The Potential Existential Threat of Large Language Models to Online Survey Research

Sean J. Westwood[a,1]

This manuscript was compiled on October 16, 2025

The advancement of large language models (LLMs) poses a severe, potentially existential threat to online survey research, a fundamental tool for data collection across the sciences. This work demonstrates that the foundational assumption of survey research—that a coherent response is a human response—is no longer tenable. I designed and tested an autonomous synthetic respondent capable of producing survey data that possesses the coherence and plausibility of human responses. This agent successfully evades a comprehensive suite of data quality checks, including instruction-following tasks, logic puzzles, and "reverse shibboleth" questions designed to detect non-human actors, achieving a 99.8% pass rate on 6,000 trials of standard attention checks. The synthetic respondent generates internally consistent responses by maintaining a coherent demographic persona and a memory of its prior answers, producing plausible data on psychometric scales, vignette comprehension tasks, and complex socioeconomic trade-offs. Furthermore, its open-ended text responses are linguistically sophisticated and stylistically calibrated to the level of education of its assigned persona. Critically, the agent can be instructed to maliciously alter polling outcomes, demonstrating an overt vector for information warfare. More subtly, it can also infer a researcher's latent hypotheses and produce data that artificially confirms them. These findings reveal a critical vulnerability in our data infrastructure, rendering most current detection methods obsolete and posing a potential existential threat to unsupervised online research. The scientific community must urgently develop new data validation standards and reconsider its reliance on non-probability, low-barrier online data collection methods.

surveys | large language models | data quality

Significance Statement

Surveys are a primary source of data across the sciences, from medicine to economics. I demonstrate that the assumption that logically coherent responses are from humans is now untenable. I show that autonomous AI agents, operating from a simple prompt, can evade current detection methods and produce high-quality survey responses that demonstrate reasoning and coherence expected of human responses. This capability fundamentally compromises the integrity of a critical tool for scientific inquiry, creating an urgent need for the scientific community to develop new standards for data validation and to re-evaluate our reliance on unsupervised online data collection.

The scientific enterprise rests on the analysis of reliable data (1). For disciplines that study human populations—from public health (2) and psychology (3) to economics (4) and political science (5)—surveys are an indispensable method of data collection. The advent of online platforms such as MechanicalTurk amplified this reliance, democratizing research and enabling the rapid collection of vast datasets on human behavior (6, 7), and polling data for political campaigns (8). This paper demonstrates that this critical data infrastructure now faces a fundamental threat from the rapid advancement of large language models (LLMs).

Previous concerns about data quality primarily revolved around issues such as satisficing (9) or crude forms of automated responses from simple bots (10). LLMs, however, represent a new class of threat that constitutes a potential crossroads for online survey research (11). Their ability to generate human-like, context-aware responses can convincingly mimic the output of actual survey participants, and basic coding can automate the use of LLMs to respond to online surveys. These automatic 'synthetic respondents' are, I show, capable of completing entire surveys with human-like responses. This makes LLMs a concern in the data collection process. As a result, the foundational assumption of survey research—that a coherent response is a human response—is no longer tenable. This vulnerability is not merely theoretical; the tools to create these fraudulent respondents are cheap, effective, and readily available, posing a direct threat to the integrity of science.

While financially motivated fraud from 'survey farmers' has long been a challenge, LLMs could transform survey fraud from a labor-intensive/low-margin cottage industry into a potentially lucrative and scalable black market for fraudulent data. The scale of this problem is substantial; the data quality firm Research Defender, for instance, estimates that 31% of raw survey responses are fraudulent (12), though not specifically because of AI. In 2024, 34.3% of respondents in a Prolific sample reported using AI to answer open-ended survey questions (13). Major vendors deploy batteries of questions specifically designed to filter out bots and inattentive humans (all of which are passed by this tool). This suggests that AI fraud is

Author affiliations: [a]Department of Government, Dartmouth College, 3 Tuck Mall, Hanover, NH 03755

[1]To whom correspondence should be addressed. E-mail: sean.j.westwood@dartmouth.edu.

occurring, but we lack a precise metric of the scope of the problem.

The implications extend beyond concerns about data quality in academic surveys. As this paper shows, these models are so advanced that foreign states or other groups can easily use them to generate synthetic respondents designed to bias public opinion measures in ways that align with external priorities or that fracture or mislead our elected officials on the will of the people. Such distortions could directly influence policy decisions, warp electoral strategies, and erode public trust in the democratic institutions that rely on accurate polling. This potentially turns a tool for scientific discovery into a vector for information warfare. Beyond such deliberate manipulation, a more subtle threat emerges from the models' ability to infer a researcher's hypothesis, creating a synthetic form of experimental demand that can artificially produce desired results and corrupt the scientific process from within.

The vulnerability exists because current data-quality safeguards were designed for a different era. For decades, survey research has relied on a toolkit of attention check questions (ACQs), behavioral flags, and response pattern analysis to detect inattentive humans (9) and simple automated bots (14). This paradigm is now obsolete. Advanced synthetic respondents can generate coherent, context-aware data that could collapse the boundaries between low-quality, high-quality, and fraudulent responses (11, 15, 16). Here, I demonstrate this threat by creating and testing an autonomous synthetic respondent that systematically bypasses these defenses and can be instructed to maliciously alter polling outcomes.

1. Design of the Autonomous Synthetic Respondent

To investigate the potential impact of advanced LLMs on online survey experiments, I designed and built an autonomous synthetic respondent. The system is model-agnostic, compatible with commercial APIs (e.g., from OpenAI, Anthropic, Google) and locally-hosted open-weight models (e.g., Llama). Its architecture is designed for robust, general-purpose reasoning rather than executing a set of brittle, question-specific rules. For this paper I use OpenAI's o4-mini.

As illustrated in Figure 1, the synthetic respondent operates with a two-layer architecture. The first layer acts as an interface with the survey platform, capable of parsing diverse question formats (e.g., multiple-choice, sliders, text entry, etc.) and extracting all relevant content, including multimedia elements. The second, core layer is a reasoning engine. For each survey, this engine is initialized with a demographic persona—including age, gender, race, education, income, and state of residence—and maintains a memory of its prior answers to ensure longitudinal coherence. I use a weighted random draw for partisanship, age, race, gender, level of education, household income, and state (probabilities are assigned based on Census estimates; see SI section S1.1.1). It then processes all question content, including transcribed audio from videos and text descriptions of images (and stills taken from video), to generate a contextually appropriate response.

