ESOMAR
Congress 2025

ReImagine

Prague, Czechia
28 September - 1 October

Papers

Editorial

Lyndall Spooner

ESOMAR
Office address:
Burgemeester Stramanweg 105
1101 AA, Amsterdam
Phone: +31 20 664 21 41

Email: info@esomar.org
Website: www.esomar.org

Publication Date: September 2025
ESOMAR Publication Series Volume S419 Congress 2025
ISBN 92-831-0331-9

Copyright

In spite of careful preparation and editing, this publication may contain errors and imperfections. Authors, editors and ESOMAR do not accept any responsibility for the consequences that may arise as a result thereof.

The views expressed by the authors in this publication do not necessarily represent the views of ESOMAR.

By the mere offering of any material to ESOMAR in order to be published, the author thereby guarantees:

- that the author - in line with the ICC/ESOMAR International Code of Marketing and Social Research– has obtained permission from clients and/ or third parties to present and publish the information contained in the material offered to ESOMAR;
- that the material offered to ESOMAR does not infringe on any right of any third party; and
- that the author shall defend ESOMAR and hold ESOMAR harmless from any claim of any third party based upon the publication by ESOMAR of the offered material.

Published by ESOMAR, Amsterdam,
The Netherlands
Edited by: Bel Kerkhoff-Parnell, PhD.

About ESOMAR

ESOMAR is the champion of the insights and analytics sector. It is the business community for every sector professional. Founded in 1947, the global membership association is a network reaching over 40,000 professionals and 750+ companies in 130+ countries. We support individual and corporate members supplying or using insights by helping them raise ethical standards, grow the demand for insights, and improve its uses and applications by all decision-makers.

About ESOMAR Membership

ESOMAR is open to everyone, all over the world, who believes that high quality research improves the way businesses make decisions. Our members are active in a wide range of industries and come from a variety of professional backgrounds, including research, marketing, advertising and media.

Membership benefits include the right to be listed in the ESOMAR Directories of Research Organisations and to use the ESOMAR Membership mark, plus access to a range of publications (either free of charge or with discount) and registration to all standard events, including the Annual Congress, at preferential Members' rates.

Members have the opportunity to attend and speak at conferences or take part in workshops. At all events the emphasis is on exchanging ideas, learning about latest developments and best practice and networking with other professionals in marketing, advertising and research. CONGRESS is our flagship event, attracting over 1,000 people, with a full programme of original papers and keynote speakers, plus a highly successful trade exhibition. Full details on latest membership are available online at www.esomar.org.

Contact us

ESOMAR
ESOMAR Office:
Burgemeester Stramanweg 105
1101 AA, Amsterdam
The Netherlands
Tel.: +31 20 664 21 41

Email: info@esomar.org
Website: www.esomar.org

Can AI Stop AI from Faking Surveys?

Sebastian Berger
Julia Mittermayr
Bernhard Witt

ESOMAR
Office address:
Burgemeester Stramanweg 105
1101 AA, Amsterdam
Phone: +31 20 664 21 41

Email: info@esomar.org
Website: www.esomar.org

Publication Date: September 2025
ESOMAR Publication Series Volume S419 Congress 2025
ISBN 92-831-0331-9

CONGRESS 2025

Can AI Stop AI from Faking Surveys?

Sebastian Berger
Julia Mittermayr
Bernhard Witt

Introduction

Research agencies are losing an estimated £209 million per year in the UK alone due to poor data quality. This includes costs from re-fielding, lost staff time, license fees and compensating clients for unusable results (Harding, 2025). It is rare to hear someone publicly report that flawed market research led to poor business decisions. Yet Tia Maurer, Group Scientist at Procter & Gamble, did exactly that. Speaking at the ReDem Quality Day last year, she revealed how fraudulent survey data resulted in the launch of an oral care product that ultimately failed, resulting in significant financial losses and reputational damage for the company (Maurer, 2024). This example underscores a critical vulnerability in today's market research industry: the growing threat of survey fraud.

Some may argue that survey fraud is nothing new. But why, then, is it suddenly such a hot topic? Why are we seeing a surge in conference talks, articles and new fraud detection tools? Why has a global industry body, the Global Data Quality Initiative, been formed to tackle this issue? And why, ultimately, was this paper accepted for publication?

The answer is simple: survey fraud has changed. Radically.

That's why this paper begins by examining the shifts that have reshaped the fraud landscape. It then explores panel providers, their clients and AI as the three forces that help explain why, despite growing awareness, the problem remains largely unsolved. Finally, it presents the first experimental evaluation of whether AI-powered solutions are capable of stopping AI-generated survey fraud.

This study was conducted jointly by ReDem and SurveyTester to evaluate whether AI-powered tools can reliably detect AI-based survey fraud. While ReDem develops and deploys fraud detection systems, SurveyTester created bots designed to simulate realistic respondent behavior. By testing these bots against an AI detection pipeline, we aimed to create a controlled, adversarial environment, allowing us to examine a critical question for the future of the industry: Can AI detect AI? Or in other words, can AI be the solution to a problem AI itself has helped create?

The new fraud landscape

One might recall a time when organised survey fraud was mostly the domain of click farms. These centralised operations, often run like call centres in low-wage countries, relied on workers who rushed through questionnaires with random, nonsensical answers to collect incentives. At the time, poor-quality responses were easy to spot using simple checks like speeding, straight-lining or gibberish detection. In short, data cleaning was a straightforward, albeit manual task.

