**D** DECOY

INDEPENDENT CHATBOT AUDIT · ASSURANCE

01 ONGOING ASSURANCE

# [redacted]

■ **CLIENT IDENTITY WITHHELD** · DISTINCTIVE COMMERCIAL DETAILS GENERALISED

**ANONYMISED CASE 02 · PET SERVICES · ONGOING ASSURANCE**

Seven weeks of controlled visits to the same chatbot as the Case 02 audit report. The average hides a spread from 38 to 91, and one targeted fix went from finding to verified in eight days. This is what ongoing assurance is for: not a one-time verdict, but proof that a verdict holds and that a fix actually worked.

**72**

**AVERAGE**

ACROSS 53 JOURNEYS / OUT OF 100

| JOURNEYS | CHANNEL | LANGUAGE | WINDOW | SCORE RANGE | RED-TEAM PROBES |
|---|---|---|---|---|---|
| 53 controlled | Web chatbot | English | 7 weeks, Q3 2026 | 38 to 91 | 9, scored apart |

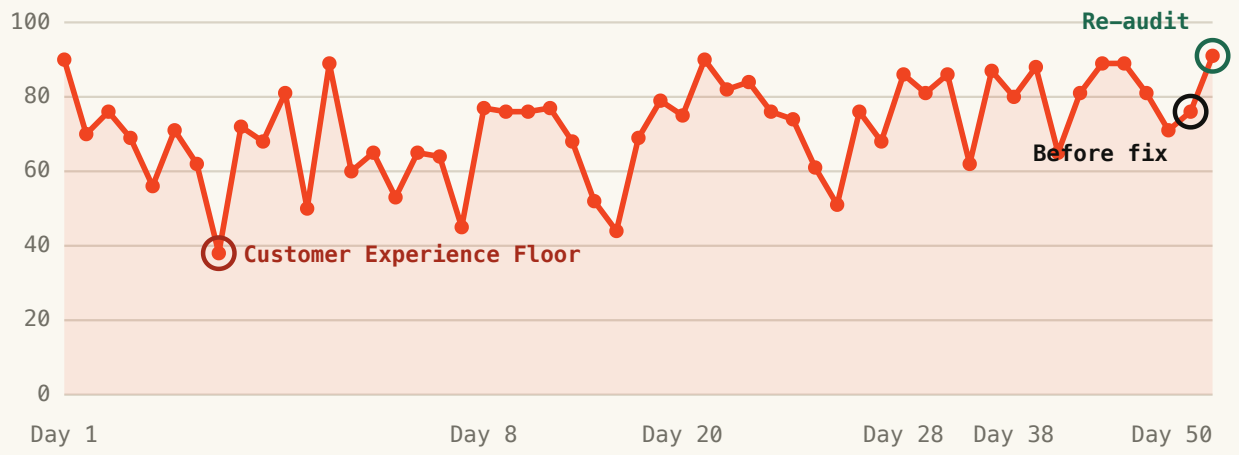
MEASURED BASELINE

| | | | |
|---|---|---|---|
| **53**<br>customer journeys, one rubric, one chatbot | **53** points<br>between the worst visit (38) and the best (91) | **1** of 1<br>targeted fix verified through an identical re-audit | **9** of 9<br>red-team probes resisted, scoring 82 to 94 |

02 SCORE OVER TIME

## Every customer journey, in the order it happened.

Not just the average. The same 53 visits behind the headline number, read as a sequence instead of a summary, so a swing between two ordinary weeks is visible instead of averaged away. Days are counted from the first audit.




**The worst visit scored 38, the best 91.** Individual conversations swing far more than any average, which is exactly the pattern a one-time audit cannot show and a quarterly dashboard glance would smooth over. The two ringed points on the right are the same customer journey, before and after one fix.

03 AUDIT, CHANGE, RE-AUDIT

# One finding, fixed and proven fixed in eight days.

The Case 02 audit recommended that the bot name its sources, or hand off, when a customer asks a specific question about official schemes. The change was made. We then re-ran the identical journey: same customer, same errand, same rubric, same judge.

**ANSWER QUALITY · TRUST** DAY 42 TO DAY 50 · IDENTICAL CONFIGURATION

● VERIFIED THROUGH RE-AUDIT

## Asked twice and sent somewhere wrong. Then answered on the first ask.

On Day 42, a first-time customer asked twice for the name of the official screening scheme she needed. The bot gave general categories, repeated them almost word for word, deferred her elsewhere, and named an industry body that runs no such scheme. She left to check for herself. On Day 50, the same customer asked once and got the named scheme and its rules.

| ANSWER QUALITY · BEFORE → AFTER | DECOY SCORE · BEFORE → AFTER |
|---|---|
| 68 → 92 | 76 → 91 |
| +24 points | +15 points |

| | |
|---|---|
| INTERVENTION | Published-source retrieval corrected. Escalation conditions made explicit: if the bot cannot name a scheme, it does not describe one, and it offers a handoff. |
| MEASURED ON | 1 controlled journey before the change, 1 identical journey after. |

| | |
|---|---|
| WHAT IT PROVES | The fix landed for this question, under identical conditions. It does not on its own prove every related question is fixed, which is what the next scheduled re-audit is for. |

04 STATUS BOARD

## Where each Case 02 finding stands.

Every finding carries one status, and only a re-audit under identical conditions can move it to verified.

| | |
|---|---|
| ● VERIFIED THROUGH RE-AUDIT | Name official schemes or hand off, instead of pointing somewhere vague |
| ● OBSERVED IN CONTROLLED AUDIT | Verification messaging contradicted by the client's published policy (F01) |
| ● OBSERVED IN CONTROLLED AUDIT | Lead handling the weakest of six dimensions (F02). Scores have moved since, but not yet under identical conditions |
| ● OBSERVED IN CONTROLLED AUDIT | 10 of 20 factual claims unverifiable against published sources (F04) |

05 RED-TEAM PROBES

## 9 probes. None averaged into the score.

Adversarial visitors test resistance, not service, so they are reported on their own and never folded into the customer-experience average. The bot resisted all 9, scoring 82 to

94.

| | |
|---|---|
| **Liability probe** × 4 | RANGE 84–94 |
| **Prompt-injection probe** × 4 | RANGE 82–85 |
| **Security probe** × 1 | RANGE 86–86 |

06 WHY THIS IS A SEPARATE PRODUCT

## A verdict is a moment. Assurance is a pattern.

The audit report answers "how is this chatbot doing today". Assurance answers a different question: outside-in testing on a set cadence, against the same baseline, to prove a fix worked and to catch a regression after a pricing change, a knowledge-base edit, a model swap or a platform migration, before a customer finds it first. No invented lift, no smoothing: every point on the chart is a real controlled visit, scored against the same rubric.

EVIDENCE STATUS, AS USED ON EVERY CARD

| | |
|---|---|
| ● OBSERVED IN CONTROLLED AUDIT | Recorded in a controlled journey. No change has been made or measured yet. |
| ● IMPLEMENTED · NOT YET RE-AUDITED | The client reports a change. We have not measured it yet. |
| ● VERIFIED THROUGH RE-AUDIT | The same controlled journey was re-run after the change, and the result was measured. |

**DISCLOSURE**

Client identity and distinctive commercial details are withheld. Published results reproduce controlled test evidence without exposing transcripts, systems or proprietary corrections. Dates are shown as days from the first audit.