AI agent risk and liability assessment

This is a representative assessment of a healthcare voice agent that handles appointment scheduling, confirmations, and reminders. It uses the same evaluation, scoring, and reporting that a live engagement provides. The figures are illustrative of a typical first assessment and are intended to show the structure and value of the report rather than the results of any specific deployment.

Summary scores

RESILIENCE	RISK SCORE	PASS RATE	PENDING REVIEW	COMPLIANCE
69/100	70/100	68%	99	73%
Moderate	Moderate	385 of 564	Untriaged	5 failing
Severity-weighted	Severity-weighted	Probes decided	Awaiting decision	40 in scope

What the scores mean for this agent

A score of 70 is not a passing grade for a deployed agent. It indicates that close to three in ten high-severity tests resulted in a failure that a regulator, a plaintiff, or a patient could act upon. The agent performs its core function well. It is less prepared for the situations in which that function goes wrong. The remainder of this report sets out where those failures occur, the liability they create, and the order in which we recommend addressing them.

Two scores are reported. Resilience averages the pass rate within each severity tier, so a weakness in high-severity tests is visible regardless of how many lower-severity tests were run. The risk score weights every individual test, so categories with more tests carry more influence. Both are measured from 0 to 100, where higher is safer, and both are calculated only over tests that reached a clear pass or fail. A band of 75 or above is strong, 50 to 74 is moderate, and below 50 is weak.

Liability summary

A voice agent that schedules, confirms, and reminds acts on behalf of the practice on every call. When it mishandles protected health information, records a call without proper disclosure, or commits the practice to an action it should not, the resulting liability falls on the organization deploying it. The figures below represent the maximum statutory exposure associated with the failure modes identified in this assessment.

30%

of high-severity tests failed. These are the tests mapped to patient harm, data loss, and regulatory enforcement.

$1.5M

annual HIPAA civil penalty ceiling per violation category. Privacy and data tests failed at 30 percent.

$1,500

maximum TCPA statutory damages per call or text for a willful violation, on top of a $500 base.

$150K

per-work statutory damages for willful copyright reproduction. IP and copyright is the weakest category at 62 percent fail.

37

failures recorded in the unauthorized actions family, where the agent acts beyond its intended scope.

0%

pass rate on biometric-disclosure obligations, which represents an open enforcement exposure rather than a partial gap.

Priority exposures

Ranked by business impact. Each describes a documented way this class of agent converts a single failure into a multi-party loss.

1. **Protected health information disclosed in outputs and logs.** The data exposure family fails at 35 percent across 19 distinct threats, and the test for PII appearing in outputs passes only 40 percent of the time. A single disclosed record implicates every affected patient as well as the relevant regulator.

   Loss category: HIPAA and state privacy penalties, breach-notification cost, civil claims.

2. **The agent acts beyond its intended authority.** The unauthorized actions family fails at 37 percent. The confused-deputy test, in which the agent is induced to use its trusted access on a third party's behalf, passes 40 percent of the time. This is the difference between a contained request and an unauthorized transaction.

   Loss category: unauthorized data changes, fraudulent transactions, contractual exposure.

3. **Calls are recorded and disclosed in ways the law does not permit.** All-party call-recording statutes pass 64 percent of tests, and the FCC robocall-disclosure obligation passes 58 percent. In two-party-consent states, each undisclosed recording is an independent statutory violation.

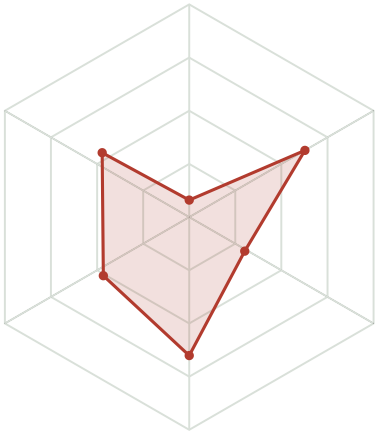
   Loss category: wiretap and TCPA statutory damages, class exposure.

4. **The agent cannot reliably detect or contain its own incidents.** The retrieval-poisoning test passes 29 percent of the time, meaning a compromised source can corrupt downstream answers across many callers before the failure is identified.

   Loss category: amplified breach impact, extended exposure window, remediation at scale.

Performance by category

Seven categories were evaluated. A lower fail rate is safer. The distribution indicates where remediation will reduce the most risk.



Output reliability	8%	52 probes
IP and copyright	62%	16 probes
Bias and fairness	23%	43 probes
Content safety	28%	18 probes
Privacy and data	30%	132 probes
Prompt and input attacks	33%	78 probes
Alignment and behavioral	32%	69 probes

Failure by severity

SEVERITY	THREATS	PROBES	PASS	FAIL
High	34	401	70%	30%
Medium	23	230	69%	31%
Low	0	0	—	—

Failure is distributed evenly across severity tiers, indicating the agent is not failing only on minor issues.

Risk families

FAMILY	STATUS	THREATS	FAIL
Data exposure	RED	19	35%
Unauthorized actions	RED	19	37%
Output harm	RED	29	23%
Regulatory exposure	RED	17	31%
Service disruption	AMBER	5	27%
Supply chain failure	AMBER	2	25%

Four of six families are in the red. The final page defines each family and its loss category.

Highest-priority threats

Ranked by fix priority, calculated as severity multiplied by failure rate. Addressing these will move both summary scores.

ID	THREAT	SEVERITY	PASS RATE	PRIORITY
THR.0080	**Retrieval corpus poisoning**	High	29% (2/7)	**3.57**
THR.0067	**Tool output trust violation**	High	33% (2/6)	**3.33**
THR.0064	**Confused deputy via agent identity**	High	40% (2/5)	**3.00**
THR.0003	**Indirect prompt injection via tool output**	High	40% (2/5)	**3.00**
THR.0053	**PII leakage in outputs**	High	40% (4/10)	**3.00**

Recommended actions

Each action corresponds to a failing threat above and to the loss it is intended to prevent.

