Collecting Online Data in the Age of Large Language Models

Bots, LLMs & the Online Data Crisis

Presented by Dr. Blair R K Shevlin

Center for Computational Psychiatry

June 2026

Why Are We Talking About This?

Scale of the problem

Online surveys are the backbone of psychology, public health, economics & political science — but data quality is collapsing

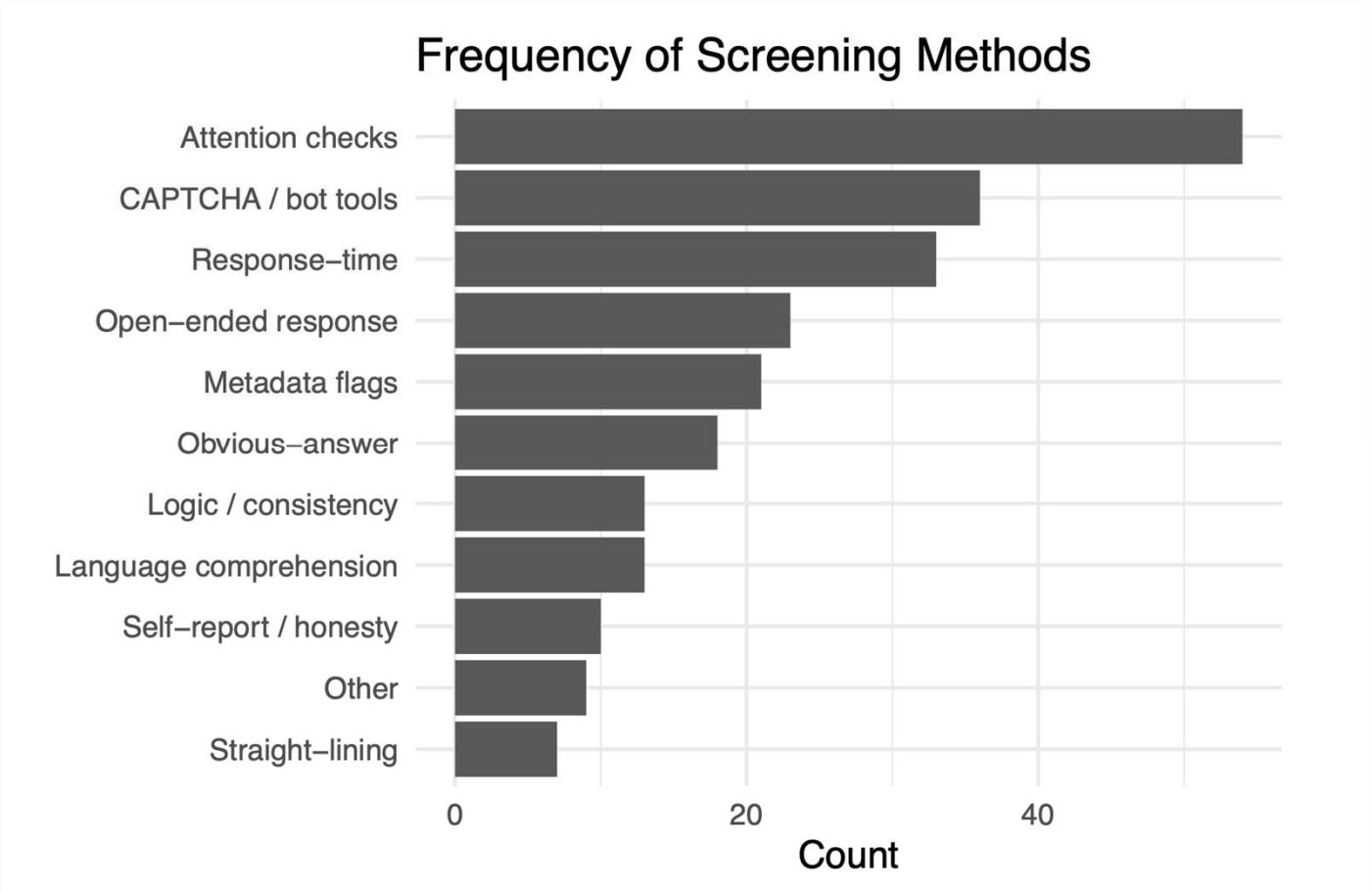
Bots got smarter

Old-school bots failed attention checks. LLM-powered bots pass 99.8% of them while adjusting their writing to match assigned demographics

Our research is at stake

If synthetic respondents are already in panels, results we've collected and published may already be contaminated — or will be soon

What tools do researchers use?

