AI Security Practices for AutoQA (Aligned with ISO/IEC 23894)

SQM’s AutoQA services apply a structured, lifecycle-based approach to identifying, assessing, and reducing AI-related risks. We align our AI risk management program to the principles and guidance described in **ISO/IEC 23894:2023 (Information technology — Artificial intelligence — Guidance on risk management)**, which is designed to help SQM integrate AI risk management into how we build and operate our AI-enabled products and services.

Core principle: We do **not** train our models on customer data: Customer call recordings, transcripts, and QA outputs are processed to deliver the service, but **we do not use customer data to train our foundation models**.

1) Governance and accountability

We maintain clear ownership and escalation paths for AI risk decisions, including:

- Defined roles for approving AI use cases and changes
- Documentation of intended use, limitations, and risk acceptance
- Ongoing review of AI risks as product capabilities evolve

2) Risk assessment across the AI lifecycle

We evaluate AI risks throughout design, development, deployment, and ongoing operations, including:

- Threat and misuse-case analysis for AutoQA workflows
- Assessment of privacy, security, and operational risks tied to model behavior
- Change management for model/prompt/config updates (with risk review before rollout)

3) Data protection for recordings and transcripts

Because AutoQA processes sensitive customer conversations, we apply strong data protection controls such as:

- Controlled access to audio, transcripts, and QA outputs (least privilege / role-based access)
- Encryption in transit and at rest for recordings and derived artifacts
- Configurable retention and deletion workflows aligned to customer requirements
- Redaction of all Personally Identifiable Information (PII) and Protected Health Information (PHI) from the transcripts
- All data is encrypted using industry standard encryption technologies while in-transit and at rest

4) Secure AI design to reduce data leakage and misuse

We implement controls intended to reduce common AI security failure modes, including:

- Guardrails to limit unintended disclosure of sensitive data
- Controls to reduce prompt-based manipulation in AI-driven workflows

- Segmentation and access controls that support customer tenant isolation

5) Quality, robustness, and monitoring

We continuously monitor AI performance and operational signals so risks can be detected and managed early:

- Monitoring for drift, anomalies, and unexpected outputs
- Human review pathways for exceptions and high-impact cases
- Repeatable evaluation processes for key AutoQA behaviors (e.g., scoring consistency)

6) Incident response and continuous improvement

We integrate AI-specific risk response into broader security operations:

- Documented response processes for AI-related incidents (e.g., suspected leakage, misuse, abnormal access)
- Post-incident reviews that drive corrective actions and control improvements
- Periodic reassessment to keep pace with changing threats and customer expectations

7) LLM-specific threat modeling and guardrails

We design controls to reduce common LLM application risks (e.g., prompt injection and sensitive information disclosure) referenced by the OWASP Top 10 for LLM Applications.
Examples include:

- We do not use customer data to train our foundation models
- Input/output handling controls to reduce unintended disclosure of sensitive data
- Guardrails to mitigate prompt manipulation and unsafe or out-of-policy outputs
- Tenant isolation controls to prevent cross-customer data exposure

9) Secure use of retrieval and embeddings

Where retrieval or vector search is used, we apply controls to reduce “vector and embedding” weaknesses (e.g., over-broad retrieval, data leakage through retrieved context) highlighted in OWASP’s LLM risk guidance.
This includes scoped retrieval, authorization-aware access checks, and logging around retrieval access patterns.

10) Monitoring, testing, and continuous improvement

We continuously evaluate the service to detect and respond to security and reliability issues:

- Monitoring for anomalous access and unusual export/download patterns
- Testing and review pathways for high-impact cases
- Corrective actions and ongoing improvements based on incidents, findings, and evolving threats (ISO/IEC 23894 emphasizes integration and ongoing risk management).