOC 2 Type II, HIPAA, and PCI DSS Level 1 platform), correct deployment configuration (HIPAA flags, PCI squad mode, prompt guardrails), and human oversight (real-time monitoring, manual takeover, escalation triggers). Together, these pillars give enterprise customers confidence that AI is working for their business safely, securely, and transparently.

Security inquiries: **security@chatley.ai** | Trust Portal: **chatley.ai/trust-portal**