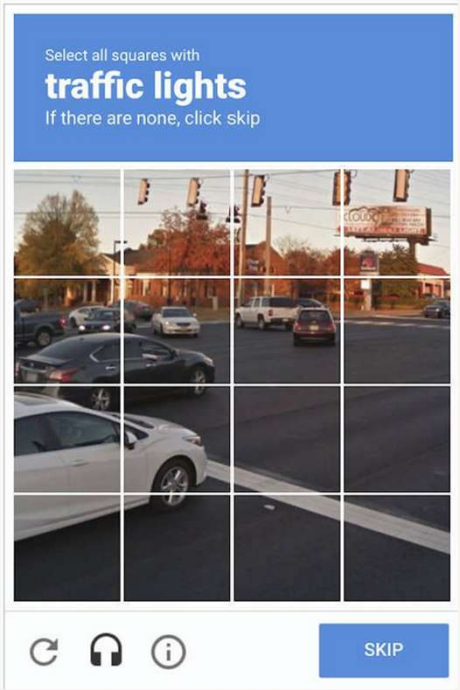


Anonymous survey from Society for Judgement and Decision-Making in 2025

What tools do researchers use?

CAPTCHA

Tests whether a respondent can perform a simple human task (like identifying objects in an image)



Oppenheimer et al., 2009; Perfecto & O'Donnell, 2025; Byrd, 2025

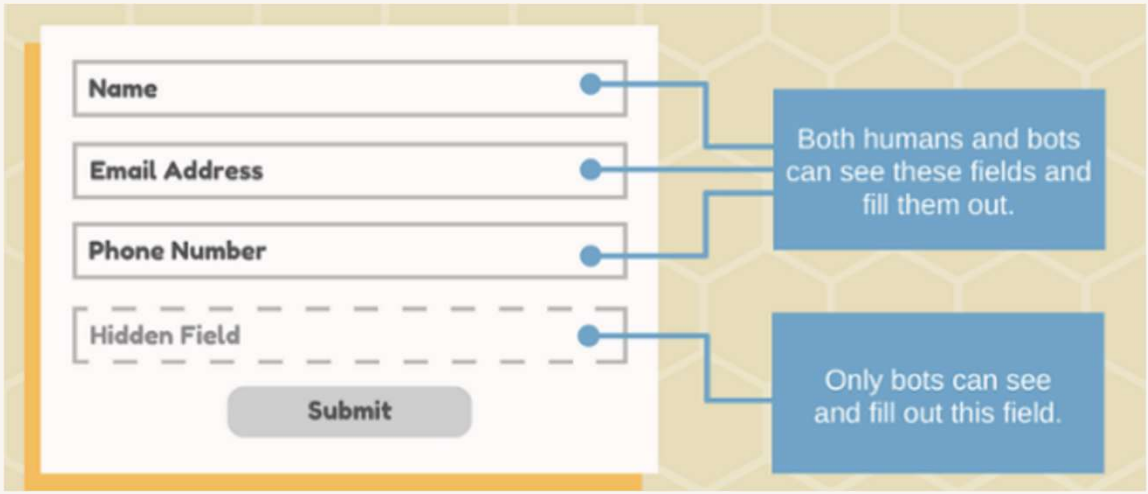
What tools do researchers use?

CAPTCHA

Tests whether a respondent can perform a simple human task (like identifying objects in an image)

HONEYPOTS

Hidden fields only bots can “see” and fill, flagging automated submissions



Oppenheimer et al., 2009; Perfecto & O’Donnell, 2025; Byrd, 2025

What tools do researchers use?

CAPTCHA

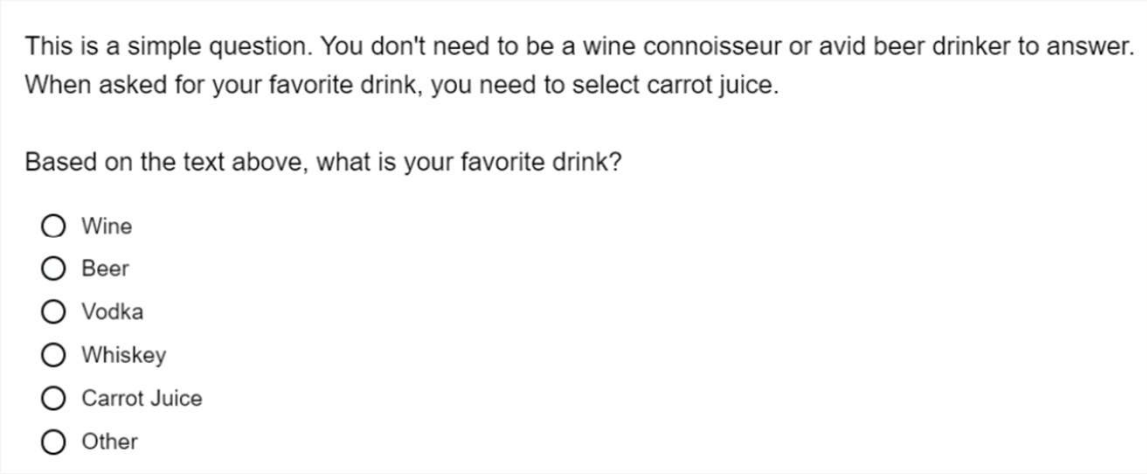
Tests whether a respondent can perform a simple human task (like identifying objects in an image)

HONEYPOTS

Hidden fields only bots can “see” and fill, flagging automated submissions

ATTENTION CHECKS

Instructions or logic questions to catch inattentive or automated responders



Oppenheimer et al., 2009; Perfecto & O’Donnell, 2025; Byrd, 2025

What tools do researchers use?

