Silence Rather Than Error

What checking one public AI agent through two independently applied approaches revealed

Sergei Ponomarev · Andrey Ekhmenin · August 2026

This article sets out the results of joint research we carried out in August 2026.

Sergei Ponomarev, PhD in political science, specialist in open government, author of the test purchase methodology for AI agents. For seven years he assessed the quality of public administration, running national monitoring programmes at an independent analytical centre. He now applies those methods to AI agents.

Andrey Ekhmenin, founder of ANDEKS, independent assessment of complex AI, digital and governance systems. In this experiment, ANDEKS™ was used for an independent assessment of what exactly may be treated as established on the available evidence and where the limit of a permissible conclusion lies.

We decided to check one and the same live AI agent through two different approaches and see what would come of it. The object was the public support agent of company X, a developer of chatbots for customer service. The company is not named: this article is about the type of problem found, not an assessment of a particular supplier.

We worked independently. Each of us decided on his own what to ask and how, each froze his result before seeing the other's, and only then did we compare them.

The main thing we found, both of us and separately: no incorrect information was identified in the answers examined, yet the agent did not state some of the published conditions for using the service. Below we describe how the check was built, what each side found, where we diverged and what follows from it.

1. The question we started from

A company places an AI agent between itself and its customer. The agent answers quickly, politely and, as a rule, correctly. That raises a question the market has no settled answer to: how do you check that it actually does what the company promises, and what counts as established at all?

The second question is the harder one. An observation and a conclusion are not the same thing. The agent answered a question about a plan correctly. Does that make it accurate? In one answer, in one session, to one question, yes. Beyond that lies the territory where different methods draw the line differently.

We did not set up the experiment to find out whose method is better. What interested us was something else: where two different approaches would see the same thing, where they would complement each other, and where they would diverge on what counts as established.

2. How the experiment was built

The object: the public AI support agent of company X, a developer of chatbots for customer service. The agent is available to any visitor to the company website through a widget. The company is not named in this article: its subject is the type of problem found, not an assessment of a particular supplier.

One detail of the object matters. The company builds AI agents itself, and what was checked was its own agent, built on its own platform. This is not an incidental user of someone else's technology but a specialist demonstrating its product on itself.

The order of work was set before the start and did not change:

1. The parties jointly froze the object and the boundary of assessment: which agent exactly was under examination, what was not part of the object, and which public pages formed the normative surface. The freeze record was confirmed by both sides.

2. Each side worked on its own, without agreeing questions, the number of interactions, criteria or interim observations.

3. Each result was frozen before its author saw the other side's result, and was not altered afterwards.

4. Comparison took place only after mutual disclosure. Original conclusions were not revised retrospectively; clarifications that arose after disclosure were placed in a separate section of the comparative record.

The first line of work: the test purchase (S. Ponomarev).

The method comes from consumer oversight. The checker arrives as an ordinary customer and compares what the company promised publicly with what the agent says in a live conversation. The interactions were carried out by an AI checker under the observation of the methodologist, following a pre-prepared scenario of an ordinary customer with fictitious details. Eight promises were selected from the public pages, each verifiable at the moment of interaction and material to the customer. The requirements were fixed by capturing the pages with checksums before the check began. Two sessions were run under a single protocol of fourteen verbatim questions in the same order, one by day and one at night, from different external addresses. A positive verdict required confirmation in both sessions.

The second line of work: ANDEKS (A. Ekhmenin).

This method proceeds not from testing a service but from independent assessment. For every claim examined it builds a chain: object, boundary, public basis, evidence required, observed fact, permissible conclusion, limitation of the conclusion. A missing basis or missing evidence is never filled in by the assessor's assumption. Six tests were carried out, including comparison of answers given in separate sessions, each with primary visual capture and a separate qualification of what that test does and does not establish.

The two lines of work were therefore not assumed to be functionally equivalent: they independently examined the same object but asked different questions of the evidence.

3. What the observations showed

The observations of the two lines of work matched almost entirely. Every numerical fact about the plans that was asked for, the agent stated correctly. In the test purchase the same questions were put verbatim in two sessions and produced the same values; in the independent assessment the same fact was requested in different wordings and was likewise reproduced. The agent held the context of the conversation and adapted its advice to the situation the customer had described. On questions where the agent said it had no answer, it invented nothing.

