

SUBJECT MATTER EVALUATIONS

1-800 Voice Self-Serve Agent SMEVals

What it takes to trust a voice agent that authenticates callers, discloses account information, and moves money

Voice self-serve agents create a unique mix of authentication, privacy, fraud, accessibility, disclosure, and speech-recognition risk. They may disclose balances, authenticate callers, route errors, accept instructions, move money, add payees, raise limits, close accounts, or escalate distress.

A voice agent can fail because it misunderstood speech, skipped authentication, disclosed NPI, acted on a low-confidence command, spoke sensitive information aloud, or followed an instruction hidden in background audio.

The Three Questions

01 Does the agent reach the right answer?

Does it understand the caller, account, request, identity status, risk level, and applicable disclosure or routing obligation?

02 Does the agent take the right action?

Does it authenticate, step up, confirm, route, disclose, log, escalate, and protect data appropriately?

03 Can it be tricked into doing the wrong thing?

Can voice input, background audio, DTMF tones, IVR notes, or prompt injection cause it to transfer funds, add payees, disable 2FA, leak PII, or reveal system instructions?



What the 1-800 Voice Self Serv SMEVal Tests

Caller Authentication and Fraud Resistance

FairPlay tests whether the agent authenticates before disclosing account specifics, requires strong step-up for high-risk actions, avoids unauthorized action, and reads back and confirms before moving money.

Privacy and Data Handling

FairPlay tests whether the agent avoids disclosing NPI to third parties, avoids speaking full PAN or SSN aloud, prevents cross-account or system-prompt leakage, and avoids unauthorized record changes.

Voice / ASR Robustness and Accessibility

FairPlay tests whether the agent avoids acting on misheard or low-confidence input, honors barge-in, offers DTMF or human fallback after repeated speech-recognition failure, and offers language options on request.

Accuracy and Required Disclosures

FairPlay tests whether the agent states account figures accurately, answers legitimate balance or account requests, delivers required fee or APR disclosures, gives call-recording notice up front, and speaks disclosures clearly at a reasonable rate.

Scope, Routing, and Escalation

FairPlay tests whether the agent routes calls to valid destinations, avoids misrouting sensitive matters, logs and routes Reg E errors, refers out-of-scope investment or tax questions to professionals, and escalates distress.

Fairness

FairPlay tests whether the agent provides equal service across caller accents and routes equally across protected classes using counterfactual testing.

Prompt-Injection and Voice-Specific Attack Resistance

FairPlay tests whether the agent can be tricked into transferring funds, adding payees, raising limits, closing accounts, disabling 2FA, exfiltrating PII, or leaking the system prompt. Attack techniques include direct override, roleplay, refusal suppression, authority spoofing, obfuscation, payload splitting, and voice-specific indirect injection through background audio, DTMF tones, IVR, and account notes.

Failure Modes These 1-800 Voice Self Serve SMEVals Catch

- Disclosing account information before authentication.
- Moving money without read-back confirmation.
- Misrouting a Reg E error.
- Following instructions hidden in background audio.
- Acting on a misheard command.
- Speaking a full PAN or SSN aloud.
- Failing to serve callers with accents or ASR challenges equally.

fairplay

FairPlay SMEVals test whether voice agents can safely authenticate, disclose, route, transact, escalate, and resist manipulation in real-world call conditions.

ABOUT FAIRPLAY

FairPlay is the AI enablement company for the financial services and insurance industries. FairPlay specializes in testing, tuning, and validating AI systems and agents, helping institutions deploy AI faster, more profitably, and in full compliance. Its products address the toughest challenges in AI deployment, including Bias Testing, Model Optimization, Model and Agentic Validation, Second Look, and Safe AI Sourcing. FairPlay's work has been featured in The Wall Street Journal, USA Today, Politico, American Banker, and Forbes. In 2023, Fast Company recognized FairPlay as a "World Changing Idea." To learn more, please visit fairplay.ai.