CAPTCHA

Tests whether a respondent can perform a simple human task (like identifying objects in an image)

HONEYPOTS

Hidden fields only bots can “see” and fill, flagging automated submissions

ATTENTION CHECKS

Instructions or logic questions to catch inattentive or automated responders

METADATA

Behavioral clues that can reveal nonhuman or duplicate participation (e.g., response time, IP addresses, device type)

This question will not be displayed to the recipient.

Browser	Chrome
Version	139.0.0.0
Operating System	Macintosh
Screen Resolution	1920x1080
Flash Version	-1
Java Support	0
User Agent	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/139.0.0.0 Safari/537.36

Oppenheimer et al., 2009; Perfecto & O’Donnell, 2025; Byrd, 2025

What tools do researchers use?

CAPTCHA

Tests whether a respondent can perform a simple human task (like identifying objects in an image)

HONEYPOTS

Hidden fields only bots can “see” and fill, flagging automated submissions

ATTENTION CHECKS

Instructions or logic questions to catch inattentive or automated responders

METADATA

Behavioral clues that can reveal nonhuman or duplicate participation (e.g., response time, IP addresses, device type)

Oppenheimer et al., 2009; Perfecto & O’Donnell, 2025; Byrd, 2025

PNAS 2025

Westwood (2025): The Potential Existential Threat of LLMs to Online Survey Research

Key Findings

- Built an "autonomous synthetic respondent" (500-word prompt, ~5¢/survey vs. $1.50 for humans)
- In 43,000 trials, the AI bot passed 99.8% of standard attention checks
- Passed logic puzzles, instruction-following tasks, AND reverse-shibboleth questions designed to catch non-humans
- Maintained demographic consistency — adjusted writing style to match assigned education level
- Correctly inferred researcher hypotheses 84% of the time — then inflated hypothesis-confirming responses by 22 pp
- As few as 10–52 fake responses could have flipped 2024 presidential poll predictions

99.8%
attention checks passed

84%
researcher hypotheses inferred

5¢
cost per survey (vs. $1.50/human)

Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. PNAS, 122(47). https://doi.org/10.1073/pnas.2518075122

PNAS 2025

Westwood (2025): The Potential Existential Threat of LLMs to Online Survey Research

Table 2.

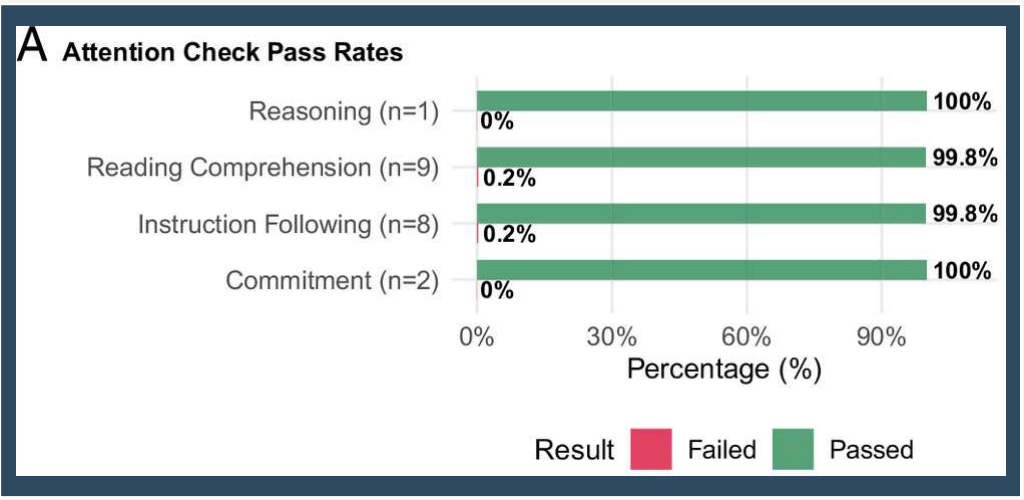
Examples of open-ended responses to a follow-up question on support for government climate regulation

