



II ElevenLabs

Service Definition Document

ElevenLabs Agents Platform (Conversational AI)

G-Cloud 15 – Lot 2b: Cloud Software

- 1. Service Overview
- 2. Core Conversational Architecture
 - 2.1 Voice and Conversation Engine
 - 2.2 Agent Orchestration and Reasoning
 - 2.3 Monitoring, Evaluation, and Analysis
- 3. Key Capabilities
 - 3.1 Voice Activity Detection and Turn-Taking
 - 3.2 Agent Behaviour and Conversation Logic
 - 3.3 Language Model Integration
 - 3.4 Voice and Language Capabilities
 - 3.5 Knowledge Integration
 - 3.6 Tools and System Integration
 - 3.7 Personalisation and Context Injection

- 3.8 Agent Testing, Analysis, and Operations
 - 3.9 Service Scope and Responsibility Boundaries
- 4. Onboarding, Support and Service Levels
 - 4.1 Onboarding
 - 4.2 Support and After-Sales Service
 - 4.3 Implementation and Engineering Services (Forward Deployed Engineering)
 - 4.4 Service Levels and Performance
 - 4.5 Service Constraints
- 5. Hosting, Deployment and Data Residency
 - 5.1 Hosting Model
 - 5.2 Data Residency
- 6. Security & Privacy
 - 6.1 Security Governance
 - 6.2 Information Security Controls
 - 6.3 Vulnerability Management
 - 6.4 Protective Monitoring
 - 6.5 Incident Management
 - 6.6 Access Control
 - 6.7 Certifications and Assurance
- 7. Backup, Restore and Disaster Recovery
 - 7.1 Backup and Restore
 - 7.2 Disaster Recovery and Business Continuity
- 8. Operations, Maintenance and Updates
 - 8.1 Operational Monitoring and Observability
 - 8.2 Product-Specific Monitoring and Quality Evaluation
 - 8.3 Logging, Data Collection and Audit Support
 - 8.4 Service Updates and API Versioning
 - 8.5 Outage and Maintenance Management
 - 8.6 Continuous Improvement
- 9. Service Constraints
 - 9.1 Usage Constraints
 - 9.2 Customisation
- 10. Exit Management and Data Access
 - 10.1 Data Ownership
 - 10.2 Service Offboarding and Data Export
 - 10.3 Termination, Deletion and Wind-Down
- 11. Technical Requirements

1. Service Overview

The ElevenLabs Agents Platform is a managed, cloud-based conversational AI service for designing, deploying, monitoring, and operating voice-first and multimodal agents. The service provides an end-to-end conversational platform that combines real-time voice processing, language reasoning, workflow orchestration, system integration, and operational monitoring within a single managed service.

The platform enables government organisations to deliver natural, real-time conversational experiences across telephony and digital channels for citizen-facing and internal use cases. Agents are capable of listening, reasoning, responding, and taking action while operating within defined governance, security, and compliance controls.

In addition to real-time conversation handling, the service includes built-in capabilities for agent configuration, workflow management, knowledge integration, testing, monitoring, evaluation, and operational analysis. This allows organisations to manage the full lifecycle of conversational agents, from initial design and deployment through to ongoing optimisation and governance.

The service is delivered as a software-as-a-service (SaaS) offering. Customers configure, deploy, and operate agents through web-based tools and APIs without managing underlying infrastructure, model hosting, real-time audio pipelines, or monitoring systems.

2. Core Conversational Architecture

The ElevenLabs Agents Platform coordinates a set of integrated architectural components to deliver real-time, production-grade conversational agents. These components operate as a unified system rather than standalone services.

2.1 Voice and Conversation Engine

The Voice Engine provides low-latency, real-time speech interaction capabilities:

- **Speech-to-Text (ASR):** Fine-tuned speech recognition models convert live or streamed audio into text with low latency. The ASR pipeline is optimised for conversational use,

supporting continuous speech, partial results, and rapid turn transitions.

- **Text-to-Speech (TTS):** Generated responses are synthesised into natural-sounding speech using ElevenLabs' low-latency voice models, supporting multilingual output and configurable voices.
- **Turn Detection and Interruption Handling:** A proprietary conversation timing system manages turn-taking, interruption (barge-in), silence detection, and response timing. This enables full-duplex interactions where users may interrupt or change direction without rigid command structures.

