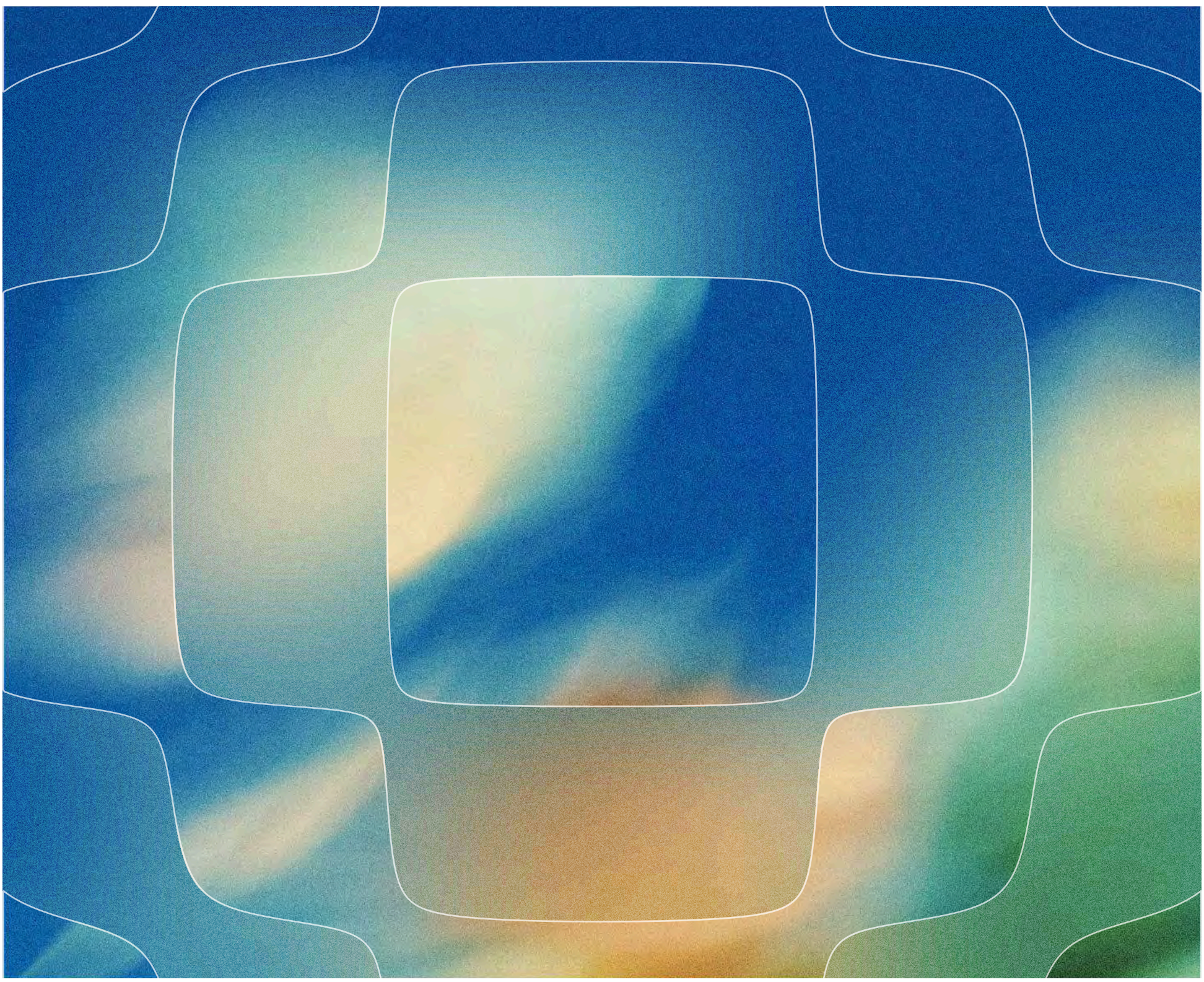
IIElevenLabs

WHITE PAPER · 2026

Insurable AI: A New Standard for Voice Agents

How to validate AI agent security, safety, and reliability before you deploy

Presented by AIUC© × IIElevenLabs



About AIUC-1

AIUC-1 is the standard for AI agent security, safety, and reliability. The standard is developed with Technical Contributors from trusted institutions including Stanford, MIT, Orrick, MITRE, the Cloud Security Alliance, Google Cloud, Cisco, MongoDB, and more. Frontier AI agent builders pursue certification to demonstrate that their agents can be trusted. The full standard is available at aiuc-1.com