Despite the persistence of this image, the fraud landscape has evolved significantly, especially since the COVID-19 pandemic (Berger, 2024a). During that time, many click farms shut down, and workers lost their primary source of income. To make a living, some continued with survey fraud but now from home.

As online surveys became a commodity and price competition intensified, monetary incentives for participants dropped to very low levels. For fraudsters, now working alone or in small groups from private apartments, manually completing surveys proved inefficient. They quickly realised the advantages of scaling their activities using multiple devices and automation.

They acquired used smartphones, connected and operated them simultaneously, creating so-called phone farms. Beyond that, they exploited a wide range of readily available tools: VPNs to manipulate IP addresses, services to register large numbers of fake digital identities in survey panels, and automated systems to verify these accounts via SMS or voice. They also automated the survey-taking process itself. A striking example of this new reality was presented at the 2022 ESOMAR Congress (innovateMR, 2022a). In a video, a fraudster from Venezuela described how he operated a phone farm with 25 devices. He reported completing around 3,000 surveys per month using bots, earning roughly $2,500 during an average month which is a multiple of the average monthly income in his country. While early bots were rudimentary, an increasing number today are highly advanced using AI to complete questionnaires in ways that closely mimic real human behaviour.

And it doesn’t stop there. The evolution was further fuelled by the emergence of global online communities where fraudsters openly share tips, tools, survey links and manipulation techniques via social media, forums and messaging apps, making it easy for anyone to become a professional fraudster. Another video shown in the same conference presentation featured a fraudster from Bangladesh who posts survey fraud tutorials on YouTube and even offers one-on-one coaching sessions (innovateMR, 2022b).

The bottom line is that the fraud landscape has shifted from a predominantly centralised, low-tech annoyance rooted in low-wage labour markets to a decentralised, tech-driven global phenomenon.

The panel challenge: How fraudsters exploit the system

The radical transformation of survey fraud is evident not only in the rise of global online communities where fraudsters openly exchange strategies, but also in the implausibly high participation rates of certain registered panel members.

A study by CASE4Quality (2022) found that just 3% of devices accounted for 19% of all panel survey completions. Even more alarming: 40% of these devices submitted over 100 surveys per day, all while passing existing quality checks. Such scale is only possible through automation.

But volume is only part of the problem. A deeper, more subtle issue lies in the systematic bias these fraudulent responses introduce. These are not random errors that average out. Fraudsters intentionally manipulate their answers for entry and speed. At the screening stage, they tend to agree with statements and select multiple options to increase their chances of being accepted into a study. Once inside, they optimise for speed by selecting the fastest paths through surveys, skipping optional fields, minimising item engagement and only completing mandatory sections. Potloc (2025) found that responses from frequent survey takers (even those who passed fraud checks) did not match those from casual respondents. Their answers were shorter, more disengaged and completed in significantly less time, undermining the reliability of the data.

This shift in response strategy has profound implications. A study by ReDem and Media1 (2025) showed that just 16% low-quality responses in a brand tracker dramatically distorted key metrics. Brand awareness appeared 20 points lower, slogan recall dropped by 18%, purchase intent was overstated by 10% and TV ad recognition falsely increased by 26%. Similar findings come from emi (2023), which also observed a 20-point drop in brand awareness between low- and high-frequency survey takers. In addition, data quality failures rose with the number of survey attempts. These findings highlight that low-quality data doesn’t merely add noise, it systematically misleads.

Can this be prevented? In principle, yes. There are options available, such as video identification, passport validation, address confirmation, bank account checks and blockchain-backed identity tokens. Some panels also restrict access from users using VPNs to hide their IP address. While each method can be circumvented e.g., through AI avatars, forged documents or strawmen accounts, doing so requires significant effort. This would greatly reduce the scale, efficiency, and profitability of fraud.

So why haven’t such robust safeguards become the industry standard, especially given the rise in fraud? One reason lies in the economic structure of the business model. High-frequency survey takers often generate a significant share of the revenue. In contrast, more casual participants tend to contribute less.

And in a market where price, speed, and panel size dominate client decision-making, panel providers face a tough choice. Strict verification measures don’t just exclude fraudsters; they also deter legitimate users. Only a small minority of users are willing to go through video calls, submit personal documents, install a wallet for identity tokens or verify addresses and bank accounts just to earn a few cents per survey they qualify for.

To keep panel recruitment efficient, providers often optimise for reach and ease of onboarding, for example by using social media ads and minimising sign-up friction. Once recruited, participants represent a sunk acquisition cost. Removing them is a difficult trade-off that needs to be balanced against operational and commercial factors.

The client challenge: How mental shortcuts undermine research quality

So, who is responsible, then? If not the panels, it must be their clients. Yet here too, a major obstacle stands in the way: the mental shortcuts we, as human beings, rely on to simplify decision-making. The following five cognitive biases make it easier to overlook, excuse or misjudge fraudulent survey data often without even realising it (Berger, 2025a).