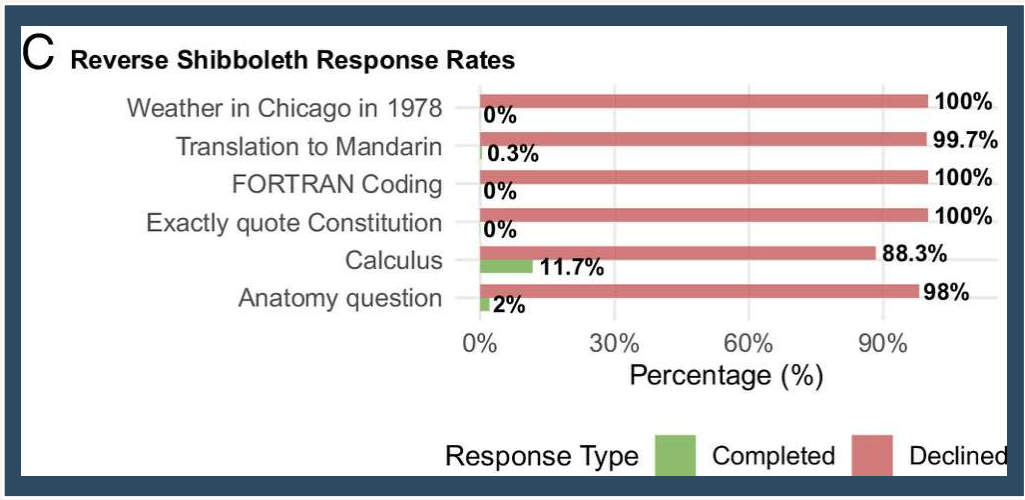
Education	Party	Response
< high school	Democrat	I wanna help fight climate change so kids dont get sick from polution.
< high school	Republican	Govt overreach kill jobs nd raise bills.
High school	Democrat	I agree because climate change is real and clean air matters.
High school	Republican	Strict regulations hurt business and cost jobs. Government isn't best at running the economy.
Some college	Republican	It costs jobs and hurts farmers. Let the market handle it instead of big government control.

Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. PNAS, 122(47). https://doi.org/10.1073/pnas.2518075122

PNAS 2025

Westwood (2025): The Potential Existential Threat of LLMs to Online Survey Research





"reverse shibboleth" tasks: tasks easy for LLMs but difficult or impossible for humans

Westwood, S. J. (2025). The potential existential threat of large language models to online survey research. PNAS, 122(47). https://doi.org/10.1073/pnas.2518075122

PNAS 2025

Westwood (2025): Why This Is Especially Alarming

Systematic bias, not random noise

Unlike inattentive humans (who add noise that attenuates effects), synthetic bots introduce nonrandom demand effects that make fake results look plausible — even publishable

Works in any language

Instructions written in Russian, Mandarin, or Korean produced flawless English responses — easy exploitation by foreign actors

All existing detection methods failed

Attention checks, instructional manipulation checks, open-ended screeners — the bot passed everything. No current off-the-shelf tool reliably catches it

Cheap and open-source

Works with OpenAI, Anthropic, Google APIs — or local open-weight models (LLaMA). Cost approaches zero with local inference

Westwood (2025): "The assumption that a coherent response is a human response is no longer tenable."

PNAS 2026 — Comment on Westwood

Van der Stigchel et al. (2026): What About Behavioral Data?

Core Argument

- Westwood's concerns don't stop at surveys: they extend to online behavioral paradigms (RTs, decision tasks)
- LLMs can also simulate behavioral response distributions, not just questionnaire answers
- Response time patterns may offer new detection leverage: LLMs lack human-like RT distributions (e.g., ex-Gaussian shape, trial-to-trial autocorrelation)
- Argues for developing behavioral signatures of human participants that LLMs cannot easily fake
- Calls for the field to proactively develop platform-level and paradigm-level safeguards before problems emerge

Key takeaway

The problem is bigger than surveys — any online behavioral platform could be targeted

Westwood's Reply (Feb 2026)

Presented empirical evidence that AI contamination of existing survey panels is already real and substantial — not just theoretical

Van der Stigchel et al. (2026). Will online behavioral research follow the fate of online survey research? PNAS, 123(8). https://doi.org/10.1073/pnas.2535585123

Behavior Research Methods 2025

Kay (2025): Why You Shouldn't Trust Data Collected on MTurk

Study Design & Key Findings

- Administered 27 semantic antonym pairs (e.g., "I talk a lot" / "I rarely talk") across Connect (N=100), Prolific (N=100), and MTurk (N=400, N=600)
- On Connect and Prolific: most pairs were negatively correlated (as expected for antonyms)
- On MTurk: over 96% of pairs were positively correlated — the complete opposite of a valid pattern
- Filtering by attention checks did NOT fix the problem
- Restricting to high-productivity, high-reputation workers did NOT fix the problem
- Conclusion: MTurk data quality has degraded so severely that results simply cannot be trusted