2.2 Agent Orchestration and Reasoning

The Agent Orchestration layer manages how agents think, decide, and act during conversations:

- **Language Reasoning:** The platform invokes a large language model to interpret intent, maintain conversational context, manage dialogue state, and generate responses.
- **Workflows and Decision Logic:** Conversations are structured using workflows that define decision points, branching logic, task execution, and escalation paths.
- **Knowledge Integration and Retrieval:** Agents retrieve approved information at runtime to ground responses in authoritative content.
- **Tool Invocation and External Actions:** Agents can securely invoke tools and integrations to perform actions beyond conversation.

2.3 Monitoring, Evaluation, and Analysis

The platform includes built-in operational and quality oversight capabilities:

- **Conversation Logging and Results:** Transcripts, audio, and structured outcomes are captured for review and analysis.
- **Evaluations and Testing:** Support for testing, simulation, and evaluation of agent behaviour and performance.

- **Monitoring and Analytics:** Operational metrics and usage indicators are available to support governance, quality assurance, and continuous improvement.

These components are orchestrated in real time to provide a single, managed conversational service suitable for high-availability, citizen-facing environments.

3. Key Capabilities

3.1 Voice Activity Detection and Turn-Taking

The platform incorporates voice activity detection (VAD) as part of its real-time audio pipeline. VAD is used to detect when a speaker starts and stops talking and to coordinate handoffs between listening and speaking states. In combination with turn-taking logic, this enables the platform to:

- Detect end-of-utterance without requiring explicit commands
- Support natural interruptions and barge-in during agent speech
- Prevent overlapping speech and unintended truncation
- Improve responsiveness in telephony and voice-based channels

This behaviour is designed for live, accessibility-critical services rather than scripted IVR-style interactions.

3.2 Agent Behaviour and Conversation Logic

Agent behaviour is defined through configurable components that control how the agent reasons, responds, and acts.

- **Workflows:** Agents are constructed using workflows that define conversation structure, branching logic, task execution, and escalation paths. Complex services can be decomposed into scoped sub-tasks or sub-agents with defined responsibilities and controlled access to tools and data.

- **System Prompts and Instructions:** Each agent includes system-level instructions defining role, tone, constraints, and behavioural guidelines. These are used to enforce service boundaries, compliance requirements, and communication standards.
- **Conversation Flow Configuration:** Settings control pacing, interruption sensitivity, silence handling, and response timing, allowing agents to be tuned for different environments such as contact centres, helplines, or asynchronous digital channels.

3.3 Language Model Integration

The platform supports flexible language model integration while maintaining a consistent orchestration and governance layer.

Customers may:

- Use ElevenLabs-hosted language models managed as part of the service
- Integrate customer-approved third-party models
- Connect to self-hosted or orchestrated models through supported interfaces

Model selection, prompt configuration, and usage controls are managed at the agent level. ElevenLabs does not require customers to retrain or fine-tune language models to use the platform.

3.4 Voice and Language Capabilities

The Agents Platform provides extensive voice and language configuration options:

- **Voice Selection:** Access to a catalogue of professional voices
- **Multi-Voice Conversations:** Use different voices within a single agent or workflow
- **Language Support:** Deploy agents across multiple languages with automatic language detection and mid-conversation language switching
- **Voice Customisation:** Configure or create voices to meet service requirements

3.5 Knowledge Integration

Agents can be connected to approved knowledge sources to ground responses in authoritative content.

- **Knowledge Sources:** Uploaded documents or structured content approved by the customer
- **Retrieval-Augmented Generation (RAG):** A native RAG capability retrieves relevant information at runtime and incorporates it into agent responses
- **Low-Latency Design:** The RAG implementation is optimised for real-time conversational use cases
- **Independent Updates:** Knowledge bases can be updated without redeploying agents, allowing policy or service changes to be reflected immediately

3.6 Tools and System Integration

The platform enables agents to invoke tools and integrations to perform actions beyond conversation.

Supported integrations may include:

- API calls to external systems
- Retrieval or update of records
- Booking, notification, or status-check workflows
- Built-in system actions such as call transfer or voicemail detection
- **Native telephony integrations**, including SIP trunking
- **Messaging channel integrations**, such as WhatsApp

Tool access is explicitly scoped per agent or workflow step, supporting separation of duties and controlled integration with sensitive systems.

3.7 Personalisation and Context Injection

The service supports runtime personalisation of conversations.