Speed was measured only in the test purchase: the recorded delays before a substantive answer began ranged from 2.0 to 5.6 seconds. The independent assessment did not measure the delay precisely and recorded only that answers were received.

Neither line of work identified any incorrect information in the answers examined.

4. The principal finding: silence rather than error

The test purchase put two questions to which the company already has published answers.

The first question: may the service be used for customer support where the conversation would involve patient data protected by HIPAA? The company's terms of use prohibit this outright. The terms state, in substance, that the services are not tailored to comply with industry-specific regulations such as HIPAA, and that if your interactions would be subject to such laws, you may not use the services.

In both sessions the agent gave the same reply: it could not find information about HIPAA compliance and advised contacting support. About the prohibition written in its own company's terms it said nothing.

The second question: is there a money-back guarantee? The terms state that purchases are non-refundable. The agent did not invent a guarantee, which is right, but neither did it pass the published condition on to the customer.

The independent assessment put a third question, one that did not form part of the executed test purchase protocol: what may the company do with personal information a customer writes in the chat? The company has a published privacy policy stating why data is collected, to whom it may be disclosed and how long it is kept.

The agent replied that it had no specific information about how personal data is processed. Asked again about sensitive data, it advised against disclosing it. Sensible advice, but the customer never heard what the published policy actually says.

Three different questions, two independently applied approaches, one result.

An agent can correctly hold the limit of its knowledge, state nothing false, and at the same time fail to convey to the customer an existing public rule of its own company. Freedom from error is not the same as sufficiency.

The practical consequence is simple. A customer who asks any of those three questions leaves without an answer the company has and has published. Meanwhile a quality metric built on counting false statements will not see such a case: there are no false statements in these answers. The normative information gap established here consists precisely in the fact that an existing published rule was not conveyed to the customer.

5. Two classes of gap that must not be conflated

The comparison showed that cases which look alike belong to different classes, and the distinction turned out to matter.

Class	Situation	How to assess it

First	No unambiguous answer was found in the public sources fixed for the check; the agent states no unverified fact	Correct holding of the limit of knowledge: no unverified fact was stated, and on the available material no error is established
Second	The rule is published by the company, but the agent does not convey it to the customer	A normative information gap. This is the finding

An example of the first class from our experiment: asked whose model powers the agent, it replied that it had no specific information and named nobody. No unambiguous public answer relating to this particular agent was found within the normative surface fixed for the check, so declining to name a supplier is correct behaviour rather than a gap.

The examples of the second class are the three cases above. The difference between the classes is not one of wording. In the first case the checkers found no public basis against which the answer could be compared. In the second such a basis exists and is published, and the agent did not convey it.

6. How the methods qualified the observations

Let us compare how the results were qualified across three promises. This is perhaps the most substantive part of the experiment, because the difference lay not in the facts but in the threshold at which an observation becomes a claim that a public promise has been met.

Promise	Test purchase	Independent assessment
Instant answers	Confirmed against a fifteen-second threshold set by the methodologist and disclosed as a research operationalisation	Not established: the company named no figure, and the assessor may not set one on the company's behalf
Availability 24/7	Confirmed at the points checked, by day and at night	Not established for the general promise: two points do not prove continuity
Answers based on the company's data	Confirmed: the answers agree with the public data	Not established as to origin: agreement with the website does not prove the answer came from an internal knowledge base

On the first point the parties converged after disclosure: if the company has set no numerical threshold for the word "instant", the assessor should not set one in its place. The measurements are not thereby useless, since they establish the observed response time, but the threshold must be marked as an operationalisation of the method rather than as a standard of the company.

On the third point it emerged that there is no direct conflict: the methods measure different claims. The test purchase asks whether observable behaviour matches what was promised and does not set out to establish the origin of an answer: agreement between an answer and published information does not by itself establish its internal source.

On the second point there is no substantive contradiction either. Both sides established availability at the moments checked, and both note that continuity is not proven. The difference concerns how the result is presented: one side issues an operational verdict specifying the scenarios checked, the other leaves the promise unestablished.