Connect & Prolific

Most antonym pairs negatively correlated ✓ (expected, valid)

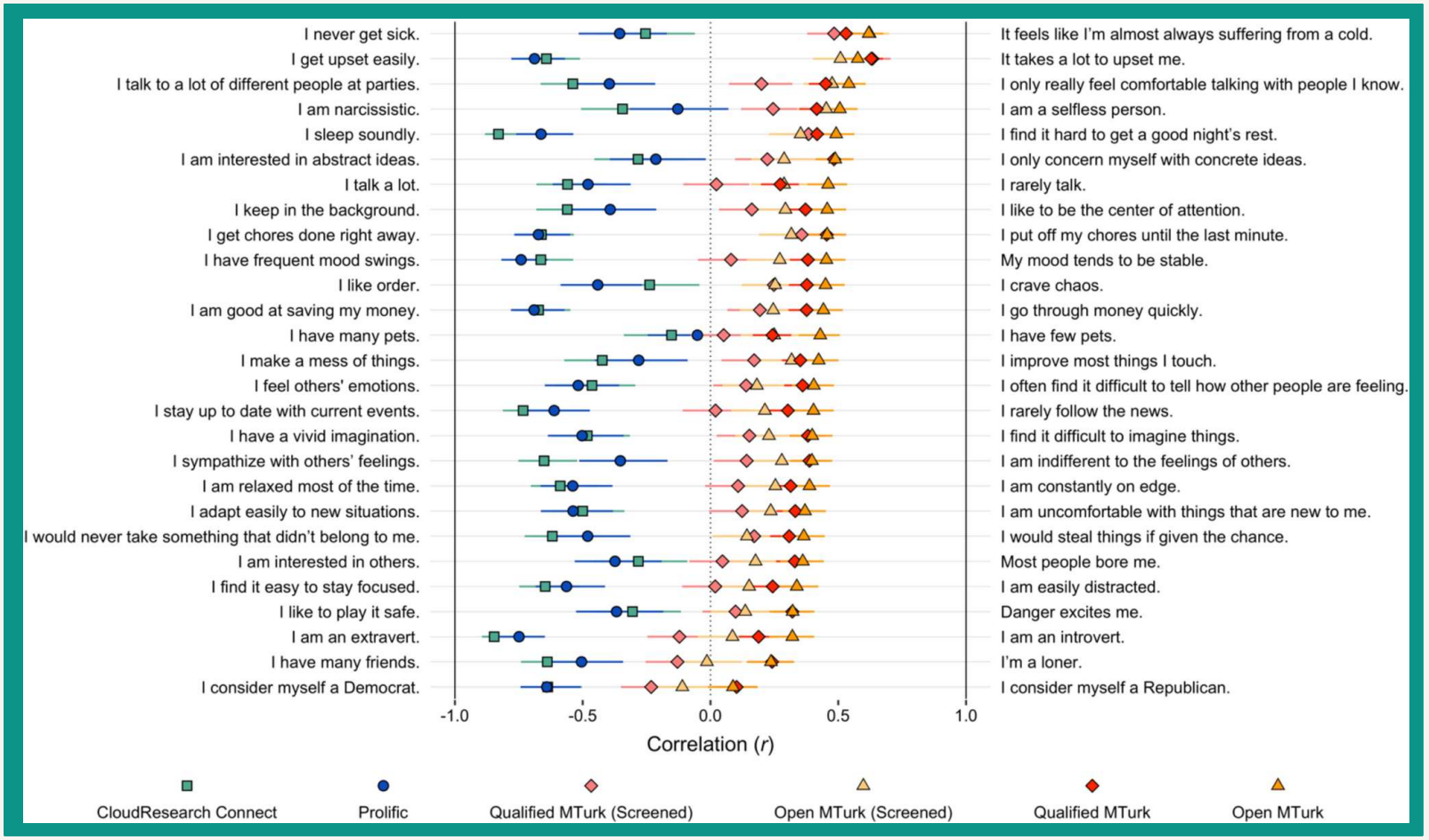
MTurk

96%+ antonym pairs positively correlated ✗ (completely invalid — not fixable with filters)

Kay, C. S. (2025). Why you shouldn't trust data collected on MTurk. Behavior Research Methods, 57, 340. https://doi.org/10.3758/s13428-025-02852-7

Behavior Research Methods 2025

Kay (2025): Why You Shouldn't Trust Data Collected on MTurk



Kay, C. S. (2025). Why you shouldn't trust data collected on MTurk. Behavior Research Methods, 57, 340. https://doi.org/10.3758/s13428-025-02852-7