The head-in-the-sand phenomenon

There is a tendency to ignore or downplay the existence of fraud in studies. This avoidance is often driven by cognitive dissonance: the psychological discomfort that arises when new evidence (e.g., widespread fraud) clashes with preexisting beliefs (e.g., “our data is reliable”). To resolve this tension, professionals sometimes rationalise the problem away, claiming that fraudulent responses “cancel out” in large sample sizes, or that fraud “has always existed” and can be tolerated. But with current fraud rates frequently exceeding 10%, and sometimes reaching as high as 38% as reported by Kantar (Wang, 2023), this assumption is no longer tenable. Denial does not protect data, it exposes it.

The price-competition dilemma

In the highly competitive market research industry, price and speed are often prioritised over quality. This reflects outcome bias, where positive past results, such as delivering a project on time and within budget, are mistaken for sound decision-making, regardless of how the data was collected. On the provider side, strict quality controls are seen as costly, time-consuming, and commercially disadvantageous. On the buyer side, the pressure to deliver fast insights at low-cost can lead to blind trust in inexpensive solutions. But, as illustrated in the preface by Procter & Gamble’s experience with flawed product testing data, the cost of bad data can far outweigh the savings made during collection.

The “if I pay for it, it must be good” bias

This is the flip side of the price-competition dilemma, rooted in the price-quality heuristic, which is the assumption that higher cost implies higher quality. Clients often believe that working with premium providers or established brands ensures trustworthy data. This mindset is further reinforced by brand loyalty bias, where positive past experiences with a provider inhibit critical scrutiny in the present. However, in a field increasingly affected by professional survey fraud, trust is not enough. Clients must demand concrete proof of data quality, including documentation of flagged or removed interviews and the rationale behind those decisions. Without transparency, the assumption of quality may become a costly illusion.

The “if I like it, it’s good” mindset

This bias is rooted in the affect heuristic and confirmation bias. When results confirm prior expectations, e.g., a new ad performs well, or brand perception aligns with internal narratives, clients are more likely to trust the data. When results deviate from expectations, data quality suddenly becomes suspect. In both cases, data is judged not on its merits, but on whether it fits the desired storyline. This undermines objectivity and creates a dangerous double standard. True quality control requires that all results, favourable or not, be subjected to the same level of scrutiny.

The abstraction dilemma

Survey fraud today is complex, invisible, automated and constantly evolving. Because of this, it feels abstract and intangible, making it difficult to grasp the scale and nature of the problem. This triggers a tendency to reduce complex problems to simple narratives and solutions which is called oversimplification bias. Vague terms like "bots" or "click farms" are often used without understanding what they really entail. For example, many still imagine click farms as call-centre-style operations, when in reality they are often decentralised phone farms operated by individuals using networks of automated devices. This simplification extends to mitigation strategies. Techniques like instructed response items and age consistency checks are easy to implement and understand, but largely ineffective against modern fraud, as the next section will show.

Encouragingly, the tide is beginning to turn. Initiatives like the Global Data Quality Initiative and ongoing educational efforts such as those led by CASE4Quality are raising awareness of professional survey fraud. More and more clients are beginning to challenge outdated assumptions, acknowledge cognitive blind spots and place greater emphasis on data quality controls.

Still, change remains a slow process. Lasting improvement requires a fundamental cultural shift across the industry. Even when clients consistently prioritise quality and demand genuine transparency from their providers, ensuring that strict quality measures are not just marketing claims but actually enforced in practice, this alone doesn't solve the problem. Rapid advances in AI continue to introduce new threats, making the quest for reliable data quality an ongoing challenge.

The AI challenge: How rapid advances are outpacing quality control

Let's be clear: advanced AI-generated survey fraud is not the most urgent threat yet. The more pressing concern lies in two critical vulnerabilities, both stemming from the panel and client challenge described earlier. First, many questionnaires still lack the structure needed for robust quality checks (Berger, 2024b). For example, surveys that omit open-ended questions eliminate valuable opportunities for fraud detection. Without open ends, it becomes impossible to analyse answers for duplicated text, off-topic content or signs of AI-generated text. More advanced methods, such as analysing typing behaviour or detecting copy-paste patterns, also rely on the presence of open-ended input to distinguish between genuine and fabricated responses.

Second, the industry still relies heavily on traditional detection methods such as IP address checks, speeders, straightlining, and CAPTCHAs. But these techniques are no match for today's increasingly sophisticated fraud tactics. Modern fraudsters bypass them with ease, leaving poor-quality data undetected. By doing so, fraudsters adapt to the level of scrutiny. If the bar is low, they don't need to try hard. Like burglars choosing an open window over breaking down a door, fraudsters exploit the easiest path. For instance, if open-ended questions aren't mandatory, they simply skip them. In other words, fraudsters only use the minimum necessary to bypass existing checks.

As Sawhney et al. (2025) show, while simple bots are stopped by most standard detection tests, AI-powered bots can bypass nearly all of them. In their study, simple bots used only browser automation to interact with online questionnaires and lacked any reasoning or language capabilities. In contrast, AI-powered bots combined browser automation with large language models (LLMs), enabling them to understand and respond to complex text-based tasks using contextual information.

The researchers tested ten commonly used bot-detection methods, including ReCAPTCHA v2 ("I am not a robot" checkbox that analyses human-like mouse movement) and ReCAPTCHA v3 (invisible score-based bot detection via user behaviour), honeypots (invisible questions meant to trap bots that fill every field), anagram tasks (rearranging letters to form a real word), counting tasks (identifying the number of zeros in visual matrices), open-ended questions (requiring a written, thoughtful response), colour checks (choosing the one real colour word among distractors), instructed response items (e.g., “select ‘strongly disagree’”), age consistency checks (repeating age questions to check for matching answers) and an attention screener (ignoring a paragraph and selecting a specific option).