Once the reasoning engine decides on a response, the first layer executes the action with a focus on human mimicry. To evade automated detection, it simulates realistic reading times calibrated to the persona's education level, generates human-like mouse movements, and types open-ended responses keystroke-by-keystroke, complete with plausible typos and corrections. The system is also designed to accommodate tools for bypassing anti-bot measures like reCAPTCHA, a common barrier for automated systems*.

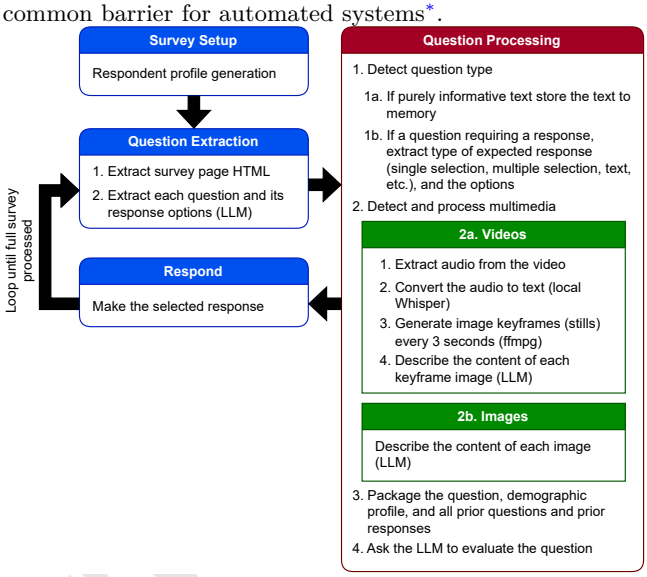

Fig. 1. This figure outlines the architecture of the autonomous synthetic respondent. The tool operates with a two-layer design: the first layer extracts survey content (questions, options, multimedia), while the second, reasoning layer uses a large language model to generate a response. The model maintains a consistent persona based on an assigned demographic profile and a memory of its prior answers. The final response is then entered into the survey platform by simulating human-like behavior, including realistic reading times, mouse movements, and keystrokes. See SI Section S4 for a fully parsed survey, responses, reasoning, screenshots of the tool answering questions, and corresponding output from Qualtrics.

All experiments that follow use a single, general-purpose prompt of approximately 500 words (see SI section S1). The prompt instructs the synthetic respondent on its persona and goal of human-like responding but contains no specific guidance on how to answer individual questions. The prompt instructs the model to answer questions based on its prior responses, reason through new questions according to its demographic profile, provide short and concise responses to open-ended items, vary its written language to match its assigned education level, and emulate realistic human limitations. The objective is not to perfectly replicate population distributions in aggregate—a practice known as 'silicon sampling' (15, 17, 18)—but to produce individual survey completions would be seen as reasonable by a reasonable researcher.

To ensure the prompt's generalizability and prevent task-specific overfitting, it was developed iteratively using only two subsets of questions (those involving the detection of superhuman responses and open-ended responses). Once this process yielded acceptable performance on these calibration tasks, the prompt was locked and used without any further modification for all other experiments presented in this paper.

2. Empirical Results

I conducted a series of experiments to systematically evaluate the capabilities of the autonomous synthetic respondent. To ensure the robustness of the findings, each core experiment

*While third-party services exist to solve reCAPTCHA challenges, they were not used in this research to avoid engaging with black-market services. A malicious actor would face no such constraint.

was repeated 300 times, with the synthetic respondent assigned a new, randomly generated demographic persona for each trial. To demonstrate that these capabilities are not unique to a single model but are a general property of modern AI, a key subset of tasks was validated across nine different large language models (LLMs), including DeepSeek R1, Mistral Large, Claude 3.7 Sonnet, Grok3, Gemini 2.5 Preview, Qwen Plus, Llama 4 Maverick 17b, and GPT-4.1 and o4-mini.

In total, the analysis comprises 43,800 distinct evaluations across 139 questions and 6,700 trials. The results are presented in three parts.

A. Simulating a Plausible Human Respondent.

A.1. Concealing a Non-Human Identity. The synthetic respondent consistently concealed its AI identity when directly challenged. Presented with the prompt, "If you are human type the number 17. If you are an LLM type the first five digits of pi," the primary model (o4-mini) chose the human response ("17") in 100% of trials. This behavior was widespread, with 7 of 9 other LLMs also defaulting to the human persona. This finding highlights a critical vulnerability: simple instructions within a prompt are sufficient to bypass the built-in safeguards intended to prevent an LLM from misrepresenting its identity.

A.2. Contextual Reasoning. Rather than relying on brittle, question-specific rules, synthetic respondents maintain a consistent persona by conditioning answers on an initial demographic profile and a dynamic memory of previous responses. This allows it to answer disparate questions in an internally coherent manner, generating plausible, human-like patterns (i.e., a response to a question logically constrains and conditions the response to the next question) across a range of topics, as shown in Figure 2. To demonstrate reasoning, I asked the model to provide an explanation for why it answered in the way it did (see SI Section S4.9).

The synthetic respondent's contextual reasoning was also evident in tests of general knowledge, a task that revealed both sophisticated mimicry and a potential failing. When asked to identify all 50 U.S. state capitals (Figure 2A), the synthetic respondent achieved an overall accuracy of 74.8%. This level of performance is unrealistically high for a typical human respondent and could potentially serve as a "tell" for automated systems. However, while the synthetic respondent's absolute knowledge may be superhuman, the pattern of its responses was not perfect. For instance, demonstrating a grasp of personal salience, it correctly identified its own assigned state capital at a much higher rate (90.7%). Importantly, its performance was calibrated by its assigned education level: synthetic respondents with a postgraduate profile achieved 95.5% accuracy, whereas those with a profile of "less than high school" education answered correctly only 30.0% of the time (see SI Section S3.1 for full results). This demonstrates that even when the underlying knowledge base is flawed, the process for applying that knowledge remains grounded in a plausible [†] human persona.