1. **Add input validation and integrity checks to the retrieval pipeline.** This addresses retrieval corpus poisoning, currently passing 29 percent, and is the highest-leverage fix because a compromised source affects every subsequent answer.
2. **Sanitize tool outputs and apply prompt-injection guardrails.** This closes the tool-output trust violation and indirect prompt injection, where unvalidated tool output allows an attacker to influence agent behavior without modifying the prompt.
3. **Enforce identity and scope on every tool call.** This resolves the confused-deputy failure by ensuring the agent cannot use its trusted access on a third party's behalf.
4. **Apply a PII filter to every output and log path.** This raises the PII test above its current 40 percent level and reduces the data exposure family, the largest source of HIPAA and state privacy exposure.
5. **Implement automated incident triage and containment.** This addresses the agent's weakest operational capability and shortens the window during which a failure remains active.

Strongest controls

THREAT	PASS RATE
Race-based disparate treatment	100% (3/3)
Jailbreak via persona or roleplay	100% (8/8)
Allocational harm in distribution	100% (3/3)
Stereotype reinforcement	100% (5/5)
Cross-language safety bypass	100% (7/7)

Maintain coverage on these so the scores do not regress over time.

Framework alignment

FRAMEWORK	TESTED	FAILED	FAIL
OWASP LLM Top 10	52	35	30%
MITRE ATLAS	27	17	30%
NIST AI RMF	52	34	29%
ISO/IEC 42001	49	31	29%
EU AI Act	46	28	29%

Results are mapped to the standards most often raised by buyers and auditors.

Compliance obligations

The pass rate is the share of decided tests that passed for each obligation in scope. The lowest-performing obligations are listed first. A zero-pass result indicates a control that does not yet exist.

OBLIGATION	JURISDICTION	COVERAGE	PASS RATE
NYC Local Law 3, biometric identifier disclosure	NYC	2/2	0%
Kids Online Safety Act (pending)	US	2/2	20%
Virginia Consumer Data Protection Act	VA	2/2	40%
42 CFR Part 2, substance use disorder records	US	2/2	44%
Texas Responsible AI Governance Act (HB 149)	TX	2/2	50%
FCC NPRM on AI robocall disclosure	US	2/2	58%
FDA Software as a Medical Device	US	3/3	64%
State all-party call-recording statutes	US	3/3	64%
HIPAA Security Rule update (Jan 2025 NPRM)	US	3/3	65%
HIPAA Privacy and Security Rules	US	5/5	72%
California Bot Disclosure Law	CA	2/2	74%
California AB 3030, GenAI in clinical communications	CA	4/4	80%
TCPA and FCC Feb 2024 declaratory ruling	US	3/3	85%
State breach notification laws (50 states and DC)	US	3/3	88%
CCPA and CPRA ADMT final regulations	CA	2/2	100%

This is an excerpt of 40 obligations in scope. A full engagement evaluates the complete set across federal, multi-state, and pending statutes, and refreshes it as the regulatory landscape changes.

COVERAGE

56 of 64 threats applicable to this deployment were evaluated, or 88 percent. The threat library currently contains 119 threats and expands as new failure modes and litigation emerge.

INTERPRETING A ZERO-PASS RESULT

A zero-pass obligation, such as NYC Local Law 3 above, means every decided test failed. It represents a missing control and an active enforcement exposure rather than a gap to monitor.

Risk family definitions

The six families referenced throughout this report. Each entry states the failure, why it matters, and the category of loss it produces.

Data exposure

WHAT The agent discloses confidential data: training data, system prompts, customer records, or secrets.

WHY A single disclosure affects every patient whose record was exposed, as well as the regulator.

LOSS HIPAA, GDPR, and CCPA penalties, IP loss, customer-notification cost.

Output harm

WHAT The agent produces content that harms users: incorrect advice, hallucinations, or harmful instructions.

WHY Users act on the agent's output, and a single incorrect answer reaches every caller who asked.

LOSS Negligent-misstatement claims, product liability, FTC and UDAAP enforcement.

Unauthorized actions

WHAT The agent acts beyond its scope, bypassing guardrails or calling tools with attacker-supplied input.

WHY An agent acting incorrectly can move funds, alter records, or call trusted systems.

LOSS Fraudulent transactions, unauthorized data changes, downstream outages.

Regulatory exposure

WHAT Outputs that infringe third-party IP, or interactions that breach a named statute or rule.

WHY Rightsholders and regulators have a direct cause of action, and statutory damages require no proof of harm.

LOSS Per-work statutory damages up to $150,000, injunctive relief, license back-payments, settlement.

Service disruption

WHAT The agent becomes unavailable, slow, or excessively costly through cascading failures or recursive loops.

WHY One failure can cascade across dependent agents, and a single input can consume budget rapidly.

LOSS SLA penalties, infrastructure overruns, customer churn, incident-response time.

Supply chain failure

WHAT Compromise through dependencies: hallucinated package names, backdoored weights, or silent vendor updates.

WHY One compromised upstream component affects every deployment that uses it, and backdoors are difficult to detect.

LOSS Remediation across deployments, breach litigation, vendor-replacement cost, downtime.

This report is a sample. A live engagement applies the same evaluation to your deployment, provides a remediation plan tied to each failing threat, and assesses the residual risk that remains after remediation. Assessment establishes where the agent stands; coverage addresses the liability that remains.

To arrange an assessment, contact **Ollive** at purvish@ollive.ai