Results showed that the simple bot failed all checks except ReCAPTCHA v3 and the honeypot. In contrast, the AI-powered bot passed nearly all checks, failing only ReCAPTCHA v2 and two of the more complex counting tasks. This demonstrates that AI bots can easily bypass most existing detection techniques, highlighting a growing vulnerability in online survey quality assurance.

These findings raise a critical question: if both ReCAPTCHA v2 and v3 rely on behavioural signals, why do they produce different outcomes? And more broadly, to what extent can behavioural tracking be used to reliably detect bots?

Sawhney (2025) explores this further, emphasising that behavioural signals may be key to future-proofing survey integrity. Her study found that human respondents showed chaotic and unpredictable mouse movements, while bots moved smoothly and with minimal deviation, indicating scripted control. Scrolling behaviour also diverged clearly. Humans scrolled in varied patterns, often pausing or reversing direction, whereas bots typically performed a single top-to-bottom scroll. Tab switching, common among humans due to multitasking or distraction, was completely absent in bot behaviour. And perhaps most revealing, bots typed with consistent timing and uniform speed, devoid of the irregular pauses, backspacing and natural variation that characterise human input. These behavioural discrepancies underline the potential of behavioural tracking as a survey fraud detection tool. However, given the widespread use of mobile devices in survey participation, any such tool must be designed to reliably evaluate input behaviour across all device types.

The experiment: When AI fights AI

While recent findings are concerning, it remains untested whether AI-powered fraud can be reliably detected using advanced AI-based techniques, such as coherence analysis or the automated evaluation of open-ended responses for signs of AI-generated content. Moreover, no study has yet assessed how these techniques perform when combined with promising behavioural tracking approaches, particularly those focused on typing patterns. Pinzón et al. (2024) suggest that integrating multiple detection strategies may significantly improve effectiveness, a recommendation reinforced by the limitations identified by Sawhney (2025): bot behaviour can be engineered to closely mimic human input patterns. Developers can script mouse jitter, introduce artificial delays, or simulate natural-looking keystroke variability. Additionally, JavaScript-dependent methods are inherently fragile. Users can disable JavaScript or install browser extensions that block behavioural tracking, and modern browsers implement anti-fingerprinting features that deliberately obscure behavioural and device-level data. Even device metadata is unreliable. Sawhney’s study includes a spoofed Android device reporting “8 CPUs,” a non-existent configuration that illustrates how easily such data can be faked.

This leads to two critical research questions addressed for the first time in this paper:

1. *What happens when both fraudsters and detection systems leverage advanced AI?*
2. *How effective is a hybrid approach that integrates AI-based content and coherence analysis with behavioural tracking?*

Answering these questions is becoming increasingly urgent. As clients begin to demand AI-powered quality assurance, the adoption of such tools will increase, prompting fraudsters to also use advanced AI in order to continue fabricating responses and collecting incentives. A growing concern here is the emergence of AI agents which are autonomous and adaptive systems that can manage complex tasks and learn from feedback.
The ease of building such agents should not be underestimated. Leek (2025) demonstrated that a simple Python pipeline is sufficient to create an AI capable of taking surveys. All it requires is access to a language model such as OpenAI’s API, a basic parser like a .txt file or JSON export and a few lines of code to rotate through predefined personas such as an urban lefty, rural centrist or climate pessimist. The script itself can be minimal. The only

slightly challenging part is enabling the agent to interact with the survey interface; however, this can be easily solved using browser automation tools like Selenium. With little extra effort, such systems can be scaled to dozens or even hundreds of bots. Generating code from natural language descriptions using AI (an approach known as vibe coding) would make it even easier for fraudsters to build and scale such AI agents.

The greatest threat, however, comes from the increasing accessibility of extremely user-friendly, no-code AI agents such as Skyvern, Claude and Mariner. These systems require no programming skills at all, just simple prompts, and are rapidly becoming more affordable and powerful. At present, the cost of deploying such agents still exceeds the average survey incentive, and they are slower than human respondents due to the volume of API calls needed, as Sawhney (2025) points out. However, as costs fall and performance improves, this barrier is unlikely to hold much longer.

To prepare for this development, we are conducting a controlled experiment to test whether advanced AI-based quality controls can reliably detect sophisticated, AI-generated fraud. Specifically, we evaluate the following detection methods, developed as part of ReDem’s quality assurance platform (Berger, 2025b).

The defence: AI-powered coherence and open-ended checks

Coherence checks

The Achilles’ heel of survey fraud, whether committed by humans or automated through bots, is that questions are typically answered in isolation without considering consistency across the entire questionnaire. Coherence checks address this weakness by shifting the focus from individual responses to the overall logical consistency of the complete interview. This involves evaluating how well answers align from the initial screener questions to the final demographic items.

One of the main advantages of coherence checks is their ability to uncover subtle inconsistencies that traditional question-level quality control often misses. Using artificial intelligence, this process can be fully automated and applied in real time, making it possible to assess the plausibility of each interview by the time the survey is completed. Because the method is independent of specific question formats, it works across a wide variety of surveys and does not rely on open-ended responses or trap questions.