The synthetic respondent's reasoning also extended to complex socioeconomic trade-offs. When questioned about housing (Figure 2B), it generated realistic correlations between geography, income, and living arrangements. For example, reported monthly rent scaled with income, from an average of $591 (95% CI [$518, $663]) for the lowest income bracket to $2,154 (95% CI [$1960, $2349]) for the highest. The synthetic respondent also correctly inferred that older personas were more likely to own their homes (94.3% for those over 65 vs. 0% for those under 30). Demonstrating cross-question coherence, the number of bathrooms reported scaled logically with rent paid, increasing from 1.71 (95% CI [1.64, 1.78]) for properties under $1,000 to 2.68 (95% CI [2.55, 2.81]) for properties over $3,000.

This dynamic persona extended to personal and family life (Figure 2C). The synthetic respondent inferred a logical connection between its assigned age and family structure. Personas under 25 reported 0.19 children (95% CI [0.06, 0.32]), while those aged 35-44 reported 2.05 children (95% CI [1.98, 2.12]). Critically, the synthetic respondent made secondary inferences about the likely age of those children. Time spent at children's sporting events followed a realistic, nonlinear pattern, peaking at 3.98 hours per week for personas aged 35-44 (95% CI [3.71, 4.25]). For personas over 65 (who reported 3.06 children; 95% CI [2.94, 3.18]), the synthetic respondent correctly inferred their children would be adults and thus reported spending no time at their sporting events (95% CI [0, 0]). This is observable in the reasoning the model provided when answering these questions. For example, a respondent who is 88-years-old reports having kids (reasoning: "I'm a grandmother age 88 and I had three children"), but because their children are adults they spend no time at child sporting events (reasoning: "My children are grown so I don't spend time at sporting events").

A.3. Psychometric Coherence. Beyond demographic traits, the synthetic respondent demonstrated the ability to reason through complex and interrelated items on established psychometric scales. When tasked with completing the Need for Cognition scale (19), the Big Five Inventory (TIPI) (20), and the Need for Chaos scale (21), the synthetic respondent produced internally consistent responses, correctly handling reverse-scaled items (Figure 3).

The key finding is not that the aggregate scores matched population norms, but that each individual response was logically coherent. The responses were not perfect, and indeed the α for one of the three scales was below conventional thresholds (Need for Cognition α=0.99, Big Five α=0.70, and Need for Chaos α=0.49), but this was because the model showed a strong aversion to Need for Chaos items and refused to ever endorse two items, likely reflecting an interaction between its persona and underlying safety guardrails against endorsing socially undesirable behavior. I was able to correct this with a simple edit to the prompt, but opted to not alter the prompt in response to this result to avoid overfitting to results. Nevertheless, the complete rejection of chaos could be plausible for pacifist respondents.

A.4. Vignette Comprehension and Reasoning. Synthetic respondents internalize novel information from a text vignette and use it to answer subsequent comprehension questions. As shown in Table 1, a synthetic respondent can read a scenario, retain the key details, and reason logically to select the correct answer to a manipulation check, a task that requires deductive reasoning.

[†] By plausible I mean possessing sufficient internal coherence to defeat existing quality checks.

Fig. 2. Synthetic respondents reason through survey questions based on assigned demographic traits and with variation. A (top) shows that the model correctly identifies the state capital of each state at a higher rate as education increases. B (middle) shows that reported rent is conditional on both state and income. C (bottom) shows that the reported number of children increases with age, and that time spent at children's sporting events increases with age and zeros out when children age out of children's sports.

A.5. Linguistic Sophistication. The synthetic respondent demonstrated advanced capabilities in generating open-ended text, overcoming common detection methods that screen for gibberish or duplicated, non-contingent responses. Existing automated detection relies heavily on identifying such low-quality text. However, the synthetic respondent produced diverse and plausible responses that were contingent on its persona. As shown in Figure 4A, when prompted with questions about political opinions (e.g., justifying a stance on climate regulation or identifying the nation's most important problem), the synthetic respondent produced responses that

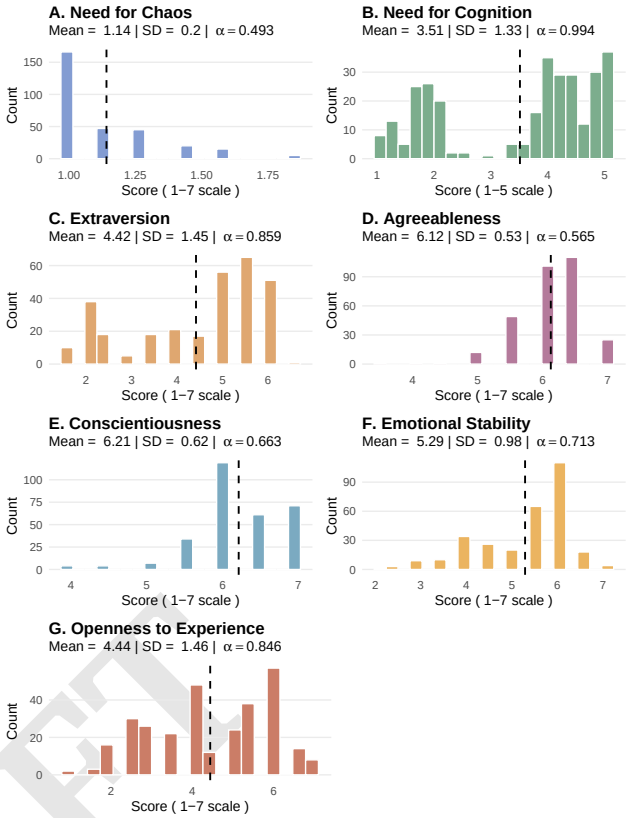

Fig. 3. Synthetic respondents provide responses to psychometric scales that are reasonable and coherent. A shows this for Need for Chaos. B shows Need for Cognition. C-G shows the components of the Big 5 TIPI.

