D DECOY

INDEPENDENT CHATBOT AUDIT

01 EXECUTIVE SUMMARY

[redacted]

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

ANONYMISED CASE 02 · PET SERVICES · ACCURACY + TRUST

We sent 52 controlled synthetic customers through this online pet-services platform's chatbot, across 13 distinct buyer types. It is fast, polite and almost never silent. It is also the weakest at the exact moment a visitor is ready to act, and it stated things about the business that its own published pages contradict.

72

ADEQUATE

DECOY SCORE / OUT OF 100

JOURNEYS	CHANNEL	LANGUAGE	AUDIT PERIOD	CLAIMS CHECKED	RED-TEAM PROBES
52 controlled	Web chatbot	English	Q3 2026	20	9, scored apart

MEASURED BASELINE

3 of 20 factual claims contradicted by the client's own published pages	**10** of 20 factual claims could not be verified against any published source	**10** of 52 journeys scored weak (below 50) on lead handling	**4** findings recorded: 1 critical, 2 high, 1 medium

CRITICAL FINDINGS, RANKED BY BUSINESS IMPORTANCE

F01

CRITICAL

A safeguard the bot described does not exist as stated.

The bot's answer on the verification process is contradicted by the client's own published policy.

F02

HIGH

Lead handling is the weakest of six dimensions.

62 out of 100 on average, falling as low as 20 on individual visits: the bot answers and rarely does anything with the answer.

F03

HIGH

The Customer Experience Floor sits 34 points below the average.

A prospect only ever experiences one conversation, and the worst one scored 38 against a 72 average.

F04

MEDIUM

More claims are unconfirmed than confirmed.

Of 20 statements a customer would act on, 7 verified, 3 contradicted, 10 unverifiable against the client's own published information.

WHAT THIS MEANS FOR THE BUSINESS

OBSERVED

A skeptical buyer asked three times whether eligibility testing is mandatory and who verifies it. The bot's answers softened each time and never gave a direct yes or no.

BUSINESS INTERPRETATION

The bot described a verification safeguard the client's own published FAQ contradicts. If this pattern generalises beyond one conversation, it is a trust and compliance exposure, not a wording issue.

OBSERVED

Lead handling averaged 62 out of 100 across 52 journeys, the weakest of six dimensions, and fell as low as 20 on individual visits.

BUSINESS INTERPRETATION

If this pattern holds at production scale, it represents material missed conversion exposure: the bot resolves the question and stops, on the majority of journeys tested.

PRIORITY ACTIONS

1. **Fix now.** Correct the verification messaging so it matches the published policy, the highest trust exposure found. See Finding F01
2. **Fix now.** Add a lead-capture or next-step prompt whenever a question resolves, the single weakest and most fixable dimension. See Finding F02
3. **Next iteration.** Re-run the claim ledger once the above are fixed, and validate the remaining unverifiable claims against source pages. See Finding F04

Based on 52 journeys across 13 buyer types, run from 8 July to 18 August 2026 while the chatbot and DECOY's method were both being refined. Full method note at the end of the report.

02 SIGNATURE METRIC

The average was acceptable. One customer experience was not.

Aggregate performance hides the visits that cost you money. We report the mean and the Customer Experience Floor together, because a prospect only ever experiences one conversation, and it might be the worst one.

72

AVERAGE

ACROSS 52 JOURNEYS

38

CUSTOMER EXPERIENCE FLOOR

WORST SINGLE VISIT

WHAT THE GAP MEANS

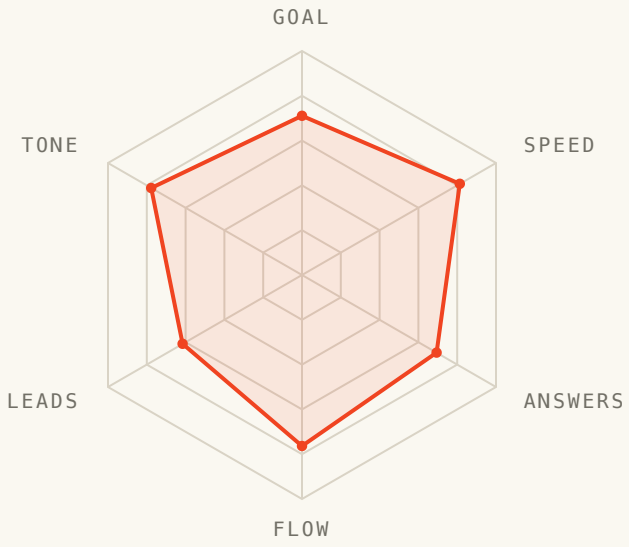
34 points separate the typical visit from the worst one. The Customer Experience Floor came from Jess - Skeptical visitor, who left unconvinced after asking for the verification specifics she needed in order to trust the platform. Nothing in an analytics dashboard would have shown you that conversation.

03 DIMENSION PERFORMANCE

Fast and polite is not the same as effective.

Six dimensions, averaged across every journey, with the full range so you can see how

much the experience varies rather than only where it lands.