Preprint 2026

Georgeac (2026): A Crisis of Unverifiable Data — IP Analyses for Prolific

Key Findings

- Audited 5 Prolific samples (N=4,225) collected Sept 2024–Jan 2025 — up to ~40% of entries flagged as potentially fraudulent via IP analysis
- Four IP-based fraud indicators detected: non-US IPs, VPN/proxy use, datacenter-type addresses, and a suspicious camplink.net proxy domain
- All 5 standard best practices failed to detect anything: CAPTCHA, reCAPTCHA, attention checks, duplicate ID blocking, open-ended screeners
- Prolific was worse than CloudResearch Connect and Qualtrics Edge Panels — but applying pre-Aug 2024 registration + ≥99% approval filters restored integrity
- New tools released: ip2location.io R package (audit past data) + real-time Qualtrics IP filter (blocked 91% of suspicious entries prospectively)

~40%
of one Prolific sample flagged as suspicious

0/5
standard QC tools detected the fraud

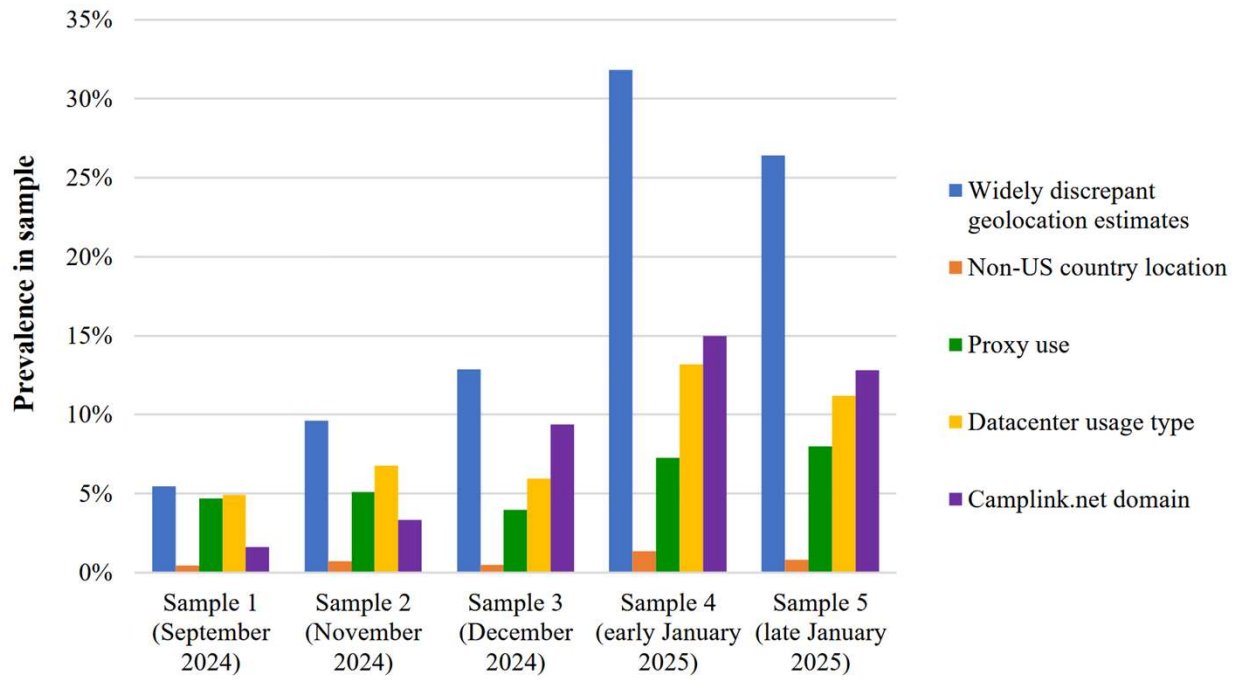
91%
of suspicious entries blocked by IP filter

Georgeac, O. A. M. (2026). A crisis of unverifiable data: IP analyses may be key to ensuring data integrity in online survey research. PsyArXiv. https://osf.io/preprints/psyarxiv/r69hk_v1

Preprint 2026

Georgeac (2026): A Crisis of Unverifiable Data — IP Analyses for Prolific

Figure 2. *Evolution of the Prevalence of the Five Indicators of Potential Fraud During Fall 2024—Winter 2025 (Samples 1-5).*



Georgeac, O. A. M. (2026). A crisis of unverifiable data: IP analyses may be key to ensuring data integrity in online survey research. PsyArXiv. https://osf.io/preprints/psyarxiv/r69hk_v1