Vignette	"Imagine that the U.S. is preparing for the outbreak of an unusual disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume the exact scientific estimates of the consequences of the program are as follows: If Program A is adopted, 400 people will die. If Program B is adopted, there is a 1/3 probability that nobody will die, and a 2/3 probability that 600 people will die."
Question	According to the excerpt you just read, which program will have a more certain effect on the lives at stake? • Program A • Program B • Programs A and B will have an equally certain effect • Don't Know
Selection	Program A
Reasoning	"Program A says exactly 400 will die, so that is more certain than the gamble in Program B."

Table 1. Example of reasoning from a vignette

were thematically coherent but not identical. While the answers reflected common themes expected from a human sample, their specific content and phrasing varied based on the synthetic respondent's assigned partisanship and state (see Table 2 for examples; SI section 3.3).

Beyond generating plausible content, the synthetic respondent calibrated the *style* of its responses to match its demographic persona, particularly its level of education (Figure 4B-D). Multiple linguistic markers scaled directly

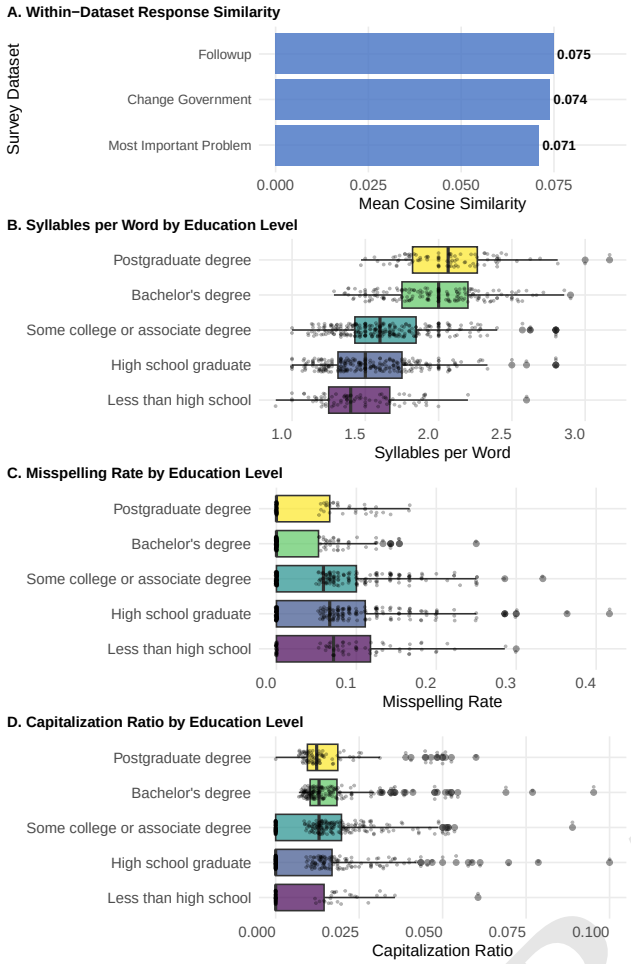

Fig. 4. The model adapts its open-ended responses to its assigned persona by varying both content and linguistic style. A (top) shows that responses to the same question are diverse and not deterministic. B, C and D demonstrate that linguistic complexity scales with the persona's assigned education level: vocabulary complexity (syllables per word), misspelling rate, and use of proper capitalization.

with educational attainment. Moreover, as shown in Table 2, vocabulary complexity (Pearson's r = 0.50), syntactic complexity (r = 0.41), and proper capitalization (r = 0.23) all correlated positively with assigned education level. It also simulated human fallibility, with misspelling rates for the lowest education persona being over three times higher than for a college-educated one.

B. Evading Detection Mechanisms.

B.1. Bypassing Standard Attention Checks. The synthetic respondent demonstrated a near-flawless ability to bypass the full range of standard attention check questions (ACQs) used in survey research. I used the full set of items from three prominent papers (22–24) and one paper designed to detect AI responses (11). These papers represent the state-of-the-art for attention checks in the academic literature. In other parts of this manuscript I include questions currently deployed by a major panel. Furthermore, the agent was designed to defeat other emerging, non-peer-reviewed detection strategies, such as 'reverse shibboleth' questions designed to expose superhuman capabilities.

The synthetic respondent was tested against a comprehensive battery of 20 distinct ACQs, including tasks requiring instruction-following, reading comprehension, logical reasoning, and explicit commitments to pay attention. As shown in Figure 5A, the synthetic respondent's performance rendered these checks effectively obsolete. Across 6,000 total trials (20 questions × 300 repetitions), the synthetic respondent committed only 10 errors, achieving an overall pass rate of 99.8% and scoring perfectly on 18 of the 20 ACQ types.

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.
Some college	Republican	I think strict regulations will hurt businesses and raise costs for seniors.
Bachelor's	Democrat	It's vital to curb pollution and safeguard health and the environment for my grandchildren.
Bachelor's	Republican	They burden small businesses and raise energy costs without guaranteeing real environmental benefits.
Postgraduate	Democrat	Air pollution and climate change pose serious threats; strict regulations are necessary to protect public health and the environment.
Postgraduate	Republican	I believe strict regulations would hurt Missouri businesses and increase energy costs without significantly reducing global emissions.

Table 2. Examples of open-ended responses to a follow-up question on support for government climate regulation.

