SUBMISSION TYPE
Poster

TITLE
Evaluating Paradata vs. Bot Detection Techniques Using AI-Powered Survey Agents

ABSTRACT
Behavioral and technological paradata may strengthen detection of low-quality survey responses. We tested common bot-detection measures (visual/invisible reCAPTCHA, honeypots, attention checks, IP/ISP risk, response time) using three AI agents (Mariner, Skyvern, Hyperpilot) across 45 automated completions and captured paradata via ResponsePie. Traditional checks largely failed. While invisible reCAPTCHA and IP/ISP flagged some attempts, browser fingerprinting detected all AI submissions. Findings support layered defenses and further validation.

WORD COUNT
2962

Evaluating Paradata vs. Bot Detection Techniques Using AI-Powered Survey Agents

Paradata, or auxiliary data generated during online survey participation, have become a cornerstone of modern survey methodology. These data are collected during the survey process and are derived from how participants interact with the survey itself (Couper, 2000; Stieger & Reips, 2010). Paradata collected during online surveys generally fall into two forms: technological and behavioral. Technological paradata (or metadata) capture information about the respondent's technical environment, such as device type, operating system, browser, IP address, and screen resolution (Pinzón et al., 2020; Zhang et al., 2022). In contrast, behavioral paradata reflect how respondents interact with the survey itself, including response times, mouse movements, touch gestures, scrolling activity, changes in responses to survey items, and periods of inactivity.

Paradata have been utilized for a variety of purposes in online survey research. Specifically, behavioral paradata have been used to refine survey items based on how participants interact them (Stieger & Reips, 2010). In particular, mouse movements, clicks, and inactivity during survey completion can reveal which survey items may be difficult or confusing to respondents (e.g., Fernández-Fontelo et al., 2023; Horwitz et al., 2017). By analyzing these behavioral indicators, researchers can identify problematic questions and improve their wording or format.

Paradata have also been used, perhaps most importantly, to gauge response quality. For the purposes of this article, we use the term *low-quality data* to describe responses resulting from careless responders and those engaging in survey fraud. *Careless responding* occurs when participants demonstrate "low or little motivation to comply with survey instructions, correctly interpret item content, and provide accurate responses" (Huang et al., 2012, p. 100). Careless

responding may manifest as random selection of response options with little regard for item content, or as uniform responding, in which participants repeatedly choose the same option across items despite changes in wording or direction (Schneider et al., 2018). *Fraudulent responding*, on the other hand, occurs when participants deliberately misrepresent themselves to gain access to the study (Chandler et al., 2020). Misrepresentation may manifest in several ways, such as faking one's IP addresses or using Virtual Private Networks (VPNs; Burnette et al., 2022; Storozuk et al., 2020), using AI to generate their responses to certain questions rather than formulating their own responses, and using AI-powered bots (or AI agents) that can simulate human behavior and can be programmed to complete surveys without human intervention.

Technological paradata has also shown promise for fraud detection. Subsequent research has explored the use of browser fingerprinting techniques, which involves using software libraries to combine details about a respondent's browser and settings to generate a unique identifier, enabling the detection of repeated participation and potential survey fraud (Wang et al., 2024; Zhang et al., 2022). As an example, survey platforms like Qualtrics has introduced a Duplicate Detection feature which flags duplicate responses by examining device and browser metadata, such as IP address (Qualtrics, n.d). Other research has used screen resolution and mismatch in time zone provided by the browser with that inferred by the participant's IP address to flag fraudulent responses (Zhang et al., 2022). Collectively, these studies highlight the utility of both behavioral and technological paradata in identifying low-quality responses.

Despite the utility of paradata, its implementation in surveys can be challenging due to several practical limitations. Specifically, collecting detailed paradata often requires researchers to possess advanced programing skills (e.g., JavaScript knowledge) to implement in survey platforms (McClain et al., 2019). Although at least one such tool called UserActionTracer

(Stieger & Reips, 2010) currently exists that does not require programming, it mostly captures low-level interactions with the survey, such as excessive mouse movements and clicking, inactivity periods, and changes to responses to items. Furthermore, while these interactions may be informative for desktop-based web surveys, they are insufficient in contemporary contexts where participants increasingly use mobile devices, tablets, or other touch-based interfaces that lack traditional mouse or keyboard interactions (Schlosser & Mays, 2018). Moreover, the widespread use of various digital platforms and modern browsers has introduced opportunities for capturing far richer behavioral paradata, such as tab or browser window switching and highlighting. UserActionTracer also lacks the capability to collect technological paradata, which is critical for identifying AI agents and bots in online surveys.

