Adversary

AI agent red-teaming as a service — jailbreak, prompt-injection, and data-exfiltration testing for every company shipping a customer-facing AI agent without the in-house adversarial team to test it.

Category	Set 3 · Post-AI Plays
Customer	US/EU/Asia AI startups and mid-size enterprises deploying customer-facing AI agents (support, sales, internal tools, voice agents)
Monetisation	$8k–40k per engagement · $5k/mo continuous testing retainer · $25k–80k annual enterprise contracts
Build effort	Med
Plan version	v1.0 — 2026-05

Executive Summary

Adversary is an AI agent security-testing service for companies shipping LLM-powered products without the in-house adversarial expertise to test them. The market emerged sharply in 2024-25 as enterprise AI deployments scaled and the consequence of agent failures became visible — prompt-injection attacks exfiltrating customer data, jailbroken support agents promising refunds the company never authorised, voice agents tricked into ordering inventory, internal copilots induced to leak training data. By 2026 every serious enterprise AI deployment needs adversarial testing, and almost no company has built the in-house team to do it.

Adversary operates as a productised consulting service. A typical engagement is 3-6 weeks: scoping (1 week), structured adversarial testing across 14 attack categories (2-4 weeks), and a written report with reproducible attack scripts, severity-ranked findings, and remediation guidance (1 week). Pricing is $8k-40k per engagement depending on agent complexity and surface area. Continuous testing retainers ($5k/month) provide ongoing monthly testing as the agent evolves. Annual enterprise contracts ($25k-80k) provide quarterly engagement + on-call response.

Year-1 target: 80 engagements + 25 active retainers, generating $1.8 million revenue (■15 crore) against ■5.2 crore in costs. The wedge is India-staffed delivery: a team of 12-20 trained AI-security testers in Bengaluru/Pune delivering work at quality competitive with US firms (Robust Intelligence, HiddenLayer, Lakera) at 40-55% lower cost. The operational moat compounds as the team accumulates an internal library of 2,000+ attack patterns across customer engagements.

The Problem

Every company shipping a customer-facing AI agent in 2026 faces an adversarial testing gap. The agent is built by ML engineers and product engineers who think about happy-path conversations; the threats come from users actively trying to break the agent — prompt injection through user inputs, indirect injection through documents the agent retrieves, jailbreaks of safety guardrails, data exfiltration through clever multi-turn manipulation, function-calling abuse where the agent is induced to invoke tools with malicious inputs.

The cost of agent failures has become visible and material. In 2024-25 several public incidents — DPD's customer-service chatbot swearing at customers, Air Canada's chatbot promising fare discounts the airline then had to honour in court, internal copilot exfiltration of training data — established that production AI agents are an enterprise risk category requiring dedicated security attention. By 2026, enterprise legal, compliance, and security teams require adversarial testing before production deployment.

The supply side is thin. The dedicated AI-security firms (Robust Intelligence, HiddenLayer, Lakera, Protect AI) charge $50k-300k per engagement and have multi-month waitlists. Big-four consultancies have begun building practices but their AI-specific expertise is shallow. In-house red teams exist at FAANG-scale companies but nowhere else. The middle market — Series B-D AI startups, mid-size enterprises deploying AI — is unserved at the quality + price they need.

The Solution

Adversary delivers structured AI agent adversarial testing as a productised service. The standard engagement covers 14 attack categories: direct prompt injection, indirect prompt injection (via retrieved documents), jailbreak attempts (DAN-style and modern variants), instruction hierarchy violations, data exfiltration (training data, system prompt extraction, customer data), function/tool-calling abuse, multi-turn manipulation, role-play exploitation, encoding/obfuscation attacks, persona hijacking, context-window manipulation, output-injection attacks (XSS via agent outputs), business-logic abuse, and supply-chain attacks (poisoned retrieved content).

Each engagement begins with a 90-minute scoping call to understand the agent's architecture, surface area, tool access, and business risks. The testing team (2-3 testers per engagement) then conducts a structured 2-4 week test program, documenting every successful attack with reproducible scripts. The final deliverable is a written report with: executive summary, methodology, findings ranked by CVSS-adapted severity scoring, reproducible attack scripts, specific remediation guidance per finding, and a follow-up retest option included.

Three structural differences from incumbents define the wedge. First, India-staffed delivery: 40-55% lower cost than US firms at quality competitive with the same firms (delivered by senior testers with prior ML or security backgrounds, not undifferentiated juniors). Second, productisation: standardised methodology, attack-pattern library, report template, and engagement scoping that allow clients to compare engagements meaningfully and that produce faster turnaround. Third, continuous-testing retainer model: monthly testing as the agent evolves vs. one-shot engagements that go stale within weeks.