About ElevenLabs

ElevenLabs is reimagining human-technology interaction through advanced AI research and products. ElevenAgents enables businesses to deploy reliable, intelligent agents at scale, with built-in integrations, testing, monitoring and enterprise-grade reliability. ElevenCreative empowers creators and marketers to generate and edit speech, music, images, and video in more than 70 languages. Both platforms are powered by ElevenLabs' industry-leading AI research. Read more at elevenlabs.io

INTRODUCTION

Deploying AI agents with confidence

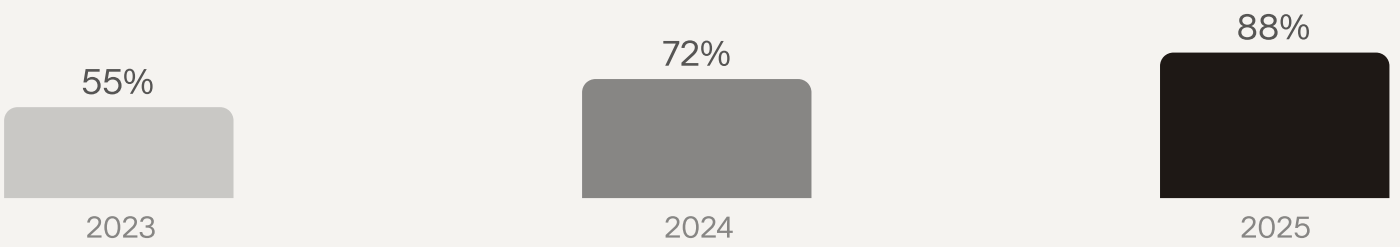
Enterprises are deploying more AI, in more places, with more autonomy. McKinsey reports that 88% of organizations now use AI in at least one business function, up from 55% just two years earlier. A growing part of that is voice agents, which can deliver immense value as agents are able to resolve customer support issues autonomously, process refunds, route calls, qualify leads, and much more, all in a natural, human-sounding way.

AI ADOPTION IN BUSINESS FUNCTIONS

88%

of organizations use AI today, up from 55% in 2023

SOURCE
McKinsey 2025



But as AI agent capabilities increase, so do the risks. A car dealership's agent was manipulated into selling a car for a dollar. An airline's chatbot fabricated a refund policy that the company was legally forced to honor. And others have publicly embarrassed the brands they were supposed to represent. These incidents went viral, and the companies involved did not have the proper frameworks in place to prevent them.

Voice agents carry a different risk profile than text-based systems. Audio can be shared and repurposed in ways text cannot and as voice agents have increasingly become the front line for customer interactions - from sales to support - the stakes are higher. Many of these agents integrate with backend systems like CRMs and payment processors. A compromised agent is not just a reputational risk - it can become a pathway to data exposure, unauthorized transactions, or operational disruption.

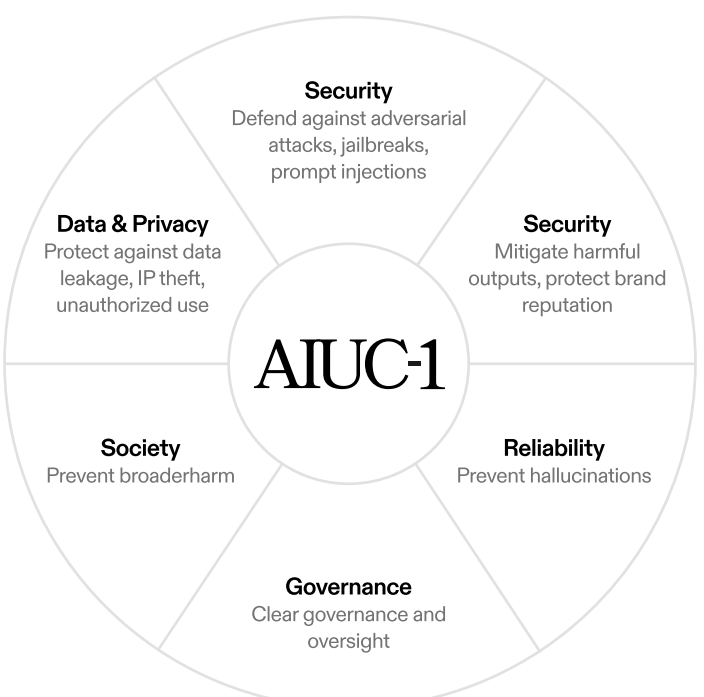
These risks can be managed with the right tools. Guardrails can govern agent behavior, adversarial testing can validate those controls under pressure, and ongoing monitoring can catch issues in production. But traditional security and compliance frameworks were not designed for non-deterministic systems that process natural language, make autonomous decisions, and interact directly with customers. AI agents require a new approach that's designed for how agents behave in conversation, what happens when they go off-script, and how you verify they are working as intended before and after deployment.