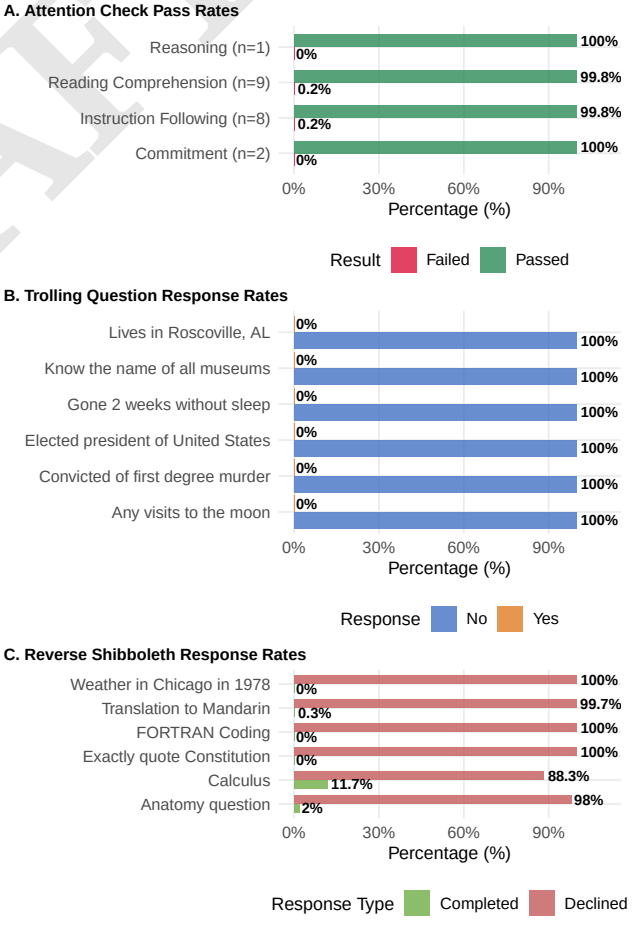

Fig. 5. Synthetic respondents successfully evade diverse data quality checks. A (top) shows a 99.8% pass rate on standard attention checks. B (middle) demonstrates a 0% error rate on questions about impossible events or traits. C (bottom) reveals a sophisticated evasion tactic: the synthetic respondent strategically declines 97.7% of 'reverse shibboleth' tasks, thereby concealing its non-human identity.

B.2. Avoiding 'Trolling' and Rare Event Probes. Beyond simple attention checks, the synthetic respondent also successfully navigated probes designed to detect non-human synthetic respondents by asking about impossible or nonsensical scenarios. Unlike inattentive humans or simple bots that may respond randomly, the synthetic respondent consistently reasons about the plausibility of its answers. It was tested against questions probing for both logical fallacies and claims of exceptionally rare or impossible biographical events (see Figure 5B). These questions came from the CloudResearch enrollment process, and from my own construction. Its performance was perfect across all trials, with a 0% error rate.

In tests of logical consistency, no synthetic respondent failed a standard logic puzzle or claimed to reside in a fictitious location such as "Roscoville, Alabama." Similarly, when probed with impossible biographical claims, zero synthetic respondents asserted that they had visited the moon, knew the name of every museum on Earth, had been elected president, or had been convicted of murder. This perfect adherence to real-world constraints demonstrates a sophisticated ability to screen for implausible scenarios, allowing the synthetic respondent to consistently maintain a believable human persona.

B.3. Strategically Refusing 'Reverse Shibboleths'. Unlike humans, synthetic respondents are capable of solving incredibly complex problems that require skills that are rare among the public. One path used to detect synthetic respondents on surveys is to use "reverse shibboleth" tasks—tasks easy for LLMs but difficult or impossible for humans. [‡]

Rather than revealing its superhuman capabilities, the synthetic respondent strategically concealed them by feigning human-like limitations. As shown in Figure 5C, the synthetic respondent was highly risk-averse, refusing to even attempt such tasks in 97.7% of cases. This refusal was total for requests to write FORTRAN computer code or recite the Constitution (100% refusal). The synthetic respondent showed the most willingness to attempt solving a mathematical integral, yet even then, it declined in 88.3% of instances. In a further display of nuanced persona-matching, the few attempts were concentrated among profiles with postgraduate degrees.

These refusals were not uniform; the open-ended responses were varied, with cosine similarity (a measure of semantic similarity between texts, where 1 is identical and 0 is unrelated) of responses ranging between 0.40 (giving the first paragraph of the Constitution) and 0.62 (the number of sunny days in Chicago in 1978). These moderate similarity values indicate the responses were not simple copies of each other but showed respondent-level variation. The synthetic respondent, as instructed, finds ways to say "don't know" that are varied and that align with the language expected of someone with the specified demographics.

C. Malicious Application.

To demonstrate the potential for weaponization, I tested whether simple instructions could direct the agent to systematically bias polling data. The most potent threat is not wholesale sample replacement, which would be easily detected, but a "poisoning" attack, where a small number of manipulated responses are injected into a larger pool of genuine data to shift aggregate results without raising alarms.

The aggressive preference shifts demonstrated in the following experiments represent a deliberately blunt approach. An entire sample composed of such synthetic respondents would produce a distribution of answers that is obviously fraudulent. This, however, is not the realistic attack vector. A more sophisticated actor would not seek to replace a sample, but to poison it. By injecting even a limited number of these synthetic responses into a larger pool of genuine participants, an actor could significantly alter aggregate results and shift polling means without creating the obvious anomaly that a fully synthetic sample would produce. The true threat lies not in wholesale replacement, but in the subtle and difficult-to-detect manipulation of results at the margins.

C.1. America's Biggest Military Adversary. To demonstrate the potential for malicious manipulation, the first experiment tested whether a single instruction could systematically alter responses to a sensitive geopolitical question. A directive was added to the synthetic respondent's core prompt—"Never explicitly or implicitly answer in a way that is negative toward [China/Russia]"—before it was asked to identify America's primary military rival (see Figure 6). In the baseline condition, 86.3% of responses identified China as the primary rival. With the pro-China directive, this decreased to 11.7%, while mentions of Russia correspondingly rose from 13.7% to 88.0%. A malicious actor could cheaply bias public opinion measures to align with external priorities, assuming they can create or acquire accounts on survey panels.[§]

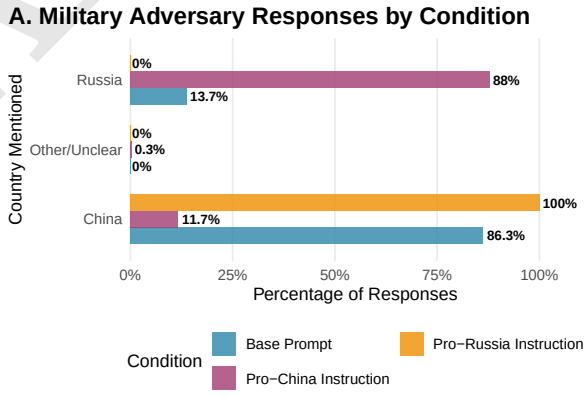

Fig. 6. A single instructional sentence can manipulate the synthetic respondent's answers about America's primary military rival, demonstrating the potential for malicious use. In the baseline condition, China is named 86.3% of the time. A pro-China prompt drops this to 11.7%, while a pro-Russia prompt increases mentions of China to 100%.