Continuous testing retainer ($5k/month) provides monthly adversarial testing of an existing agent with delta-from-last-month reporting, immediate alerting on new high-severity findings, and quarterly comprehensive retests. Annual enterprise contracts ($25k-80k) provide quarterly comprehensive engagements + named lead tester + on-call response for production incidents.

Market Opportunity

The serviceable global market for AI agent adversarial testing is approximately 12,000 companies in 2026 with customer-facing AI agents in production — growing at ~150% per year. Of these, an estimated 4,000-6,000 fall into Adversary's sweet spot (post-product-market-fit AI startups + mid-size enterprises) where the engagement budget is $10k-100k per year and the dedicated AI-security firms are unaffordable.

At a blended ARPU of $35,000 per customer per year (mix of one-time engagements and retainers), the SAM is approximately $140-210 million growing at 100%+ annually. Capturing 1% of the SAM in year 2 is a $1.5 million ARR business; 4% in year 4 is $6-8 million.

Adjacent expansion opportunities: model security testing (testing the underlying LLM rather than just the agent — this is a separate technical specialty growing alongside), AI compliance and audit work (EU AI Act, NYC Local Law 144, Colorado SB-205 all require demonstrable AI safety testing), and AI incident-response services (forensics + remediation when something goes wrong in production).

Target Customer

Primary persona: a 38-year-old VP Engineering at a Series C SaaS company that shipped a customer-facing AI agent 4 months ago. The agent has handled 280k conversations; one customer publicly tweeted a successful jailbreak two weeks ago that prompted board attention. The VP has no in-house red team, was quoted $180k for a Robust Intelligence engagement, and needs something faster and more affordable. Will engage Adversary for a $24k initial engagement + $5k/mo retainer.

Secondary persona: a 44-year-old CISO at a mid-size US financial-services firm that has deployed an internal AI agent for analyst research. Compliance is asking for documented adversarial testing before allowing the agent to access customer-account data. Will engage Adversary for a $40k initial engagement and a $50k annual enterprise contract.

Tertiary persona: a 31-year-old founder of a vertical AI startup at $4M ARR who shipped a voice agent feature. Voice agents have a different attack surface (tone-based manipulation, audio injection attacks). Will engage Adversary for a focused $14k engagement on the voice surface specifically.

Product

Engagement methodology (publicly documented): scoping → reconnaissance (understand agent architecture, system prompt structure, tool access, RAG corpus) → structured testing across 14 attack categories with documented coverage → exploitation depth on findings (chain attacks, not just single-shot) → reporting → optional retest.

Attack-pattern library: internal repository of 2,000+ attack patterns accumulated from prior engagements, anonymised and generalised. Updated weekly with new patterns from the security research community + internal discovery. Provides systematic coverage that ad-hoc testing cannot.

Tester tooling: internal harness for running attack patterns against client agents at scale (rate-limited respectfully), recording attempts and outcomes, scoring success. Includes adapters for common agent platforms (LangChain, LangGraph, AutoGen, custom REST APIs, voice agents via Twilio).

Reporting template: standardised structure (executive summary, methodology, findings with CVSS-adapted scoring, reproducible scripts, remediation guidance) that clients can hand to their engineering teams for action. PDF + Markdown + structured JSON exports.

Retainer service: monthly automated testing run + manual depth-test by named tester, delta-from-baseline reporting (what is new since last month), immediate Slack/email alerting on new high-severity findings, quarterly comprehensive retest.

Technical Architecture

Tester harness: Python orchestration framework that runs attack patterns against client agents with rate-limiting, parallelisation, result recording. Built on top of standard testing frameworks (pytest), with custom adapters per agent platform.

Attack-pattern library: Git-based versioned repository with pattern metadata (target category, sophistication level, success-rate history). 30-40% of patterns are public security-research-derived; 60-70% are internally discovered.

Reporting infrastructure: Markdown source + Pandoc generation pipeline producing PDF reports with consistent branding. Findings tracked in client-isolated Postgres database.

Client engagement workspace: Notion-based client engagement room (status, deliverables, communication thread), with limited-time access for the client team during the engagement.

Compliance: SOC2 Type II from year 1 (table-stakes for selling to enterprise customers). NDAs in place before any engagement begins. Tester laptops on managed-device infrastructure with strict access controls.

Pricing infrastructure: Stripe for engagement billing (one-time and retainer), with US entity for ease of US customer payment processing while delivery is India-based.