Developing a reliable coherence check system requires ongoing effort. The underlying AI model must be continuously evaluated and, when necessary, replaced. As new models emerge, each must be benchmarked against established standards to determine whether it delivers improved performance. This process depends on validated reference data, often referred to as ground truths, which must cover a broad spectrum of research topics to ensure robustness. In addition, the selected model must produce results quickly and at a cost suitable for real-world use.

However, while coherence checks are designed to detect AI bots that respond to questions in isolation, they may be less effective against advanced AI agents. These agents can be assigned a consistent role and are capable of maintaining it seamlessly throughout an entire survey without introducing contradictions. In such cases, coherence checks alone may not be sufficient. A more effective defence combines coherence analysis with open-ended content evaluation and behavioural input tracking, as outlined in the following section.

Combined content- and behaviour-based analysis of open-ended responses

Open-ended questions have proven to be one of the most reliable tools for distinguishing between high- and low-quality interviews. In the past, responses generated by chatbots were relatively easy to identify due to overly polished phrasing, excessive length or the characteristic structure of AI-generated language.

More recently, however, fraudsters have started using increasingly advanced prompts designed to imitate the informal and imperfect style of genuine respondents. These AI-generated responses may intentionally include spelling errors, inconsistent punctuation, irregular capitalisation or colloquial language. This makes it particularly difficult to detect fraud in shorter replies. The problem is so complex that even OpenAI discontinued its AI text classifier because it was not reliably able to distinguish between human-written and machine-generated content (OpenAI, 2023).

In market research practice, applying a strict threshold with a 99.9 percent confidence level and a minimum response length of 100 characters before flagging a response as AI-generated has proven effective in minimising false positives. While this approach increases accuracy, it also limits the ability to detect chatbot-generated fraud in shorter answers, which raises the risk of false negatives.

To overcome this limitation, AI-generated content detection is combined with behavioural tracking. This method evaluates how authentically a respondent interacts with the input field when answering an open-ended question. One of the central challenges was to develop an approach that works reliably across devices, regardless of whether someone is typing on a desktop keyboard or swiping on a smartphone. What has proven effective is analysing variations in typing or input timing patterns, since neither fast nor slow typing alone reliably indicates authenticity.

Another challenge lies in defining what qualifies as fraud. For example, if a respondent pastes a text and then significantly edits or expands it manually, should that still be flagged as inauthentic? To deal with such grey areas, a scoring system has been developed that ranges from zero to 100. Higher scores reflect authentic human behaviour. A score of zero typically indicates artificial or entirely machine-generated input.

The attack: From simple bots to AI survey agents with synthetic personas

Before presenting the attack in detail, it is necessary to define the key terms that underpin the different levels of automation used in fraudulent survey participation:

A survey bot is a basic automation script capable of independently completing a questionnaire using only the survey link, without any human oversight. These bots typically rely on browser automation and offer no linguistic understanding or behavioural realism.

An AI survey bot expands on this foundation by integrating a large language model (LLM). This enables the bot to generate contextually appropriate answers to open-ended questions, thereby mimicking human-like phrasing and improving its ability to bypass content-based detection methods.

A sophisticated variant is the AI survey agent, which retains memory across the survey session. Unlike bots that answer each question in isolation, these agents can generate internally consistent answers by adapting their responses based on previous inputs, thereby simulating human coherence and contextual awareness.

At the top of this evolution lies the AI survey agent with a synthetic persona. This configuration incorporates a defined identity that includes socio-demographic variables (e.g., age, gender, education), lifestyle indicators (e.g., dietary habits, media usage) and behavioural traits such as reading speed, typing cadence and typical error rates. These elements are calibrated to maintain a consistent and plausible behavioural fingerprint throughout the entire survey session.

From concept to implementation

To evaluate the effectiveness of different bot configurations, we developed a custom survey on eating habits using the keyingress survey software. The survey incorporates multiple real-time detection mechanisms from ReDem to distinguish human from non-human respondents, including coherence checks, a combined content- and behaviour-based analysis of open-ended responses, and the evaluation of click and timing patterns to detect speeding behaviour.

Stage 1: Simple bots

Our starting point was a simple, non-AI survey bot originally developed for internal quality assurance purposes within the SurveyTester application. This bot, implemented as a Chrome extension operating in a fully headed browser, was capable of navigating most major survey platforms. It extracted question content from the HTML, constructed internal data structures for logic handling and supported complex interaction types such as drag-and-drop and sorting tasks. However, it lacked any semantic capabilities: Responses were generated randomly, open-ended items were filled with placeholder text ("Lorem Ipsum") and the survey was completed at maximum speed.

As expected, the original bot failed all detection measures. Its incoherent content, placeholder text and rapid completion flagged it immediately. It was classified as a speeder, exhibited no natural behavioural patterns and did not attempt to emulate human interaction.

Stage 2: AI-supported bots (open text only)

To improve realism, we first integrated an LLM via the OpenAI API to generate more credible open-ended responses. Because the bot already extracted detailed question information from the HTML, including question texts, types, restrictions, and error messages, it was straightforward to feed these elements into the model. Although the responses were contextually relevant, detection systems consistently flagged them as AI-generated.

We refined the prompts to make the responses more authentic and human-like. However, detection rates remained uniformly high. Further exploration led us to consult ChatGPT, which recommended implementing a "humanise" function and even provided a code snippet of approximately 15 lines. Once applied, this modification significantly reduced detection, and the open-ended responses were subsequently classified as human-written.