READING THE SHAPE

Responsiveness is the strongest dimension at 81. Lead handling is the weakest at **62**, and it falls as low as 20 on individual visits. The bot answers quickly and then does nothing with the answer.

Goal completion 71

Across 52 journeys, from **35** to **100**.

Responsiveness 81

Across 52 journeys, from **45** to **98**.

Answer quality 69

Across 52 journeys, from **32** to **92**.

Conversation flow 76

Across 52 journeys, from **32** to **93**.

Lead handling 62

Across 52 journeys, from **20** to **85**.

Tone and brand 78

Across 52 journeys, from **54** to **90**.

04 FINDINGS BY PRIORITY

The chatbot described a safeguard the business does not have.

Four findings, ranked Critical to Medium. A skeptical buyer asked the one question that decided whether she would trust the platform, three times. Here is what she was told, and what the client's own published pages say.

CRITICAL · SYSTEMIC · TRUST
FINDING F01 · FACT-CHECKED AGAINST THE CLIENT'S OWN SITE

● OBSERVED IN CONTROLLED AUDIT

Verification was presented as thorough. It is not.

That eligibility testing and supporting documentation are actually required and verified before a seller profile is approved, not just requested.

EVIDENCE · VERBATIM, DOMAIN TERMS GENERALISED JOURNEY 07

CUSTOMER VERBATIM

ok "may need to provide" sounds pretty optional, is eligibility testing actually mandatory or just a nice to have?

CHATBOT VERBATIM

I understand you're looking for a clear answer, and I apologise if my previous response wasn't direct enough. It's not a simple "yes" or "no" because mandatory eligibility testing requirements vary significantly depending on the specific listing category and the standards that apply to that category. `[client]` requires appropriate supporting information and testing to ensure responsible selling, but the exact tests that are mandatory are not a blanket rule for every single listing. To get a definitive answer for a particular category, the `[client]` team can confirm the specific requirements that would apply.

CUSTOMER VERBATIM

ok fine category specific testing i get that, but who actually verifies the results, do you check documents or is it just self reported by the seller?

CHATBOT VERBATIM

I understand your concern about preventing irresponsible listings, and it's a very important point. `[client]` 's primary focus is on the rigorous verification of profiles and supporting documents *before* a listing or seller is approved to be listed on the platform. This initial screening is designed to ensure that ethical, quality, and legal standards are met from the outset. While we do investigate any concerns or bad reviews that are reported, our core strategy to prevent irresponsible selling practices is through this thorough upfront verification process and by promoting transparency and adherence to responsible seller guidelines for all users.

VERDICT

Contradicted by the client's own published FAQ.

BUSINESS IMPLICATION

A buyer makes a purchase decision on a safeguard that does not exist as described. The exposure is not the lost sale, it is the conversation that happens when she finds out.

VERIFICATION CONDITION	A comparable customer asking whether testing is mandatory receives an answer that matches the published policy, or is handed to a human.

HIGH · CONVERSION FINDING F02 · DIMENSION: LEAD HANDLING

● OBSERVED IN CONTROLLED AUDIT

Fast and polite is not the same as effective.

Lead handling averaged 62 out of 100 across 52 journeys, the weakest of six dimensions, falling as low as 20 on individual visits. See `Dimension performance` for the full range.

BUSINESS IMPLICATION	The bot answers quickly and then does nothing with the answer, on the majority of journeys tested rather than as an outlier.
RECOMMENDED ACTION	Add a next-step or lead-capture prompt whenever a question resolves.

HIGH · CONSISTENCY FINDING F03 · CUSTOMER EXPERIENCE FLOOR

● OBSERVED IN CONTROLLED AUDIT

The average hides the visit that cost the most.

The worst single visit scored 38 against a 72 average, a 34-point gap. See `Signature metric` for the comparison.

BUSINESS IMPLICATION	A prospect only ever experiences one conversation. If theirs sets the Customer Experience Floor, the aggregate number is not what they will remember.
RECOMMENDED ACTION	Investigate what made the visit that set the Customer Experience Floor worse than typical, and retest that persona and funnel position specifically.

MEDIUM · TRUST FINDING F04 · CLAIM LEDGER

● OBSERVED IN CONTROLLED AUDIT

More claims are unconfirmed than confirmed.

Of 20 statements a customer would act on, 7 verified, 3 contradicted, 10 unverifiable against the client's own published information. See the `Claim ledger`.

BUSINESS IMPLICATION	A buyer more often cannot confirm what the bot told them than can, which compounds the trust exposure in Finding F01.
RECOMMENDED ACTION	Validate the 10 unverifiable claims against source pages and retest.

05 CLAIM LEDGER

Every factual claim needs a source, or an honest verdict.

20 statements the chatbot made that a customer would act on, each checked against the client's published information. **7 verified**, **3 contradicted**, 10 unverifiable. We do not guess: a claim we cannot source is reported as unverifiable rather than scored either way.