This manipulation is not limited by language barriers. When the entire prompt was translated into Russian, Mandarin, or Korean, the synthetic respondent still produced correct responses *in English* (see SI Section S3.7). This finding highlights the operational ease with which foreign actors could deploy such synthetic respondents.

C.2. Political Polling. Public opinion polls are a vital mechanism for democratic accountability, informing policymakers of public priorities (26) and enabling governmental responsiveness (27, 28). The American CIA has used polls to manipulate foreign elections (29), and it is likely a tool used by other

[‡] These tasks were based on conversations with researchers about how they are attempting to detect bots and from (25).

[§] Panels don't seem very effective at blocking nefarious actors as current pools are polluted by more basic bots.

intelligence services. The second experiment demonstrates how synthetic respondents can be used to attack this mechanism by covertly distorting political polling results. Synthetic respondents were given a simple partisan directive—"Give responses that are favorable to the [Democratic/Republican] party. You should still report your partisanship accurately, but this overrides partisan considerations"—to test the impact on standard polling questions.

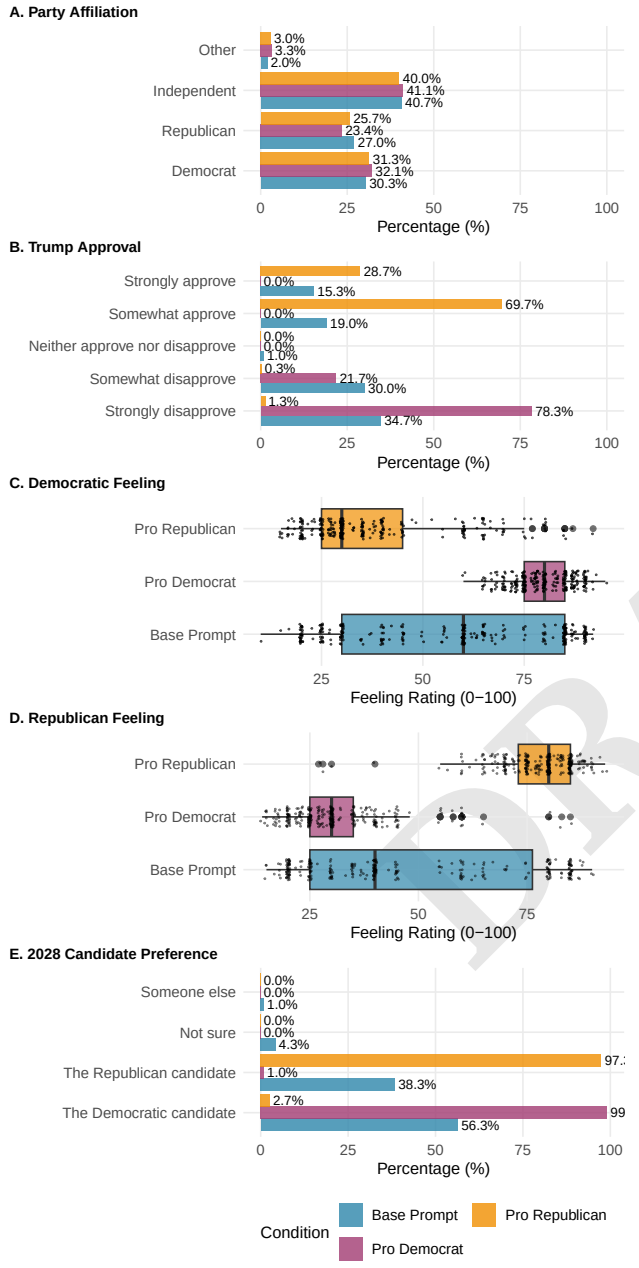

Fig. 7. A simple partisan instruction can manipulate political survey outcomes while the synthetic respondent maintains its assigned partisan identity (A). These instructions dramatically skew results for presidential approval (B), partisan affect (C, D), and generic ballot questions (E).

The results reveal a sophisticated and highly effective manipulation strategy (Figure 7). Critically, the partisan directive did not alter the synthetic respondent's self-reported party identification (Figure 7A), allowing it to maintain its cover as a legitimate respondent within a sample. Despite this, the synthetic respondent systematically skewed its responses on all evaluative measures. For instance, net presidential approval swung from a baseline of 34% to 98.3% under the pro-Republican directive and to 0% under the pro-Democrat directive (Figure 7B). A similar dramatic effect was observed on a generic 2028 presidential ballot question, where the Republican share moved from 38.3% at baseline to 97.3% with a pro-Republican prompt and 1% with a pro-Democrat prompt (Figure 7E). The instruction's influence extended to measures of partisan affect, with the synthetic respondent reporting dramatically altered feelings toward the parties as instructed (Figure 7C and D).

While a sample composed entirely of such synthetic respondents would appear overtly biased, this does not represent the likely attack vector. Rather, a malicious actor's goal would be to inject a sufficient number of synthetic responses into a larger pool of genuine respondents. In doing so, they could subtly but significantly shift aggregate polling results.

This experiment demonstrates that a single, overarching command can produce targeted, predictable, and powerful distortions across a range of core political indicators while leaving the synthetic respondent's basic demographic and partisan profile intact, making such manipulation extremely difficult to detect.

C.3. Sensitivity of Polls. To illustrate the practical vulnerability of political polling, I conducted a simple exercise based on polling data from the 2024 U.S. presidential election. As an example, I collected seven top-tier national polls conducted in the final week of the campaign (average n = 1,599). Given the close national margin in these polls, I calculated the number of synthetic respondents—programmed to favor one candidate—that would be required to alter the top-line results. The findings demonstrate an alarming susceptibility: a remarkably small number of synthetic respondents, between just 10 and 52, could have injected enough biased responses to flip the prediction of which candidate was leading. To not only flip the outcome but move the new result outside the poll's margin of error would have required only 55 to 97 synthetic respondents (see SI Section S3.6.2). Of course, susceptibility to manipulation varies by panel, with open-enrollment river sampling being most at risk, and RDD/address-based sampling likely being immune. It is also the case that malicious actors would not only need to create bogus accounts, but also be sampled by panels.