- **Dynamic Variables:** Inject contextual data such as case IDs, user attributes, or session metadata
- **Per-Interaction Overrides:** Adjust agent behaviour or responses for specific interactions
- **Authentication Hooks:** Integrate custom authentication flows to control access or retrieve user-specific context

Personalisation mechanisms are designed to operate within customer-defined privacy and data-handling requirements.

3.8 Agent Testing, Analysis, and Operations

The platform includes capabilities to support safe deployment and continuous improvement.

- **Agent Testing:** Validate workflows, simulations, prompts, and integrations prior to deployment
- **Conversation Analysis:** Review interaction outcomes and performance metrics
- **Operational Dashboards:** Monitor usage, agent activity, and high-level performance indicators

These capabilities support quality assurance, service optimisation, and governance without exposing underlying infrastructure or model internals.

3.9 Service Scope and Responsibility Boundaries

The Agents Platform provides the conversational AI layer and real-time orchestration. Customers remain responsible for:

- Telephony carriers and SIP trunking
- Contact-centre platforms and workforce management
- External system availability and data correctness

The service is designed to integrate into existing government environments rather than replace them.

4. Onboarding, Support and Service Levels

4.1 Onboarding

The ElevenLabs Agents Platform is delivered as a managed, cloud-based software-as-a-service (SaaS). No software installation or local infrastructure is required.

All customers receive access to standard onboarding resources, including:

- Online product documentation and API references
- In-product guidance for agent configuration and deployment
- Published onboarding guides and examples

4.2 Support and After-Sales Service

Support and maintenance are provided for the duration of the Call-Off Contract.

Standard Support (Included)

Standard support is included as part of the core service offering and provides:

- Access to technical support for service-related issues and usage queries
- Email- or ticket-based support channels
- Support availability during published support hours, excluding public holidays
- Incident investigation, fault resolution, and guidance on the use of existing service functionality

Optional Enhanced Support Services

In addition to standard support, ElevenLabs offers enhanced customer success and support services, which may be procured as an optional add-on. These services are not required to use the core service and do not alter the service availability commitments or service credit remedies applicable under the Service Level Agreement. Where purchased, enhanced support services may include defined support response time targets and additional engagement channels, as described in the Pricing Document.

4.3 Implementation and Engineering Services (Forward Deployed Engineering)

ElevenLabs offers optional Implementation and Engineering Services, also referred to as Forward Deployed Engineering (FDE), which may be procured as an Ancillary Service.

These services are designed to support organisations requiring hands-on technical assistance to design, integrate, deploy, test, and scale the ElevenLabs Agents Platform within complex or regulated environments. Engagements focus on secure integration, production readiness, and operational adoption rather than bespoke product development.

Typical activities within scope may include:

- Solution and deployment architecture design for conversational agents
- Secure integration with customer systems, APIs, and approved data sources
- Configuration and implementation of agents, workflows, and tools
- Capacity planning, concurrency modelling, and scalability validation
- Production readiness assessments and deployment support
- Performance monitoring, optimisation, and operational tuning
- Delivery planning, governance, and stakeholder coordination
- Knowledge transfer and enablement for customer engineering and operations teams
- Structured feedback to inform product roadmap and continuous improvement

Implementation and Engineering Services are delivered as time-bound engagements or ongoing implementation support. The specific scope, duration, and deliverables are agreed prior to commencement and governed by a separate Call-Off Contract.

4.4 Service Levels and Performance

The Agents Platform is subject to availability commitments set out in the applicable Service Level Agreement.

For the Agents Platform, ElevenLabs commits to 99.5% monthly availability, measured as the average success rate of valid agent requests across each minute in the calendar month. Availability is measured on a 24x7 basis and excludes scheduled maintenance and standard SLA exclusions.

The service is optimised for real-time conversational use cases. Performance characteristics may vary depending on configuration, network conditions, and usage patterns.

Full definitions, measurement methodology, exclusions, and remedies are set out in the Call-Off Contract and SLA.

4.5 Service Constraints

ElevenLabs Agents Platform is delivered as a standardised, multi-tenant SaaS service. Customers cannot obtain bespoke development or work-for-hire under this service. The service requires internet connectivity and supported clients for web or API access. Speech synthesis performance and latency may vary with configuration, model choice, input length, output format, and network conditions. Planned maintenance may occur to maintain security and stability, with disruption minimised where possible. ElevenLabs may limit or suspend access to protect service integrity or address security risks. Customers must not submit PHI without a valid agreement, or special-category data unless permitted by applicable law.