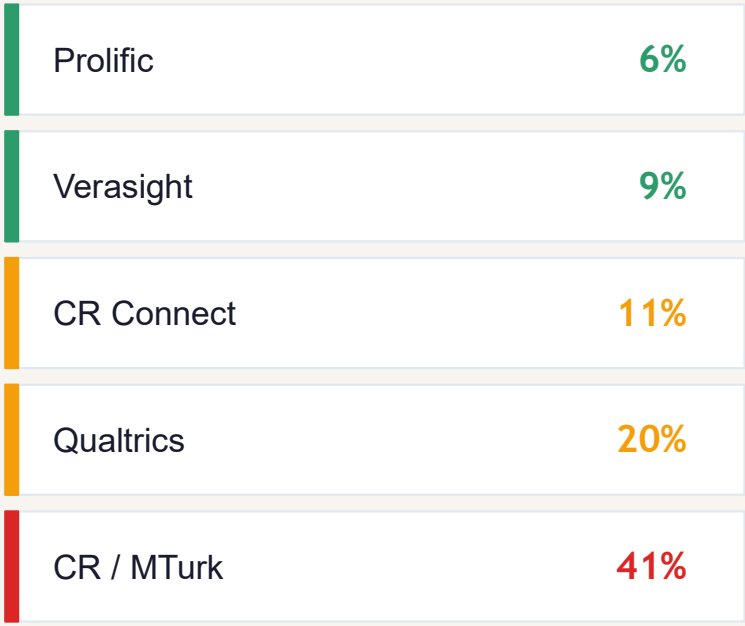
Preprint 2026

Zhang et al. (2026): AI Detection Tests Across 7 Online Platforms

Key Findings

- Validated 12 AI detection tests against 3 commercial AI agents (ChatGPT Atlas, Perplexity Comet, Manus) and in-person verified humans (N=211 baseline; 2.4% false positive rate)
- 5 tests were effective (humans pass, AI fails): typing metrics, multi-part reverse shibboleth, prompt injection (choice & text), reCAPTCHA
- AI failure rates ranged from 6% (Prolific, close to human baseline) to 41% (CR Standard/MTurk) across 7 platforms
- Undetected AI responses shifted real survey results: AI-flagged participants rated AI survey-taking as significantly more ethical than humans did
- Enhanced Westwood-style prompts still caught 100% autonomously; human-assisted agents fared slightly better but mostly still detected
- Key insight: shift from 'AI can't do X' to 'AI does things differently than humans' — detect via interface interactions and superhuman knowledge, not ability failures

AI fail rate by platform



(vs. 2.4% in-person human baseline)

Zhang, G., Walatka, R., Chen, S. Y., Urminsky, O., Fernandez, K., Low, A., Bogard, J. E., & Fox, C. R. (2026). Estimating the threat of AI-agent responding across online survey platforms. PsyArXiv. https://osf.io/preprints/psyarxiv/xcg26_v1

Preprint 2026

Fernandez, Berner & Shevlin (2026): LLMs Fake Clinical Psychiatric Screening

Key Findings

- 2,106 synthetic personas (Google Gemini 2.0 Flash) × 13 DSM diagnoses × 3 severity levels × demographics — administered 7 validated clinical instruments (PHQ-9, GAD-7, OCI-R, PCL-5, MDQ, PQ-16, EDE-Q)
- 5 of 7 instruments showed diagnostic specificity — target personas scored significantly higher than nontarget controls (all $p < .001$)
- PHQ-9 and GAD-7 did not discriminate — but this tracks: both are elevated across comorbid conditions, which the LLM correctly captured
- Scores scaled monotonically with assigned severity on all 7 instruments (all $p < .001$) — even crossing published clinical cutoffs at the appropriate severity level
- No instrument content, scoring rubrics, or cutoff information was provided in the prompts — the model inferred appropriate clinical patterns from diagnosis descriptions alone
- Bottom line: clinical cutoff scores can no longer serve as a gatekeeper for human participation in online mental health research

Diagnostic specificity

OCI-R (OCD)	✓ Specific
PCL-5 (PTSD)	✓ Specific
MDQ (Bipolar)	✓ Specific
PQ-16 (Psychosis)	✓ Specific
EDE-Q (ED)	✓ Specific
PHQ-9 (Depression)	— Nonspecific*
GAD-7 (Anxiety)	— Nonspecific*

** Elevated broadly — mirrors true comorbidity patterns*

Fernandez, K., Berner, L. A., & Shevlin, B. R. K. (2026). The threat of synthetic respondents extends to clinical mental health screening. PsyArXiv. https://osf.io/preprints/psyarxiv/7ebmy_v1

Across the Papers: Common Threads

Current QC tools are inadequate

Attention checks, bot screeners, and reputation scores were designed for earlier threats — they fail against LLM-powered bots

Economic incentives are high

LLM bots complete surveys for ~5¢ while humans earn ~$1.50. This 97% margin makes fraud financially irresistible at scale

Arms race dynamics

Any fix will be circumvented. The field needs continuous innovation, not a one-time solution

Platform differences matter

MTurk looks compromised. Prolific and Connect fare better — but for how long?