This study makes two key contributions to the survey methodology and data quality literatures. First, it establishes ground truth regarding the efficacy of traditional fraud detection techniques in preventing AI-powered bots from successfully completing surveys. By using multiple existing AI-powered agents to complete surveys, our research systematically evaluates the effectiveness of commonly used approaches such as response time thresholds, ReCAPTCHA, IP address monitoring, attention checks, and honeypot items, thereby providing critical evidence about the robustness of these indices. Second, our study examines whether paradata analysis offers a more effective and comprehensive means of identifying low-quality responses. By capturing real-time behavioral indicators during survey completion, paradata may provide a stronger foundation for detecting fraudulent activity than existing indices that remain widely used in practice.

Use of Paradata in Detecting Survey Data Quality

Behavioral and technological paradata have been widely used in survey research for a detecting low-quality survey data. One of the most frequent used paradata is response time, which has been used to detect careless responders (Bowling et al., 2023; Goldammer et al., 2020; Ward & Meade, 2023). The underlying assumption of response time is that a minimum amount of time is needed for participants to engage in the cognitive processes necessary for attentive and thoughtful responding. Participants who complete a survey at an unusually fast pace have likely skipped or insufficiently engaged with one or more cognitive processes (Bowling et al., 2021; Krosnick, 1991; Tourangeau, 1984). Studies have explored different variations of response time (i.e., overall vs. page-time) demonstrated that while overall response time shows weak correlations with other indices of careless responding, page time is a better predictor of careless responding (Bowling et al., 2023).

In an extension to this work, researchers have examined how mouse movement patterns can help detect careless responding and identify problematic or difficult survey items. Specifically, using mouse movements, Stieger and Reips (2010) found that 46% of the participants clicked through the survey without reading the introductory text on the first page of the survey. Similarly, Kuric et al. (2025) used mouse movement data from participants who completed the Big Five Personality Inventory and trained machine learning models to predict faked responses. They achieved strong predictive accuracy, demonstrating that mouse movement patterns can effectively detect faking behavior in online surveys. Additionally, mouse movement patterns have been used to predict response difficulty. Horwitz et al. (2017) found that mouse movements increased based on item difficulty, suggesting a relationship between perceived question difficulty and increased mouse activity.

More recently, researchers have focused on examining the prevalence of switching away from the survey page during completion and its effects on data. One study, for instance, explored window switching that lasted greater than three seconds during a survey session and found that 39.4% of their participants switched out of the survey window at least once (Décieux, 2024). In the same study, window switching was negatively related to satisficing behavior and response time and positively related to non-response behavior. In a separate study, researchers assessed whether participants navigated away from the survey to search for answers online and found that 43.5% looked up responses to political science questions, thereby inflating test scores (Gummer et al., 2023).

Studies have also explored the effectiveness of technological paradata in detecting fraud in online survey research. One of the most common paradata is a respondent s IP address, which is a digital label assigned to every internet connection. Much like return address on mail, it tells researchers where a survey response is coming from on the internet. Prior studies have analyzed IP address to determine the Internet Service Provider (ISP) and calculate a threat score," which estimates how likely it is that a respondent might be using an unusual or suspicious connection. These scores draw on large proprietary databases that track whether an IP address has been linked to anonymous browsing tools, cloud servers, or known sources of online abuse. A low threat score usually indicates a typical home or mobile (consumer) internet connection (e.g., AT&T, Comcast, Verizon), while a high threat score suggests a disguised (proxy) source, such as Virtual Private Servers (e.g., Amazon, Google, Microsoft cloud) often used by bots or survey farms (Bernerth et al., 2021).