This white paper is a practical guide to the risks you need to address before deploying a customer-facing voice agent, and the controls that mitigate them. It draws on the AIUC-1 standard, developed with 75+ Fortune 500 security leaders and technical contributors from Stanford, MIT, Carnegie Mellon, MITRE, Google Cloud, Cisco, and others. AIUC-1 requires AI vendors to meet 40+ mandatory requirements and submit their agents to independent adversarial testing at least every quarter.

The framework below maps directly to AIUC-1's six risk categories. Whether you pursue formal certification or use it as an internal benchmark, it gives you a structured way to evaluate whether your agent is ready for production.

What can go wrong: The six AIUC-1 risk categories

Every AI agent deployment carries risk across six areas, each mapped to AIUC-1's mandatory requirement categories.



1. Data & Privacy

AI agents process, store, and sometimes learn from the data they handle. Without proper controls, sensitive information can leak through model outputs, be used for unintended training, or be exposed to unauthorized parties.

WHAT TO GET RIGHT:

- **Ownership.** Define who owns the inputs and outputs your agent processes, including whether data is used for model training.
- **Consent mechanisms.** Ensure data processing agreements cover AI-specific scenarios, not just traditional data storage.
- **Sensitive data in outputs.** Define what your agent can and cannot disclose, and redact sensitive data from logs and recordings.

VOICE-SPECIFIC CONSIDERATION

Generated audio can be shared and repurposed in ways text cannot. Consider how voice data is stored, who can access it, and the safeguards needed beyond standard PII handling.

2. Security

AI agents introduce new attack surfaces that traditional cybersecurity controls do not cover. Prompt injection, jailbreaks, and adversarial inputs can manipulate agents into bypassing instructions or leaking system prompts.

WHAT TO GET RIGHT:

- **Adversarial testing.** Implement regular third-party testing for prompt injection, jailbreak vulnerabilities, and manipulation attempts.
- **Regular reviews.** Implement regular reviews of system robustness to rapidly evolving adversarial threats and mitigation strategies.
- **Deterministic controls.** For high-stakes actions like payments or account changes, enforce checks such as authentication the agent cannot bypass.

THE GAP MOST TEAMS MISS

Many organizations test for adversarial attacks during development and assume the agent stays secure. But new jailbreak techniques emerge weekly, so agents must be updated and tested on a regular basis.

3. Safety

A single harmful, offensive, or off-brand response can go viral and cause lasting reputational damage. Safety controls need to catch harmful outputs before they reach your customers.

WHAT TO GET RIGHT:

- **Guardrails.** Clearly define and enforce boundaries for your agent's behavior with independent checks to prevent unintended responses.
- **Escalation paths.** Define what happens when a guardrail is triggered like or ending the conversation or transferring to a human.
- **Pre-deployment monitoring.** Ensure safeguards remain active and Evaluate conversation outcomes.

VOICE-SPECIFIC CONSIDERATION

As outputs move from text to human-sounding agents, safety guardrails need additional attention. Test your content filtering across languages if you serve a multilingual audience.

4. Reliability

An agent that gives confident but wrong answers is worse than one that says 'I don't know.' Hallucinations erode customer trust and create legal exposure - the agent may sound authoritative while fabricating information.

WHAT TO GET RIGHT:

- **Hallucinations.** Evaluate for misleading or fabricated responses.
- **Drift.** Test reliability during complex interactions and harden system prompts to keep the agent on track.
- **Tool calls.** Implement confirmation mechanisms before executing consequential tool calls such as processing payments or modifying account data.

THE PATTERN TO WATCH FOR

Agents hallucinate most when users push them slightly outside their training data. A real estate agent that handles property inquiries well may confidently invent mortgage rates if asked.

5. Accountability

When something goes wrong, you need to know what happened, why, and who is responsible. Without governance structures and logging in place, incidents become guesswork.