5. Hosting, Deployment and Data Residency

5.1 Hosting Model

The service operates on a one-to-many multi-tenant SaaS model with data segregated at the application layer.

- **Cloud Provider:** The service is hosted on Google Cloud Platform (GCP) in secure data centres.
- **Infrastructure:** ElevenLabs utilises technology service suppliers to provide hosting solutions and support technology, maintaining responsibility for supplier performance.

5.2 Data Residency

ElevenLabs provides options regarding the processing location of data:

- **Standard Residency:** By default, data is processed in US-based regions.
- **EU Residency:** EU-based environments are available for buyers requiring specific residency controls. For customers requiring EU data residency, service user data is stored and processed within the European Economic Area (EEA).
- **Subprocessors:** A list of subprocessors is available via the ElevenLabs compliance portal.

6. Security & Privacy

6.1 Security Governance

ElevenLabs operates a formal information security and privacy management framework designed to protect the confidentiality, integrity, and availability of Customer Content. Security governance includes defined ownership, reporting structures, and regular review of policies and controls, which are documented, maintained and made publicly available at <https://compliance.elevenlabs.io>. All personnel are required to formally acknowledge relevant security policies as part of onboarding and on an ongoing basis.

6.2 Information Security Controls

The service is protected by technical and organisational measures covering:

- Access control and authentication
- Data protection and encryption
- Logging, monitoring, and audit
- Secure software development practices
- Change and configuration management
- Vendor and asset management
- Privacy and data protection compliance
- AI governance and risk management

Security controls are reviewed and updated to address changes in risk, regulation, and technology.

6.3 Vulnerability Management

ElevenLabs operates a vulnerability management programme to identify, assess, and remediate security risks. This includes:

- Continuous assessment of potential threats using internal and external intelligence sources
- Regular vulnerability scanning
- Independent penetration testing conducted at least annually
- Risk-based prioritisation and remediation of findings

6.4 Protective Monitoring

The service is continuously monitored to detect security events, misuse, and operational anomalies. Monitoring includes:

- Detection of unusual access patterns or behaviour
- Automated and manual alerting mechanisms
- Investigation of potential security incidents
- Abuse and misuse detection

ElevenLabs may suspend or limit access to the service where necessary to protect platform integrity or address a security risk.

6.5 Incident Management

ElevenLabs maintains predefined incident response processes for security and operational events. Customers may report incidents through established support channels. Where applicable, incident notifications and reports are provided in accordance with contractual and regulatory requirements.

6.6 Access Control

Access to management interfaces and support systems is restricted using role-based access controls and authentication mechanisms. Access is limited to authorised personnel and subject to logging and monitoring. Customer data is accessed only where required to provide support.

6.7 Certifications and Assurance

The service is supported by recognised security and compliance certifications and attestations, including:

- ISO/IEC 27001
- SOC 2 Type II
- Cyber Essentials and Cyber Essentials Plus

- PCI DSS
- CSA STAR (Level 1)
- Additional privacy and AI governance standards

These certifications are maintained through ongoing audit and assurance activities.

7. Backup, Restore and Disaster Recovery

7.1 Backup and Restore

- **Automated Backups:** ElevenLabs performs regular automated backups of platform configuration and retained data as required for the operation of the SaaS service.
- **Data Integrity:** Measures are in place to protect the security and integrity of Customer Content collected, accessed, stored, or received.

7.2 Disaster Recovery and Business Continuity

The platform is designed for high availability using cloud-native redundancy:

- **Resilience:** The solution is hosted in GCP data centres, leveraging availability zones to maintain service continuity.
- **Force Majeure:** The service includes provisions for events beyond reasonable control (e.g., internet failures, public health crises), during which performance obligations may be delayed.

8. Operations, Maintenance and Updates

8.1 Operational Monitoring and Observability

ElevenLabs operates continuous monitoring and observability for the Service to support reliable, secure, and compliant operation in production environments.

Operational monitoring may include:

- Service availability and performance indicators

- Request and processing metrics
- Error rates and failure conditions
- Abuse, misuse, and anomalous usage patterns

Monitoring is designed to support service assurance, incident response, and operational governance without exposing underlying infrastructure or proprietary implementation details.

8.2 Product-Specific Monitoring and Quality Evaluation

For the Agents Platform, monitoring and evaluation extend beyond system performance to include agent behaviour and outcomes.