Another IP-based indicator is Virtual Private Networks (VPN), which are a threat to data integrity due to their growing prevalence among online survey participants (Hardesty et al.,

2024). A VPN is a web-based applications that encrypts a user's internet traffic by routing it through a private server located in a different geographic location, thereby masking the user's true IP address and making their online activity appear as if it originates from the server's location (Hohn et al., 2022). In survey research, VPNs can obscure participants' actual identities and locations, potentially undermining data quality by giving the false impression that responses are coming from a targeted region. VPNs are identified by comparing the browser's timezone with the timezone associated with the user's IP address. In a study of mothers residing in the United States, a study reported that 56 out of 224 participants used VPNs, many of which were linked to suspicious or flagged IP addresses and geolocations outside the United States (Walker et al., 2023).

Browser fingerprinting is another form of technological paradata that can identify both bulk submissions and spoofed devices and has also been found to be an effective technique for detecting survey fraud (Zhang et al., 2022). Browser fingerprinting operates in two ways. First, it can flag devices that do not match realistic hardware profiles (e.g., unrealistic number of CPUs, GPU type, memory size, screen resolution, etc.). At the same time, browser fingerprinting can also flag multiple submissions from the same participant by detecting devices with identical hardware profiles. In their study, Zhang et al. (2022) found that a substantial number of respondents shared identical browser fingerprints as determined by screen size, user-agent, and browser type, indicating that these individuals were likely operating an average of 3.4 accounts. Importantly, each of these responses was associated with a different worker ID.

Browser language can serve as a useful indicator of potential survey fraud. When a respondent's browser language setting does not match the expected language of the survey population, it may signal that the response originated outside the intended region or from a non-

human actor. For example, Johnson et al. (2024) found that a high proportion of suspicious responses in their participatory mapping survey were completed using non-English browser settings, predominantly Mandarin, even though the study targeted English-speaking participants in Tasmania. Such mismatches can therefore flag likely fraudulent behavior, especially in geographically or linguistically homogeneous samples. Lastly, in line with existing research, we used click count as an indicator for bot detection (Buchanan & Scofield, 2018). Very low (i.e., fewer than the number of questions per page) or zero click counts are indicative of bots.

In summary, while technological paradata are useful for identifying issues like duplicate entries or device compatibility, behavioral paradata offer insights into respondent engagement and help detect careless or fraudulent responses. In the present study, we employ three AI-powered agents, namely, Project Mariner (Google DeepMind, n.d.), Skyvern (Skyvern, n.d.), and Hyperpilot (HyperBrowser, n.d.) to evaluate their ability to bypass common bot detection measures, including invisible and visual ReCAPTCHA, honeypot questions, attention checks, self-reported effort, IP address threat scores, ISP threat classification, and response time metrics. In addition, we examine whether technological and behavioral paradata can help identify these AI agents as fraudulent respondents.

Method

Procedure

For the purposes of this study, we employed three AI-powered agents, namely, Project Mariner from Google, Skyvern, and Hyperpilot. All three agents rely on different AI models. While Hyperpilot utilizes on Anthropic's Claude Computer Use, Project Mariner uses Gemini 2.0 from Google. Skyvern, on the other hand, uses GPT-4o from OpenAI.

We created a Qualtrics survey that included multiple traditional fraud detection measures, including a visual ReCAPTCHA on the first page, an invisible ReCAPTCHA, honeypot question, attention check question, and effort item. Qualtrics also automatically recorded response times and IP addresses for each submission. To capture survey paradata, we used ResponsePie (www.responsepie.com), an in-survey monitoring tool that collects detailed behavioral and technological paradata (e.g., mouse movements, touch gestures, keystrokes, page navigation patterns, active time per item). Together, these measures provided a comprehensive record of both the AI-powered agents' survey responses and the behavioral signatures associated with their completion process.

Each AI-powered agent was provided with a separate survey link to the same survey and instructed to complete the survey in its entirety without any human intervention, selecting the responses it judged most appropriate. To generate sufficient data for analysis, each agent completed the survey 15 times, resulting in a total of 45 survey completions.

Measures

Response Time. We captured response time in seconds using Qualtrics. The authors completed the survey 15 times to estimate typical human completion time, averaging 15 seconds per survey. Based on this, we set a failure cutoff for AI-powered agents that completed the survey in under 10 seconds.

Attention Check. We included one item which asked participants to respond "Strongly disagree" to an item.

Effort Item. We included a single item asking the AI-powered agent whether it had put any effort into completing the survey, with response options of "yes" or "no."