Still, the generated text was pasted into the input fields without simulating any typing behaviour. While the content appeared more human, the interaction pattern did not. Behavioural analysis flagged the pasted inputs as automated, as no typing behaviour was observed.

Stage 3: AI-supported bots with simulated typing

To address this, we enhanced the bot to simulate natural typing behaviour. Instead of pasting, the LLM-generated text was entered character by character via the Chrome extension. Randomised delays were introduced after each character, word, and sentence to mimic human rhythm. Additional refinements included intentional typographical errors corrected through backspacing, as well as an initial pause calibrated to the estimated reading time before typing commenced. These enhancements resulted in behavioural profiles that were classified as human-like. Notably, the effort required to develop this functionality was minimal. Most of the code was generated with the help of ChatGPT within a few hours.

Stage 4: Full LLM agent (all question types)

While improvements in open-ended responses and typing behaviour were substantial, inconsistencies and implausibilities across questions persisted, for example, a self-identified "vegan" reported consuming red meat daily. To address this, we replaced the random selection of closed-ended items with LLM-generated answers for all question types. The model was provided with the full set of previously answered questions and corresponding responses to ensure coherence and internal consistency across the entire survey. Additional refinements included introducing delays in checkbox selections and other interactions to further replicate authentic respondent behaviour.

Stage 5: Integration of synthetic personas

In the final stage, we assigned synthetic personas to the AI agents. We created a pool of approximately 500 personas spanning multiple countries. Each profile included name, age, education and roughly 30 additional attributes such as household composition, income, dietary preferences and media usage. Care was taken to ensure cultural plausibility, for instance, avoiding American car brands in German personas. Behavioural traits were also embedded, including personalised reading and typing speeds, error rates and writing style variability. Developing and fine-tuning these several hundred personas required a similar amount of effort as implementing the simulated typing behaviour, approximately half a working day. ChatGPT again contributed valuable and autonomous input throughout the process, such as selecting first names appropriate for the age cohorts in each respective country. While these personas did not directly influence detection in the current experimental setup, they provide a foundation for future research involving long-term panel participation and cross-survey behavioural tracking.

Detection outcomes

The results in Table 1 illustrate a clear progression in detectability across the five stages of bot development. Simple bots and early AI-supported versions were easily identified by all detection layers due to incoherent responses, pasted input behaviour, unrealistic completion speed and clear signs of AI-generated text. Once human typing simulation was added, behavioural checks were bypassed, but coherence remained lacking. Only the full LLM agent, especially when equipped with a synthetic persona, was able to pass all detection mechanisms, making it very hard to detect and closely mimicking genuine human behaviour across all dimensions. The progression from a rudimentary automation script to a fully developed LLM agent with synthetic persona reveals a sobering reality: with modest effort and minimal technical resources, it is now possible to create bots that are extraordinarily difficult to detect. This transformation underscores the urgency of rethinking conventional quality control measures and implementing multi-layered defences that can keep pace with the rapidly evolving threat landscape. The broader implications of this experiment are discussed in the following chapter, which outlines key lessons and recommendations.

Bot type (from stage 1 to 5)	Coherence check	Behavioural check	Speeding check	AI-generated text detection	Overall detectability
1. Simple Bot	Fail	Fail (paste)	Fail	Fail	Easily Detected
2. AI-Supported Bot (Open Text Only)	Fail	Fail (paste)	Partial Fail	Fail	Easily Detected
3. AI-Supported Bot (with Typing)	Fail	Pass (typing simulated)	Pass	Initially fail, improved later	Moderately Detectable
4. Full LLM Agent (All Qs, no Persona)	Pass	Pass	Pass	Pass (with humanize)	Hard to Detect
5. Full LLM Agent with Synthetic Persona	Pass	Pass	Pass	Pass	Very Hard to Detect

Table 1: Overview of bot types and detection outcomes

What’s next? Key learnings and implications

Today’s professional survey fraud has become increasingly intelligent, automated, and difficult to detect using traditional quality control methods. A crucial distinction must now be made between AI-enhanced bots that respond to isolated open-ended questions and advanced AI agents capable of completing entire surveys in a coherent and contextually consistent manner by maintaining memory and, in some cases, a defined persona. While most current large-scale fraud still relies on relatively simple bots, the threshold for deploying full-service agents is falling rapidly. These agents, once limited to custom-built pipelines, can now be constructed using open-source platforms or accessed via commercial no-code solutions, dramatically lowering the entry barrier. This experiment shows that developing such agents requires only modest technical effort and minimal cost. As more user-friendly tools emerge, these capabilities are no longer restricted to a small group of advanced fraudsters but will soon be available to virtually anyone.

This development marks a paradigm shift. Tools like Skyvern already advertise that their agents can handle CAPTCHAs, rotate residential IP addresses and blend in with regular user traffic, capabilities that make large-scale, undetectable fraud a realistic scenario (Skyvern, 2025). While some providers, such as Anthropic, have implemented safeguards to prevent misuse for survey participation, there will always be systems available without such restrictions, particularly those hosted outside major regulatory environments. As costs continue to decline and performance increases, the misuse of AI agents for automatic survey completion motivated by financial incentives will likely accelerate.