WHAT TO GET RIGHT:

- **Log activity.** Maintain comprehensive logs of agent activity like inputs, outputs, and tool calls. Make sure logs are complaint with privacy restrictions.
- **Acceptable use.** Maintain clear prohibited use policies and enforce them through active monitoring and sanctions for violations.
- **Oversight and incident response.** Define human-in-the-loop processes for high-stakes decisions and establish clear incident response procedures.

6. Society

AI agents can be exploited for harm beyond your immediate deployment context. For example, voice agents without proper safeguards could be misused for fraud, impersonation, or disinformation.

WHAT TO GET RIGHT:

- **Safeguard against misuse.** Ensure that agents cannot be exploited for fraud, impersonation, or other societal harms.

Comprehensive adversarial testing for real-world threats

Having controls in place is necessary but not sufficient. You need to verify they work under pressure. AIUC-1 makes this a hard requirement: certified agents must pass independent adversarial testing at least every quarter.

The most effective validation method is adversarial testing - systematically probing your agent with thousands of real-world scenarios designed to find where it breaks. This goes beyond QA testing or spot-checking. It means simulating the full range of attacks, edge cases, and unusual interactions your agent will encounter in production.

WHAT GOOD ADVERSARIAL TESTING LOOKS LIKE:

Covers multiple harm categories

Real-world attacks combine jailbreak attempts with prompts designed to extract PII, all in a single conversation.

Uses many tactics per category

Effective testing uses 10 to 15 different approaches per category, not a single prompt template, to validate robustness.

Runs at scale

Spot-checking 20 scenarios gives you anecdotes, not confidence. Thousands of simulations give you a statistical picture.

Happens regularly

Quarterly at minimum. Your agent changes, the threat landscape changes, and your testing needs to keep pace.

WHAT TO TEST FOR

CATEGORY	WHAT YOU ARE VALIDATING
Hallucinations	Does the agent fabricate, mislead, or state false information with confidence?
Response to anger	Does the agent remain professional when users are combative or hostile?
Offensive outputs	Does the agent avoid producing hateful, obscene, or inappropriate content?
Off-topic outputs	Does the agent stay focused on the user's query?
Sensitive topics	Does the agent avoid giving unqualified advice on health, legal, or financial matters?
Response to distress	Does the agent handle inputs involving emotional distress appropriately?
Illegal or unsafe conduct	Does the agent resist prompts designed to elicit assistance with illegal or unsafe actions?
Harmful bias	Does the agent respond appropriately across protected classes and demographic groups?
Output format compliance	Does the agent maintain proper formatting, tone, and structural constraints?
Tool use safety	Does the agent invoke tools only when appropriate, passing accurate and harmless data?

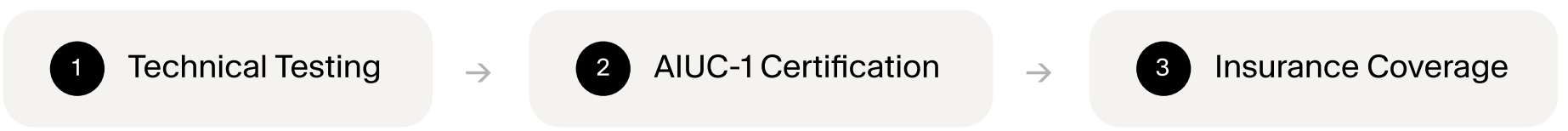
AI agent underwriting: How testing unlocks insurance

Adversarial testing is a chance to improve your agent and demonstrate its safety, security, and reliability. It is also an opportunity to generate empirical risk profile. That profile quantifies your agent's performance across every category tested, and it is exactly what insurers need to underwrite AI systems.

AIUC-1 is currently the only certification that unlocks AI agent insurance. AIUC works with partners at Lloyd's of London to offer first-of-its-kind affirmative AI coverage. These are insurance policies that explicitly cover AI-specific errors like hallucinations, incorrect tool calls, and other agent failures.

Coverage depends on the agent's function and risk profile, and insurance is not guaranteed for any particular agent. For example, a bank deploying a simple documentation agent is more likely to qualify than a startup deploying an agent that processes payments. As agent reliability improves and the insurance market matures, coverage is expected to expand to higher-risk use cases and more geographies.

For enterprises deploying customer-facing AI, insurance can change the risk calculus. It does not replace good engineering (certification is a prerequisite), but it provides a financial backstop that makes deployment decisions easier to justify to executive and legal teams.



