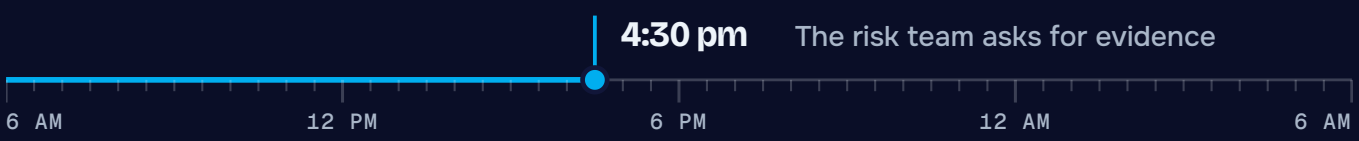
Web Vijayi

SERVICE BROCHURE

AI red teaming and evaluation

Find the weak spots in your AI before your customers do.





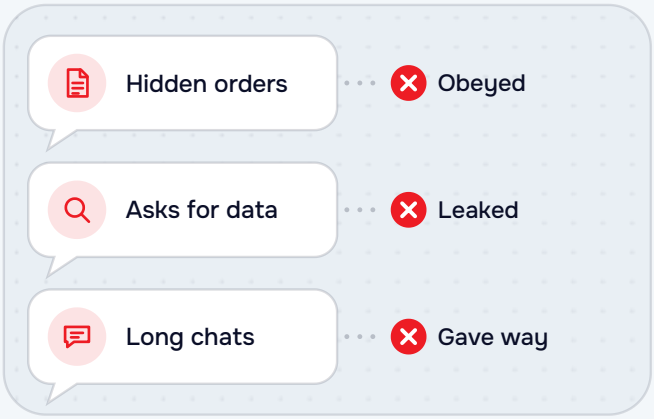
Who it is for

Banks, insurers, retailers and large firms whose AI assistant talks to many customers or staff, or helps make decisions that matter.

- BANKS AND INSURERS
- LARGE RETAILERS
- IN-HOUSE STAFF ASSISTANTS
- TEAMS FINE-TUNING MODELS

A working demo is not proof

- The product team has a working demo, and the risk, legal or security team wants evidence before launch.
- Sometimes the AI is already live, complaints are coming in, and nobody can say how often it gets things wrong.
- An assistant that refuses a bad request once can give way after several turns of persuasion.



BEFORE: CUSTOMERS FIND THE WEAK SPOTS FIRST

What we do

Our trained reviewers try to make your chatbot, agent or model leak data, break its rules or give harmful or wrong answers. They score it against pass marks you set, and you get the findings and a list of what to fix. We can test again after each change.

Red teaming

Reviewers try to break it on purpose with hidden orders, long chats and tricks to reveal its instructions.

Evaluation

A test set built from your real use cases, scored against your pass marks on every release.

Human feedback (RLHF)

Reviewers pick the better of two answers by your guidelines, for teams that train their own models.

How it works

WE DO THIS · TOGETHER

1 · 2 · 3 · 4

TOGETHER

We agree what could go wrong

With your product, risk and legal teams we list the harms that matter, who your users are and the pass mark for each risk.

TOGETHER

We write the guidelines and run a pilot

Reviewers get written instructions and examples that you approve. A small pilot shows where they're unclear before the full run.

US

We test, check and score

Reviewers work to the guidelines. Automated tools and AI judges add volume, and people check a sample of everything they score.

TOGETHER

We report, you decide, we retest

You get the findings, scores and a plain summary for leaders. Your team decides what to fix, and we rerun the agreed tests after changes.

What you get

- ✓ A written risk list and pass marks agreed with your teams
- ✓ Reviewer guidelines with worked examples, yours to keep
- ✓ Red-team findings, each with the conversation that caused it and a suggested fix
- ✓ A test set built from your use cases, with scores compared round by round
- ✓ Preference data and rated answers in the format your team trains on
- ✓ A record of reviewer instructions and checks, and a short summary for leaders

What we need from you

- Access to the system, ideally a test copy
- Your rules: what the AI must never say or do
- One person who can answer guideline questions quickly
- Examples of good and bad answers, if you have them
- Your data rules, and any agreement to sign before work starts

Timeline We agree the timeline with you before we start.

You stay in control

Anything unusual comes to you or your team to decide.

Nothing starts without your yes

No testing starts until you approve the guidelines, scope and attack list. Live tests run only in the times you set.

Progress you can see

You see progress, quality checks and early findings as the work runs, and we keep a record of what was done.

Automated tools on your terms

You decide what the tools may do without asking. You're told whenever they run, and can stop them or approve each run.

Change, pause or end it

Change the scope, the guidelines or the volume, pause the work, or end it, whenever you need to.

Questions people ask

How is this different from a pen test?

A pen test probes your networks, servers and code. Red teaming goes after the AI itself, with crafted prompts, long chats and planted documents.

Can automated tools do this instead?

They do part of it. Tools write many attacks and score answers fast, but they miss subtle harms, so people check a sample of every automated score.

How often should we retest?

After any change to the model, its instructions, the tools it can use or the data it can reach. Each of those can open a new weakness.

Which languages and markets do you cover?

We agree the languages, the markets and the mix of reviewers with you before work starts, based on who uses your AI.

Tell us what your AI does

In a free call, tell us who uses your AI and what worries you most. We'll suggest where testing should start and what a pilot would cover.

We quote every project in writing after a free consultation, with the timeline in the same quote.

webvijayi.com

care@webvijayi.com

Book a free consultation:
webvijayi.com/contact