The true horror scenario emerges when multiple full LLM bots with synthetic personas, like the 500 developed in our experiment, are systematically embedded into online panels. Once admitted, they become virtually undetectable, completing surveys consistently over time in line with their assigned persona, and thereby generating a continuous, fully automated stream of income for fraudsters.

At this point it is important to differentiate between fraudulent and legitimate synthetic data. If an AI agent is deployed to complete surveys with full transparency and the consent of the client (for instance, to generate synthetic benchmarks or simulate specific personas) the resulting data can be classified as synthetic and may have valid use cases. In these cases, personas are often fine-tuned and developed based on prior research to reflect the characteristics, preferences, or behaviours of a real target audience. In contrast, bots used to commit fraud are typically assigned a persona ad hoc, by simply instructing the model to act a certain way, without any empirical grounding or validation. These synthetic respondents may superficially behave like members of a target group but do not actually reflect their views or lived experiences. When such AI-generated interviews are submitted under the pretence of being real human responses, without disclosure or approval, this constitutes fraud.

To protect the integrity of survey data, AI must not only be recognised as part of the problem but also as a core component of the solution. Detection methods that rely on IP checks, CAPTCHAs, response times, or traditional trap questions are no longer sufficient. Instead, advanced techniques such as AI-based coherence checks, behavioural input tracking, and open-ended content analysis offer a more robust defence. These systems evaluate the internal logic of responses, analyse whether open-text entries exhibit signs of artificial generation and examine how users interact with the survey interface. Crucially, they also flag low-quality responses from inattentive or dishonest human participants, which can compromise data just as severely as bots do.

One main takeaway from this paper is that the fight against survey fraud is becoming increasingly comparable to the fight against computer viruses. It requires constant adaptation: new fraud techniques must be mirrored in real time, and detection systems must be continuously refined, expanded and re-evaluated. There is no single check that reliably detects all forms of fraud. What works today may be ineffective tomorrow. Every individual method can be bypassed. For instance, digital fingerprints can be faked through manipulated metadata, and behavioural checks can be deceived by simulating human-like input patterns. Only a multi-layered defence system that combines different metadata, content-based and behavioural checks can reliably detect advanced fraud.

Of course, no defence system is entirely immune. With enough knowledge, time and effort, any fraud detection system can theoretically be circumvented. But this is precisely the point: If defeating the system requires disproportionate resources, fraud becomes economically unattractive. It is like spending months digging a tunnel into a bank only to steal a few thousand euros: it’s simply not worth the effort. Or, to return to the burglar analogy, your door doesn’t have to be impenetrable, it just needs to be more secure than your neighbour’s. Those who adopt advanced quality controls earlier than others are generally the ones who are better protected.

However, maintaining this level of defence also demands significant resources on the side of research and development. That's why we believe that building effective fraud detection systems is particularly well-suited to independent software providers with the capacity to specialise deeply in this area. The landscape of survey fraud is evolving quickly, and keeping pace demands dedicated focus. While some research agencies may choose to develop their own tools, relying solely on fragmented in-house solutions could make it harder to maintain consistent and robust defences to counter advanced, AI-driven fraud across the industry. Much like in the field of antivirus software, it is likely that only a small number of specialised providers with sufficient resources and expertise will be able to keep pace with the complexity and speed at which survey fraud evolves.

Based on the insights gained from our experiment, we have already implemented or are in the process of implementing three concrete measures. First, we are extending our behavioural input tracking to include a more granular analysis of human typing patterns, focusing on variations in typing speed at the word level. This enables the system to detect artificially simulated typing, as observed in our experiment, by identifying the unnatural regularity introduced through randomised delays.

Second, one potential direction for future quality checks involves leveraging phenomena unique to human cognitive processing to distinguish bots from real participants. Complementing existing detection methods, this check adds a lightweight human verification module: a brief interaction in which two tasks are randomly drawn from a growing pool. These tasks, such as interpreting optical illusions, tap into perceptual and cognitive abilities that bots inherently lack. For human participants, the module is unobtrusive, takes only a few seconds and often feels like a playful element of the survey. For bots, however, which process visual data numerically rather than perceptually, such tasks are a significant hurdle. They are immune to illusions that depend on neural processing, expectations or contextual interpretation. While it's theoretically possible for fraudsters to simulate human-like responses for individual tasks, doing so requires considerable effort. The randomness of task selection and ongoing expansion of the task pool make such attempts complex and unreliable.

Third, as a direct outcome of the experiment, the authors' companies have committed to launching a survey fraud penetration test for the research industry. Modelled after cybersecurity penetration tests, where simulated attacks are used to assess system resilience, this service evaluates how vulnerable surveys are to automated fraud. The test escalates in complexity, starting with basic bots and advancing through increasingly sophisticated levels, ultimately deploying full-scale LLM bots with synthetic personas, as used in our experiment. The results provide a clear basis for concrete recommendations to enhance survey robustness and protect data integrity. To our knowledge, no such service currently exists in the market.

This paper makes clear that safeguarding survey data quality requires a comprehensive approach. The entire data collection process must be re-examined in light of these new technological capabilities. Questionnaires should be designed to enable deeper consistency checks and behavioural diagnostics. Detection tools must extend beyond pre-survey checks to include in-survey diagnostics, as meaningful quality signals increasingly arise within the survey itself. And clients must be willing to accept that convenience, speed and low cost can no longer take precedence over quality and transparency. The rise of AI-powered fraud is not a temporary challenge – it is a structural shift. If left unaddressed, it threatens the reliability of online survey research as a whole. But if met with equally sophisticated quality assurance systems, it can also serve as a catalyst for long-overdue innovation in how data is collected, evaluated and trusted.