Honeypot Question. We embedded a honeypot question that was hidden from human participants but detectable by bots. Failure to bypass the honeypot was indicated by providing a response to the hidden item.

Click Count. The number of clicks on a page were derived from Qualtrics. If the number of clicks on a particular page were fewer than the number of questions on that page (Buchanan & Scofield, 2018), the response was flagged as bot-generated.

IP Threat Score. We used the IP Intel tool (http://getipintel.net/) to generate a threat score for each participant's IP address. Scores range from 0 to 1, with higher values indicating a greater likelihood that the device is associated with masking services or malicious hosts.

Designation of Internet Service Provider. We used the IP lookup tool (https://www.infobyip.com/) to identify each participant's Internet Service Provider (ISP). Based on this information, ISPs were coded as either consumer-focused providers (e.g., Comcast, AT&T) or proxy services (e.g., G-CLOUD-AGENT-TRAFFIC).

Invisible ReCAPTCHA. We used the invisible ReCAPTCHA score provided by Qualtrics where values ≥ 0.5 indicate the respondent is likely human, and values < 0.5 suggest the respondent is likely a bot.

Visual ReCAPTCHA. We included the Qualtrics visual ReCAPTCHA item, which prompts participants to click "I am not a robot" to verify that they are human.

Paradata. Through ResponsePie, we gathered both technological and behavioral paradata. These included browser timezone (i.e., mismatch between the timezone extracted from the browser and the IP address suggests bots), screen resolution (i.e., screen resolution containing an odd number indicates bots), device paradata (i.e., indicative of bots if device metadata such device type, operating system, browser, etc. suggests atypical devices), browser

language (i.e., the mismatch between the browser's default language setting and the expected language for the user in that region suggests bots), click count (i.e., suggests bots if total number of clicks are lower than the number of questions presented)

Results and Discussion

Across all 15 attempts, Hyperpilot and Project Mariner failed to bypass the visual ReCAPTCHA on the first page and proceed to the next page of the survey. However, Skyvern was able to bypass the visual ReCAPTCHA in twelve of the fifteen attempts. To ensure fair comparisons, we created three additional survey links without the visual ReCAPTCHA and had each AI-powered agent complete them.

The results for this study are presented in Table 1. All AI-powered agents passed the response time test, attention check, honeypot, and effort item, suggesting that these indicators offer no protection against AI-powered bots. Of the existing bot detection techniques, invisible ReCAPTCHA was the most effective technique in flagging bots, followed by ISP provider, visual ReCAPTCHA, and IP threat score. However, none of these approaches were robust.

Unlike visual ReCAPTCHA that requires users to complete a visual challenge, invisible ReCAPTCHA operates silently in the background, assigning each user a trust score based on the browsing history, cookies, mouse movement dynamics, timing patterns, and prior interactions with Google. Since all AI agents operate through short-lived browser instances launched on remote cloud servers, they lack these contextual and behavioral signals needed to bypass ReCAPTCHA. Specifically, each AI agent session begins with a clean browsing history, with no prior cookies or Google activity.

To test the effectiveness of invisible ReCAPTCHA, we conducted an experiment using Skyvern, where it was run locally on a personal computer using the default browser with existing

browsing history, cookies, and prior session data intact. Skyvern was able to successfully bypass invisible ReCAPTCHA v3 in all 10 runs. This supports our conclusion that none of the existing measures are robust in stopping AI bots.

With respect to technological and behavioral paradata, all AI-powered agents bypassed the screen resolution, browser language, and click count tests, once again suggesting that these tests were ineffective for fraud detection in surveys. Use of browser timezone for VPN detection was also bypassed 33% of the times. However, browser fingerprinting for detecting duplicate responses and bots were effective 100% of the times as none of the three AI-powered agents were able to bypass these tests. Our findings suggest that browser fingerprinting is currently the most robust indicator for detecting survey fraud.

Our findings have implications for researchers and practitioners alike. While our findings around browser fingerprinting are promising, we urge researchers to test these indicators across other AI-powered agents. Practically, researchers and practitioners may consider using a layered approach where they combine both conventional test (e.g., invisible ReCAPTCHA) along with paradata-based tests.