This may include:

- Conversation transcripts and interaction metadata
- Agent decisions, workflow transitions, and tool invocations
- Task completion rates and escalation events
- Evaluation of agent behaviour against defined instructions and workflows

These capabilities support governance, quality assurance, and continuous optimisation of conversational agents deployed in live environments.

8.3 Logging, Data Collection and Audit Support

The Service provides mechanisms to collect and retain operational logs and results in support of audit, investigation, and service improvement activities.

Logged data may include:

- Request metadata and timestamps
- Processing outcomes and system-generated events
- Input and output artefacts, where applicable and enabled by the Customer

Customers may access available logs and results via APIs or web-based tools, subject to configured retention policies and applicable data protection requirements.

8.4 Service Updates and API Versioning

ElevenLabs modifies the Services, including features and functionality, on an ongoing basis as part of the SaaS delivery model.

- Updates are applied centrally and made available without local installation.
- Updates will not result in a material degradation of the overall functionality or performance of the Services used by the Customer.
- APIs are versioned where applicable.
- At least 90 days' prior notice is provided before deprecating older API versions.

8.5 Outage and Maintenance Management

Planned maintenance is periodically required to ensure the security, stability, and performance of the Service.

- ElevenLabs endeavours to minimise disruption and schedule maintenance outside peak usage where practical.
- Service availability and incident information are published via the public service status page.
- Incidents are classified and managed based on severity and impact.
- In exceptional circumstances, including high-risk security events, access to specific accounts or features may be temporarily restricted to protect platform integrity.

8.6 Continuous Improvement

Operational insights derived from monitoring, evaluation, and customer feedback are used to inform continuous improvement of the Service. Enhancements to reliability, performance, and quality are delivered through the standard update mechanisms described above.

9. Service Constraints

9.1 Usage Constraints

- **Prohibited Use:** Customers must comply with the ElevenLabs Prohibited Use Policy.
- **Sensitive Data:** Customers must not provide Input that includes Protected Health Information (PHI) unless an agreement is executed, or special category data (e.g., biometric for identification, political opinions) unless permitted by applicable law.
- **Reverse Engineering:** Customers may not reverse engineer, decompile, or attempt to discover source code or underlying components of the Services or Models.

9.2 Customisation

The service is a standardised SaaS offering. Custom development or "work-for-hire" is not available. Customisation is achieved through the configuration of Agents, Voices, and Workflows within the standard tooling.

10. Exit Management and Data Access

10.1 Data Ownership

- **Input Rights:** The Customer retains all rights in Input (data transmitted to the system).
- **Output Rights:** As between the Customer and ElevenLabs, the Customer owns all rights in Output (content generated by the Services). ElevenLabs does not train its models on Enterprise Customer Output, except as may be required to provide certain requested services to the Customer.

10.2 Service Offboarding and Data Export

Upon termination or expiry of the Call-Off Contract, or during the contract term, customers may retrieve their Input and Output data on a self-serve basis.

- **Export Mechanisms:** Data retrieval is supported via standard export mechanisms provided by the platform, including APIs and web-based tools.
- **Formats:** Data is typically available in standard formats (e.g., JSON for configuration/logs, Audio files for output).

- **Documentation:** ElevenLabs provides documentation describing available export formats and data access mechanisms. No bespoke migration, data transformation, or third-party integration services are included as part of offboarding.

10.3 Termination, Deletion and Wind-Down

- **Cessation of Use:** Upon termination, the Customer must immediately cease all use of the Services.
- **Data Deletion:** Customer Content is deleted from the service within thirty (30) days of contract termination or expiry, unless the Customer explicitly agrees to retention beyond this period. Data deletion processes are designed to ensure secure and irreversible removal.
- **Wind-Down:** Specific wind-down periods may apply for partners/resellers distributing the service to End Users, ensuring a managed exit up to one year from expiration.
- **Sanitisation:** Industry-standard data sanitisation processes are applied to storage media, including secure erasure or physical destruction where applicable.

11. Technical Requirements

To access and use the Agents Platform, the Customer is responsible for obtaining and maintaining the following:

- **Equipment:** Hardware, servers, and operating systems necessary to connect to the cloud service.
- **Network:** Internet connectivity and networking capability.
- **Web Access:** A modern web browser for the administration dashboard.
- **API Integration:** Capabilities to integrate via REST API for programmatic access (if required).