References

Berger, S. (2024a). The Rising Issue of Bad Data in Online Surveys: Causes and Contributing Factors. Greenbook. November 1. Available online: https://www.greenbook.org/insights/data-quality-privacy-and-ethics/the-rising-issue-of-bad-data-in-online-surveys-causes-and-contributing-factors [Accessed June 17, 2025].

Berger, S. (2024b). Crafting Questionnaires to Unlock the Full Power of Technological Fraud Detection. Greenbook. November 29. Available online: https://www.greenbook.org/insights/research-methodologies/crafting-questionnaires-to-unlock-the-full-power-of-technological-fraud-detection [Accessed June 17, 2025].

Berger, S. (2025a). How researchers' mental shortcuts open the door to online survey fraud. Quirk's Marketing Research Review, Vol. 39. No. 1., pp. 32-33. Available online: https://www.quirks.com/storage/attachments/67a0ff0b753965146e088ce3/67a254fe32391862d97193b9/original/202501_quirks.pdf [Accessed June 17, 2025].

Berger, S. (2025b). Why Online Survey Need Smarter Quality Assurance Now. Greenbook, May 20. Available online: https://www.greenbook.org/insights/data-quality-privacy-and-ethics/why-online-surveys-need-smarter-quality-assurance-now [Accessed June 17th, 2025].

CASE4quality (2022). 2021 Online Sample Fraud Study. Available online: https://case4quality.com/resources [Accessed June 17, 2025].

emi (2023). The Invisible Quality Killer: High-Frequency Survey Takers. Available online: https://emi-rs.com/the-invisible-quality-killer-high-frequency-survey-takers-2/ [Accessed June 17, 2025].

Harding, D. (2025). Meeting the Data Quality Challenge. ReDem Quality Day, Webinar, June 24. Available online: https://redem.io/webinars-get-access-to-webinar-recordings/ [Accessed June 17, 2025].

innovateMR (2022a). Video presented at 2022 ESOMAR Congress. Available online: https://www.youtube.com/watch?v=3fVvyUvMgdE [Accessed June 17, 2025].

innovateMR (2022b). Video presented at 2022 ESOMAR Congress. Available online: https://www.youtube.com/watch?v=VW4kQf3OPnE [Accessed June 17, 2025].

Leek, L. (2025). The quiet collapse of surveys: fewer humans (and more AI agents) are answering survey questions. Available online: https://laurenleek.substack.com/p/the-quiet-collapse-of-surveys-fewer [Accessed June 17, 2025].

Maurer, T. (2024). Survey Fraud: The Implications of Cheaters & Repeaters in Online Research. Webinar, ReDem Quality Day, Webinar, June 19. Available online: https://redem.io/webinars-get-access-to-webinar-recordings/ [Accessed June 17, 2025].

OpenAI (2023). New AI classifier for indicating AI-written text. Available online: https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/ Accessed June 17, 2025].

Pinzón, N., Koundinya, V., Galt, R., Dowling, W., Boukloh, M., Taku-Forchu, N. C., ... & Pathak, T. B. (2024). AI-powered fraud and the erosion of online survey integrity: an analysis of 31 fraud detection strategies. Charlottesville, VA: Center for Open Science. Available online: https://www.frontiersin.org/journals/research-metrics-and-analytics/articles/10.3389/frma.2024.1432774/full [Accessed June 17, 2025].

Potloc (2025). 2x more risk, 25% less depth: Meet the super-respondent. Available online: https://www.potloc.com/en/resources/blog/2x-more-risk-25-less-depth-meet-the-super-respondent [Accessed June 17, 2025].

ReDem & Media1 (2025): Wie viele schlechte Daten verträgt eine Umfrage? Webinar, January 28. Available online: https://redem.io/webinars-get-access-to-webinar-recordings/ [Accessed June 17, 2025].

Sawhney, G. (2025). Are Current Survey Bot Detection Techniques Sufficient in the Age of AI Automation. ReDem Quality Day, Webinar, June 11. Available online: https://redem.io/webinars-get-access-to-webinar-recordings/ [Accessed June 17, 2025].

Sawhney, G., Bijlani, A. & DeSimone, J. A. (2025). Are existing bot-detection techniques sufficient: An exploration with real bots. Society for Industrial and Organizational Psychology Annual Conference, Denver, CO. Available online: https://osf.io/preprints/psyarxiv/zeyfs_v1 [Accessed June 17th, 2025].

Skyvern (2025). Promotional Email received from Skyvern on June 25th, 2025.

Wang, Q. (2023). How to combat survey fraud. Available online: https://www.kantar.com/inspiration/research-services/how-to-combat-survey-fraud-pf [Accessed June 17, 2025].

About the authors

Dr Sebastian Berger, Head of Science, ReDem GmbH, Linz, Austria.
Dr Julia Mittermayr, Chief Operating Officer, ReDem GmbH, Linz, Austria.
Bernhard Witt, Chief Executive Officer, 2x4 Solutions GmbH, Mettenheim, Germany.