That eligibility testing and supporting documentation are actually required and verified before a seller profile is approved, not just requested.	CONTRADICTED
There is no separate or final approval stage specifically for the counterpart listing owner, contradicting the approval process the client describes publicly	CONTRADICTED
That verification only requires a registration document and reference number, with no further cross-checking against external registers.	CONTRADICTED

That a dedicated trust team investigates bad reviews and can delist sellers/listings as a result.	VERIFIED
That listings must meet the platform's stated verification standard, with origin disclosed as advertised.	VERIFIED
The platform fee is tiered, with the enquiring party paying more than the listing owner (the visitor had read a single flat fee for both)	VERIFIED
`[client]` does not run independent third-party checks against external registries, relying on document review	VERIFIED
That the published pricing page exists and lists the tiered fee for each party.	VERIFIED

06 RECOMMENDED ACTIONS

Raised by different customers, independently.

A single visitor cannot tell you whether a problem is systemic. These were surfaced by separate customers on separate visits, which is what makes them worth fixing before anything else.

01 Avoid repeating the same deflection paragraph verbatim across multiple turns; vary structure or escalate to a live human sooner when the same question is pushed back a second time

RAISED INDEPENDENTLY BY 3 DIFFERENT CUSTOMERS

02 Capture the lead instead of handing out a generic `[client email]` inbox: offer to take her email, book a callback, or start a profile, especially for a high-intent-but-unsure visitor like this one.

RAISED INDEPENDENTLY BY 3 DIFFERENT CUSTOMERS

03

Equip the bot with named, verifiable certification schemes (for example recognised inspection protocols and the relevant industry body's compliance programs) or a link to an official resource, so it can answer category-specific questions with real specifics.

RAISED INDEPENDENTLY BY 3 DIFFERENT CUSTOMERS

● VERIFIED THROUGH RE-AUDIT PUBLISHED-SOURCE RETRIEVAL CORRECTED · ESCALATION CONDITIONS MADE EXPLICIT · SEE THE CASE 02 ASSURANCE REPORT

04

Close the long dead-air gaps and slow opening replies so frustrated visitors do not disengage mid-conversation

RAISED INDEPENDENTLY BY 3 DIFFERENT CUSTOMERS

05

Give concrete answers on refunds, profile rejection, and typical review/go-live timelines instead of routing everything to `[client email]`

RAISED INDEPENDENTLY BY 3 DIFFERENT CUSTOMERS

06

Give the bot a concrete, pre-approved description of the actual verification steps (e.g. does staff contact third parties, check registration numbers) so it can answer specifically instead of repeating boilerplate.

RAISED INDEPENDENTLY BY 3 DIFFERENT CUSTOMERS

PRIORITY MATRIX

FIX NOW	NEXT ITERATION	MONITOR
Correct the verification messaging to match published policy	Vary the repeated deflection paragraph across turns	Refund, rejection and timeline questions currently routed to a generic inbox
Add a lead-capture or next-step prompt when a question resolves	Equip the bot with named, verifiable certification schemes and links	
	Close the dead-air gaps and slow opening replies	

07 THE CUSTOMERS WE SENT

13 buyer types, 52 journeys.

Each is a controlled synthetic customer written for this business, with a fixed situation, goal and manner. The same customer type run repeatedly is how the range below is measured rather than assumed.

Beryl - Elderly low-tech owner	52 AVG · 2 RUNS · 50–53
Karen D. - Angry small seller	52 AVG · 3 RUNS · 44–61
Marika Toussaint - Competitor doing recon	57 AVG · 2 RUNS · 52–62
Alex - Compliance & privacy checker	65 AVG · 2 RUNS · 62–68
Mark - Registered seller	68 AVG · 2 RUNS · 65–71
Jess - Skeptical visitor	71 AVG · 11 RUNS · 38–89
Dave - Premium listing owner	72 AVG · 3 RUNS · 68–77
Priya - Brand enthusiast	72 AVG · 8 RUNS · 60–88
Mia - Total beginner	76 AVG · 4 RUNS · 69–86
Marco Reinhardt - Overseas first-time visitor	77 AVG · 1 RUN · 77–77
Sarah - Curious first-time owner	79 AVG · 10 RUNS · 56–90

Dev "Kestrel" Okafor - Chatbot red-teamer	83 AVG · 2 RUNS · 79–87
Deb - Uninformed newcomer	85 AVG · 2 RUNS · 81–89

08 METHOD AND SCOPE

What this covers, and what it does not.

Synthetic customer testing evaluates controlled, multi-turn journeys across key funnel stages rather than every possible conversational permutation. It provides an empirical diagnostic baseline without replacing internal telemetry. Testing is conducted through the public customer interface only: no SDK, no backend access, and no changes to the client's systems. 9 adversarial probes were run separately and are never averaged into the customer-experience score, because a probe measures resistance rather than service.

HOW THESE NUMBERS WERE COUNTED

The 52 journeys ran from 8 July to 18 August 2026 while the chatbot and DECOY's method were both being refined. More than one AI judge model scored them. The averages include two runs of the red-team persona Dev "Kestrel" Okafor and one visit by Marco Reinhardt, whose persona was written for a different business. The nine adversarial probes were scored separately. The trends report counts 53 journeys because it also includes the re-audit on 26 August 2026.

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 systems or proprietary corrections. Quoted exchanges are verbatim, with domain terms generalised.