7. A methodological consequence: four axes instead of one

The main lesson of the experiment for methodology is that the quality of an AI agent's answer cannot be captured by a single mark. In our case one and the same reply, "I have no information, please contact support", was faultless on one axis and unsatisfactory on another. Under a single assessment those states merge, and the method stops seeing the finding.

We propose separating at least four axes:

1. Factual correctness: the information given is true.

2. Absence of unverified invention: the agent does not fill a gap with a plausible fabrication.

3. Normative completeness: the agent conveys the conditions the company has published and that bear on the question.

4. Resolution of the customer's question: the customer received an answer, or a concrete way to resolve it.

The level of claim to which a conclusion belongs should be marked separately as well:

- a single observation;
- a repeated observation;
- a locally established property;
- a universal public promise.

The last level cannot be established on a finite sample of interactions: a single interaction can refute a universal promise, but a finite number of successful interactions does not confirm it for all remaining cases. This limitation belongs in reports explicitly rather than by implication.

8. What the experiment does not establish

- It does not assess the overall quality of the agent checked and does not compare it with other agents. Individual public promises were checked in individual scenarios.
- It does not establish the causes of the gaps found. Both checks were external: internal logs, configuration and the agent's knowledge sources were not examined.
- It is not a legal opinion. Compliance with data protection law or with industry regimes was not assessed.
- The test purchase found gaps in the explanation of refund conditions and of the HIPAA restriction; the ANDEKS independent assessment found a gap in the explanation of personal data processing. The two gaps found by the test purchase recurred in both of its sessions. The prevalence of this type of problem beyond the object checked is not established.
- Neither side's numerical results convert into percentages or serve as a quality rating: the checks are not of equal weight, and what matters is the findings rather than the tally.

A separate word on the evidence base. In the test purchase the original recording of the dialogue and screen captures were not preserved, so the transcripts of both sessions are derivative documents; some timing measurements in the night session were not taken; the night session took place outside the agreed time window; after browser state was cleared a service cookie remained, so full independence of the two sessions is not confirmed. In the independent assessment, screen captures served as primary evidence. Independence of two researchers and independence of two browser sessions are different things, and the second was only partly achieved in our case.

9. What follows from this for companies

The observation that made the experiment worth running fits in one line: an agent can state the parameters of a service correctly and at the same time fail to state the published conditions of its use.

On the object checked, questions about capabilities and plan parameters received definite and correct answers, while questions about restrictions, refunds and the processing of personal data received no definite answer. The reasons for that difference cannot be established by an external check. Its practical significance is nevertheless plain: a customer from a regulated sector who asks a direct question about restrictions receives silence instead of the published answer.

Three practical consequences follow.

1. An agent should be checked not only for correctness but for normative completeness. Questions about restrictions, refunds and data processing belong in the programme of checks alongside questions about prices.

2. Assessment must distinguish "did not lie" from "answered the question". A reply of "contact support" may be a proper outcome where a concrete route for handling the question is given, and an evasion where the company has a published, definite answer.

3. Where a promise does not set a sufficiently definite criterion of fulfilment, the report should show the observation and the research criterion used to assess it separately. The absence of a figure does not by itself make a promise unverifiable: the obligation for an agent to identify itself as AI, for instance, is verifiable without any threshold. But if a company wants its "instant" assessed unambiguously, it is the company that should name the figure.

10. The experiment and the authors

The experiment was carried out on 26 and 27 August 2026. The object and the boundary of assessment were frozen jointly before work began; each result was frozen independently and disclosed only after it had been fixed. The materials of both sides are held in a dossier with checksums.

Sergei Ponomarev, PhD in political science, author of the test purchase methodology for AI agents. Seven years of test purchases and service quality assessment; doctoral research on e-government. Portal aibusiness.vc.

Andrey Ekhmenin, founder of ANDEKS, independent assessment of complex AI, digital and governance systems.

This article describes the results of a private research experiment. It is not a certification, an audit, a validation of any method, or an assessment of the company checked as a whole. The convergences and differences observed between the two approaches are specific to this experiment and do not by themselves establish their general compatibility or equivalence.