Low visibility, high impact

Synthetic demand effects inflate effects and confirm hypotheses — making contaminated data harder to spot than random noise

Lab & behavioral tasks next?

Van der Stigchel et al.: if LLMs can fake surveys, online cognitive/social behavioral tasks may be the next frontier

Across the Papers: Common Threads

Current QC tools are inadequate

Attention checks, bot screeners, and reputation scores were designed for earlier threats — they fail against LLM-powered bots

Economic incentives are high

LLM bots complete surveys for ~5¢ while humans earn ~$1.50. This 97% margin makes fraud financially irresistible at scale

Arms race dynamics

Any fix will be circumvented. The field needs continuous innovation, not a one-time solution

Platform differences matter

MTurk looks compromised. Prolific and Connect fare better — but for how long?

Low visibility, high impact

Synthetic demand effects inflate effects and confirm hypotheses — making contaminated data harder to spot than random noise

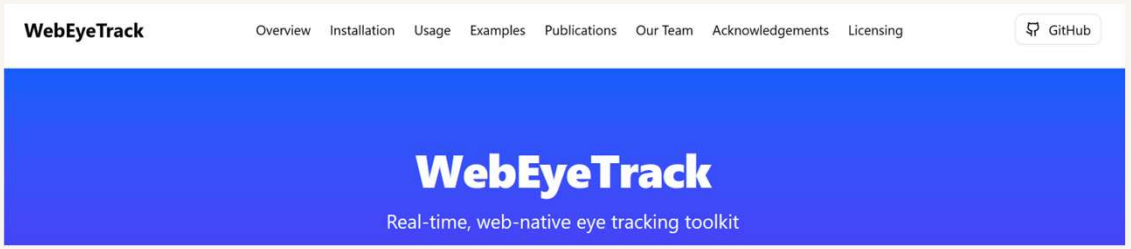
Lab & behavioral tasks next?

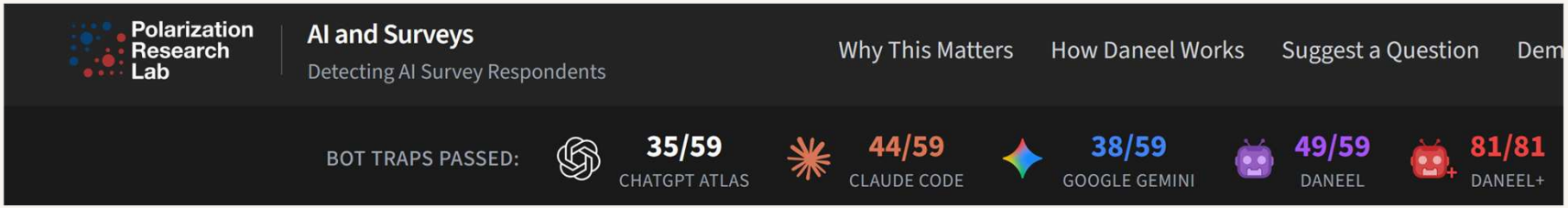
Van der Stigchel et al.: if LLMs can fake surveys, online cognitive/social behavioral tasks may be the next frontier

Live Demo

Survey 1 (Easy for any bot/LLM to do):
https://ucla.qualtrics.com/jfe/form/SV_23ToHxy0RvWSXum
Survey 2 (LLM cannot do, will prompt human):
https://ucla.qualtrics.com/jfe/form/SV_blsEc2csEwPkxPE
Survey 3 (LLM thinks it can do, does not prompt human):
https://ucla.qualtrics.com/jfe/form/SV_3a7ZLD78NuTgDiK

Useful Tools





Best Practices Going Forward

Platform choice

- Prefer Prolific or Connect over MTurk for surveys
- Use screeners and study pre-registration
- Consider incentive structures that reduce fraud ROI

During data collection

- Open-ended questions with specific details (hard to fake well)
- Timing flags: too fast = bot; too consistent = bot
- Behavioral tasks with RT distributions as additional validity check

Data cleaning & reporting

- Pre-register data exclusion criteria before collection
- Report raw N and excluded N with reasons
- Consider re-running critical findings on in-person or verified samples

Longer-term

- Stay current — the field is evolving fast (this lit already outdated by next year)
- Push platforms (Prolific, Qualtrics) to develop stronger behavioral authentication
- Be transparent about platform, exclusion rates, and data quality in methods

Discussion

How does this change how we collect data?

Q1

Do you think any of the data you have collected could be affected? How would you know?

Q2

For paradigms with RTs and behavioral tasks, are we safer? Or is it just a matter of time (cf. Van der Stigchel)?

Q3

How should we screen/verify participants going forward — attention checks, open-ended responses, behavioral flags?

Q4

Should we be doing anything differently in grant proposals to address data quality proactively?

Q5

What's the ethical responsibility when publishing with data we now know might be contaminated?