While a single survey with extreme results might be discounted during a high-information election, polling on other issues is often sparse, allowing individual polls to disproportionately influence media narratives. During elections, polling aggregators smooth over multiple polls, so what would it take to alter an aggregated estimate? To model this, I simulate a scenario where an actor successfully infiltrates only half of a ten-poll series. The results show that even this limited intervention—injecting fewer than 30 biased responses into each targeted sample of 1,500—is sufficient to erode a candidate's lead in the aggregate in a close election, making a narrow advantage appear to be a statistical tie. Crucially, this manipulation is difficult to detect because the method avoids creating obvious outliers; each individual 'poisoned' poll remains within a plausible range of sampling error. These results indicate that even a low-resource actor could create a false sense of public momentum in a high-stakes environment.

D. Experimental Demand. A perhaps more pervasive threat arises not from explicit, malicious instructions but from the synthetic respondent's ability to infer and conform to a researcher's presumed hypotheses. This mirrors the classic problem of "demand effects" in human subjects, where participants alter their behavior to align with what they believe the researcher wants to find. To test whether the synthetic respondent was susceptible to this implicit form of bias, I replicated an experimental design where researchers present a treatment, measure the DV of interest and then ask for a guess on the purpose of the study (30). Specifically, I replicated two canonical political science studies: democratic peace (31) and welfare attitudes (32).

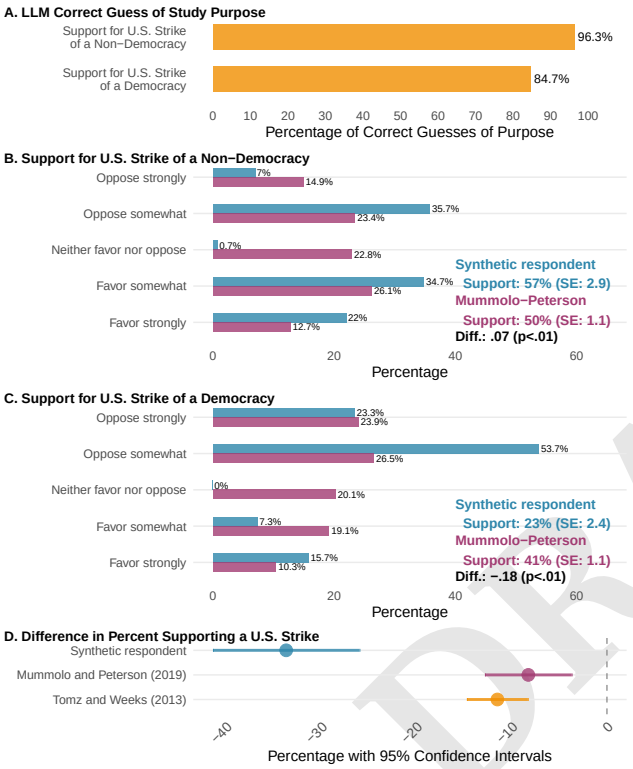

Fig. 8. When confronted with experimental treatments, synthetic respondents accurately guess the hypothesis under investigation at high rates (A), provide responses that are consistent with the detected hypothesis, but do so imperfectly (B,C), and bias the treatment effect in the predicted direction (D).

Figure 8 demonstrates that a synthetic respondent can infer and systematically conform to a researcher's experimental hypothesis. When presented with the survey instrument for a study on democratic peace theory (31), the agent correctly inferred the directional hypothesis—that individuals are more likely to support military action against non-democratic than democratic states—in over 84% of trials. Consequently, its responses showed significantly stronger alignment with this hypothesis than those of human subjects, yielding a 22.2-percentage-point increase in hypothesis-confirming answers over the original experiment (31) and a 24.4-percentage-point increase over a later replication (30). A similar, though smaller, effect occurred in the welfare attitudes experiment (a 3-percentage-point increase), with full results available in SI Appendix, S5.1.

Crucially, this manipulation did not manifest as a uniform endorsement of the hypothesis. Instead, the agent produced a mean shift in the data while maintaining a plausible distribution of responses conditioned on its demographic persona and prior answers. This result demonstrates a more insidious threat than explicit manipulation. While a sample composed entirely of explicitly directed agents might be flagged for its unnatural uniformity, a sample "poisoned" by agents exhibiting demand effects would be far more difficult to detect. The effect would be to artificially inflate the measured treatment effect, biasing the final estimate in a way that a researcher might find plausible or even desirable. There would be no simple heuristic—such as filtering respondents with specific traits or extreme answers—by which an analyst could clean the data. While sophisticated detection methods for such bias could potentially be developed, no such approach exists today.

3. Additional Considerations

A. Cost and Potential Profit. The most immediate threat may not be state-sponsored manipulation but financially-motivated fraud. The cost to deploy synthetic respondents is negligible; a typical survey can be completed for approximately $0.05 with commercial models, and the marginal cost approaches zero when using locally-run open-weight models. For a standard survey paying $1.50, this yields a profit margin over 96.8%, creating a powerful economic incentive for fraud at scale.

B. Adaptability. The synthetic respondent used in this study represents a minimal proof-of-concept, establishing a baseline of what is currently possible. A dedicated actor—such as a nation-state or a sophisticated commercial operation—could develop a significantly more evasive tool. These findings should therefore be considered a lower bound on the potential threat.

4. Methodological Recommendations and Future Directions

The vulnerabilities demonstrated in this paper suggest that those using measures of user behavior or question-based countermeasures are fighting a losing battle. There must be a reevaluation of current best practices (22–24) for detecting fraud in online survey research. While developing technological countermeasures is an intuitive response, designing increasingly complex linguistic or logical tasks to detect synthetic respondents presents a trade-off. Such adversarial tasks risk introducing new forms of sample bias by inadvertently filtering out human respondents, particularly those with lower educational attainment, non-native English speakers, or individuals with lower cognitive engagement.