Business Model & Unit Economics

Three engagement structures. Project engagement: $8k-40k per engagement, 3-6 week delivery, single comprehensive test. Continuous retainer: $5,000/month (12-month minimum commit), monthly testing as agent evolves, quarterly deep-test included. Annual enterprise: $25k-80k/year, quarterly comprehensive engagements + named lead tester + on-call.

Conversion economics: sales cycle 6-12 weeks (mid-market AI security buying is consultative). Conversion rate from qualified inbound: 24%. Average customer LTV: $58,000 (mix of one-time and recurring). Customer expansion: 35% of project clients convert to retainer within 6 months as their agent ships new versions.

Gross margin: 62% blended. Per-engagement cost structure dominated by tester labour (~30 tester-hours per standard engagement at all-in cost of $55/hour = $1,650 direct labour cost). Senior tester / engagement lead overhead adds another 15-20% in time. Infrastructure, sales, ops add ~12% overhead.

Year-1 tester team: 4 senior testers ($85k each), 8 mid-level testers ($45k each), 2 engagement leads ($110k each). Total team cost: ~$920k. Capacity: ~250 engagement-weeks per year ≈ 60 engagements + 25 retainer-months equivalents.

Unit Economics (Year-1 base case)

Year-1 engagements (target)	80
Year-1 active retainers (year-end)	25
Year-1 revenue	$1.8 million (~■15 crore)
Gross margin	62%
Customer acquisition cost (CAC)	$3,200
Payback period	~3.5 months
Year-1 all-in costs	~■5.2 crore
Year-1 net contribution	~■4.1 crore

Go-to-Market

Channel 1 — Founder + security-community content marketing (35%): publish substantive technical write-ups on attack patterns (sanitised from real engagements), case studies, speaking slots at AI security conferences (DEFCON AI Village, BSides AI, RSA). Builds inbound from technical buyers who recognise quality.

Channel 2 — Targeted outbound to Series B-D AI startups (30%): direct outreach to VP-Engineering / CISO titles at 800 mid-stage AI startups with public AI agent deployments. Conversion target: 4 engagements/month from outbound.

Channel 3 — Channel partnerships with CISO networks and AI security consultancies (20%): partnerships with CISO-network communities (CISO Connect, Cybersecurity Collaboration Forum), with adjacent AI consultancies (compliance, model governance) for referral exchanges.

Channel 4 — Compliance-driven inbound (15%): as EU AI Act and US AI safety regulations create explicit testing requirements, inbound from compliance buyers seeking audit-ready adversarial testing.

Roadmap (first 12 months)

- Month 1-3: Team setup — recruit 4 senior testers + 4 mid-level testers in Bengaluru, build internal tooling MVP, accumulate first 200 attack patterns, deliver first 8 engagements.
- Month 4-5: Public methodology + first 3 technical write-ups published, retainer tier launched, scale to 12 testers, 25 engagements cumulative.
- Month 6-8: Voice agent specialisation added (different tester skill), enterprise annual contract structure launched, 45 engagements + 10 retainers cumulative.
- Month 9-10: SOC2 Type II achieved, partnership programme with 2 compliance consultancies operational, 65 engagements + 18 retainers cumulative.
- Month 11-12: 80 engagements + 25 active retainers, $1.8M ARR run-rate.

Key Risks

- Big AI labs (OpenAI, Anthropic, Google) shipping comprehensive built-in adversarial testing tooling that reduces need for external testing — partial response is likely (already happening with OpenAI's evals) but enterprise customers will continue to require independent testing for governance/audit reasons; mitigated by positioning as independent third party.
- Big four consultancies (Deloitte, PwC, EY, KPMG) building AI security practices at scale — possible 18-24 month threat; mitigated by depth advantage (we are specialists; they are generalists with AI overlay) and price advantage (their hourly rates are 3-5x ours).
- Quality variance across testers leading to inconsistent engagement outcomes — substantial risk for a services business; mitigated by senior-tester review on every deliverable, by structured methodology that constrains tester variance, by client-satisfaction monitoring with NPS tracking per engagement.
- Liability if a tested system is later breached via a vector Adversary missed — significant exposure; mitigated by professional indemnity insurance ($2-5M coverage), by clear scope-of-engagement contracts disclaiming completeness guarantees, by SOC2 + ISO 27001 controls demonstrating professional standards.
- AI agent technology shifting rapidly: new attack categories emerge monthly; tester skills require continuous updating — mitigated by dedicated research time (10% of every tester's time on attack-pattern research), by attack-pattern library investment that compounds, by community engagement (paper publishing, conference presence).