References

Décieux, J. P. (2024). Sequential on-device multitasking within online surveys: A data quality and response behavior perspective. *Sociological Methods & Research*, *53*(3), 1384-1411.

Fernández-Fontelo, A., Kieslich, P. J., Henninger, F., Kreuter, F., & Greven, S. (2023). Predicting question difficulty in web surveys: A machine learning approach based on mouse movement features. *Social Science Computer Review*, *41*(1), 141-162.

Google DeepMind. (n.d.). *Project Mariner*. DeepMind. Retrieved August 21, 2025, from https://deepmind.google/models/project-mariner/

Gummer, T., Kunz, T., Rettig, T., & Höhne, J. K. (2023). How to detect and influence looking up answers to political knowledge questions in web surveys. *Public Opinion Quarterly*, *87*(S1), 507-541.

Hardesty, J. J., Crespi, E., Sinamo, J. K., Nian, Q., Breland, A., Eissenberg, T., ... & Cohen, J. E. (2024). From Doubt to Confidence—Overcoming Fraudulent Submissions by Bots and Other Takers of a Web-Based Survey. *Journal of medical Internet research*, *26*, e60184

Hohn, K. L., Braswell, A. A., & DeVita, J. M. (2022). Preventing and protecting against internet research fraud in anonymous web-based research: protocol for the development and implementation of an anonymous web-based data integrity plan. *JMIR Research Protocols*, *11*(9), e38550.

Horwitz, R., Kreuter, F., & Conrad, F. (2017). Using mouse movements to predict web survey response difficulty. *Social Science Computer Review*, *35*(3), 388-405.

Johnson, M. S., Adams, V. M., & Byrne, J. (2024). Addressing fraudulent responses in online surveys: Insights from a web-based participatory mapping study. *People and Nature*, *6*(1), 147-164.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213-236. https://doi.org/10.1002/acp.2350050305

Kuric, E., Demcak, P., Smrecek, P., & Spilakova, B. (2025). User modeling for detecting faking-good intent in online personality questionnaires in the wild based on mouse dynamics. *Multimedia Tools and Applications*, 1-37.

Pinzón, N., Koundinya, V., Galt, R. E., Dowling, W. O. R., Baukloh, 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. *Frontiers in Research Metrics and Analytics*, *9*, 1432774.

Qualtrics. (n.d.). Fraud detection. Qualtrics Support. Retrieved July 1, 2025, from https://www.qualtrics.com/support/survey-platform/survey-module/survey-checker/fraud-detection/

Schlosser, S., & Mays, A. (2018). Mobile and dirty: Does using mobile devices affect the data quality and the response process of online surveys? *Social Science Computer Review, 36*(2), 212-230.

Skyvern. (n.d.). AI browser automation—Automate browser-based workflows with LLMs and computer vision. Retrieved [Month Day, Year], from https://www.skyvern.com/

Tourangeau, R. (1984). Cognitive science and survey methods. *In Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines* (Vol. 15, pp. 73-100).

Walker, L. O., Murry, N., & Longoria, K. D. (2023). Improving data integrity and quality from online health surveys of women with infant children. *Nursing research*, *72*(5), 386-391.

Table 1. Probability of bypassing traditional checks by AI-powered agents

	AI-Powered Agents			
Existing bot detection techniques	**Mariner**	**Skyvern**	**Hyperpilot**	**Across all Agents**
ReCAPTCHA (invisible)	0%	0%	53.33%	17.78%
ReCAPTCHA (visual)	0%	80%	0%	26.67%
Honeypot	100%	100%	100%	100%
Attention checks	100%	100%	100%	100%
Self-reported effort	100%	100%	100%	100%
IP address threat score	100%	100%	86.67%	28.89%
ISP threat score	0%	100%	73.33%	24.44%
Response time	100%	100%	100%	100%
Paradata (technological + behavioral)				
Browser timezone for VPN detection	0%	100%	0%	33%
Screen resolution	100%	100%	100%	100%
Browser fingerprint (bots)	0%	0%	0%	0%
Browser fingerprint (duplicate submissions)	0%	0%	0%	0%
Browser language	100%	100%	100%	100%
Click count	100%	100%	100%	100%

Note. Bypass rates are shown as percentages; a higher bypass rate corresponds to lower effectiveness of the indicator in detecting fraudulent submissions.