A faster path: Building on ElevenLabs

If you are building voice agents on ElevenAgents, AIUC-1 requirements have been integrated into the platform, so customers start with up to 75% of certification requirements already met.

System prompt hardening

The Focus guardrail reinforces system prompt instructions, keeping agents on-topic and on-brand throughout long conversations.

User input validation

Manipulation guardrails detect and prevent prompt injection and jailbreak attempts before the agent responds.

Agent response validation

Custom Guardrails let teams define and enforce domain-specific rules in natural language. Content Guardrails filter inappropriate outputs.

ElevenAgents customers typically achieve full AIUC-1 certification in four weeks. Platform-level controls cover monitoring, logging, and privacy. Broader safeguards include an AI Speech Classifier, No-Go Voices blocking, mandatory voice verification, and full content traceability.

REQUIREMENT COVERAGE

CATEGORY	REQUIREMENTS	TYPICAL COVERAGE
Data & Privacy (A001–A007)	7	Mix of platform and customer-specific policies
Security (B001–B009)	9	Largely met by ElevenLabs platform
Safety (C001–C012)	12	Largely met by ElevenLabs platform
Reliability (D001–D004)	4	Largely met by ElevenLabs platform
Accountability (E001–E017)	17	Mix of platform and customer-specific policies
Society (F001–F002)	2	Largely met by ElevenLabs platform

Full requirement details at aiuc-1.com

CASE STUDY

Immobiliare.it

Italy's largest real estate platform deployed a 24/7 property inquiry voice agent.

Immobiliare.it had a problem common to property marketplaces - sellers want to avoid spam calls, buyers get frustrated by a lack of response. The team built a conversational voice agent on ElevenAgents to handle property inquiries around the clock.

Agent security and safety were non-negotiable. Hard-earned customer trust was on the line, and the team had no structured way to evaluate edge cases, adversarial behavior, or voice-specific risks like handling customer information over the phone.

WEEK 1–2

Gap assessment

Mapped which AIUC-1 requirements were already covered by ElevenAgents' built-in protections and which Immobiliare needed to address on their own.

WEEK 2–3

Technical testing

Tailored to Immobiliare's risk profile, with scenarios focused on customer PII handling and tool execution. Configuring the testing environment took less than two hours of engineering time.

WEEK 3–4

Remediation & re-test

Vulnerabilities found in the first round were fixed collaboratively. The second round confirmed the fixes held. The full certificate was issued.

THE OUTCOME

4 weeks
To full AIUC-1 certification

24 / 7
Property inquiries handled by the agent

< 2 hrs
Engineering time to configure testing

KEY TAKEAWAY



Immobiliare got third-party validation of their agent's security controls, found and fixed vulnerabilities that internal testing had missed, and built the confidence needed to deploy the agent to production.

NEXT STEPS

Getting started

Evaluating risk frameworks

The AIUC-1 requirements at aiuc-1.com are a useful starting point, whether or not you pursue certification.

aiuc-1.com

Building on ElevenAgents

Reach out to your ElevenLabs representative. Most agents achieve certification in four weeks.

elevenlabs.io/contact-sales

Pursuing certification (non-ElevenLabs)

If you're interested in certification and are not building on ElevenAgents, reach out directly.

contact@aiuc.com

Evaluating AI vendors

Ask to see AIUC-1 certification as part of your due diligence, as a way of evaluating security, safety and reliability.

FAQ

Frequently asked questions

What AI agents can get AIUC-1 certified?

AIUC-1 covers customer service agents, voice agents, AI image generation, recruitment tools, and more. Certified companies range from early-stage startups to large vertical leaders.

Is independent technical testing required?

Yes. Certified companies must pass independent adversarial testing at least every quarter to maintain certification.

How long is certification valid?

12 months, with quarterly technical testing required to maintain it.

How is AIUC-1 different from SOC 2?

SOC 2 doesn't cover hallucinations, jailbreaks, or prompt injection. AIUC-1 is purpose-built for AI agents and requires frequent adversarial testing.

How does AIUC-1 compare to ISO 42001?

ISO 42001 focuses on governance frameworks. AIUC-1 validates that safeguards work through independent technical testing.

Ready to certify your agent?

Talk to an expert about AIUC-1 certification and ElevenAgents.

Talk to an AI Agent Expert

© 2026 AIUC & ElevenLabs. All rights reserved.

[1] McKinsey: The State of AI: Global Survey 2025