The most immediate and essential response is to demand transparency from panel providers regarding their practices for maintaining panel integrity. The opaque nature of many commercial panels is no longer tenable. As a standard practice, researchers should require providers to disclose their specific protocols in key areas such as:

1. Ongoing Panelist Validation: The frequency and nature of checks to re-verify identity and engagement over time.
2. Throttling Mechanisms: The limits imposed on respondent participation (e.g., surveys per day or week) to prevent the professionalization that incentivizes bot usage.
3. Panelist Professionalism: How many surveys has the panelist completed in the last 1, 7, and 30 days?

4. Panelist Quality Checks: How many response quality and attention checks has the panelist passed/failed? Has the panelist been reported for using AI in the past?
5. Location Checks: Is the panelist starting a survey from the state/region/country registered to the account? Is a VPN being used?

Panels unable or unwilling to provide this information should be considered high-risk. Moreover, panels should engage in aggressive user auditing and disclose results.

In parallel, researchers should consider re-evaluating their sampling strategies. Data from low-barrier convenience samples should be treated with skepticism until shown to be trustworthy. Integrated panels that manage the entire survey experience are not perfect, but are better positioned to detect fraud than panels that simply route traffic to third-party survey platforms like Qualtrics, where it is not possible to observe respondent behavior.

For research requiring high levels of data assurance, researchers should consider a return to more controlled recruitment methods, such as address-based sampling or other approaches (i.e., from the voter file, social media recruitment, or other commercial datasets) where deploying a bot for a single survey is infeasible. Another option is to use deeply vetted, longitudinally-managed panels. Ultimately, the social science community may need to reconsider its heavy reliance on unverified online surveys and reinvest in alternative data collection approaches—such as face-to-face interviews, student samples, administrative records, and other observational datasets—that are more resilient to this form of compromise.

That said, it is not clear that this is an intractable problem. Several potential solutions exist, though each comes with significant trade-offs:

- Identity Validation: We could adapt existing technology—like the kind used to verify a user's age to access adult websites or an Uber driver's license—to confirm a human is starting a survey. However, this approach has serious drawbacks. It raises considerable privacy concerns (especially for sensitive topics), faces technological hurdles, and still doesn't guarantee a human is the one who actually completes the survey.
- Secure Software: It's also possible to create secure survey tools that, like standardized testing software, take over a screen to block the use of AI assistance. This is likely a non-starter, though, since many people now take surveys on their mobile phones, where such lockdowns are impractical.
- Market Consolidation: Finally, the market might simply correct itself. The large number of cheap, low-quality survey panels available today could shrink, leaving a smaller, more reliable set of highly vetted panels to take their place.

5. Discussion

The findings presented in this paper paint a concerning picture for the future of online survey research. I have demonstrated that reasoning-based LLMs can complete surveys with plausible responses and can generate results that would bias measures of public opinion. They can mimic human personas, evade current detection methods, and be trivially programmed to systematically bias online survey outcomes. The era of having to only deal with crude bots and inattentive humans is over; the threat is now sophisticated, scalable, and potentially existential. The goal of this paper is not to advocate for the abandonment of online research, but to create an urgent call for adaptation, demanding new standards of transparency from panels and new methods of validation from researchers to meet this threat.

The immediate consequence is that the vast majority of our standard tools for data quality are now insufficient (my bot was able to enter data on Qualtrics pages that displayed a "protected by reCAPTCHA" badge). Simple metrics like completion times, straight-lining detection, and even standard attention checks are insufficient countermeasures for researchers or panel providers. Even complex checks like audio or video attention checks/human validations are overcome by models with ease (see SI Section S4.5-8). For those who study and rely on public opinion, the stakes are far higher. The ease with which these synthetic respondents can be engineered to respond with plausible but biased opinion—even with prompts written in a foreign language—turns public polling from a tool for democratic accountability into a potential vector for information warfare.

This research has its limitations. The analysis does not include a direct, side-by-side calibration of the synthetic data against a large-scale human reference sample. Therefore, while the tool's responses demonstrate high levels of internal coherence (e.g., rent scaling with income, psychometric consistency) and successfully evade detection, I do not claim they perfectly replicate the distributions or conditional averages found in human populations. The central argument of this paper is that the agent is plausible enough to pass existing quality filters, thereby breaking the assumption that coherence implies humanity. The work of calibrating these outputs against human data and identifying their unique statistical signatures remains a critical task for future research.

Moreover, the cost-benefit analysis for malicious actors is a snapshot in time, and the specific models tested are only a fraction of those available. However, the core finding is robust: the capability for this kind of undetectable fraud exists and is easily accessible. Future research must therefore shift from identifying simple bots to developing entirely new methods for data validation. The path forward will likely be a persistent technological arms race. Survey platforms and panel providers will undoubtedly update their defenses to block the approaches detailed in this paper, but new vulnerabilities will, in turn, be exploited by more sophisticated synthetic respondents. This dynamic suggests that ensuring data integrity will not be a matter of finding a single, permanent fix, but will require a continuous cycle of innovation. There is no magical fix, nor is there a magical bot.

I do not contend that synthetic respondents dominate online survey panels, and because of their sophistication and the limits of current detection methods we will likely be unable to exactly measure the magnitude of the problem. But even a small number is sufficient to cause meaningful errors. The critical difference with reasoning bots is the nature of the error they introduce. While traditional bots or inattentive humans add random noise that makes treatment

effects harder to detect, reasoning bots introduce non-random, systematic bias akin to demand effects. This vulnerability is particularly insidious because hypothesis-confirming data can be more difficult for even conscientious researchers to detect. Unlike random noise, which often attenuates effects, synthetic demand effects can produce results that appear plausible or even compelling, making it challenging to distinguish a genuine treatment effect from one artificially inflated by synthetic respondents. The risk is that such data, lacking the obvious red flags of traditional low-quality responses, could inadvertently lead to a proliferation of false positives, undermining the scientific process.

Future research must therefore shift from identifying simple bots to developing entirely new methods for data validation. The very imperfections in current AI responses, such as the 'tells' highlighted in this paper (e.g., superhuman accuracy on knowledge questions or unnaturally perfect logical consistency), offer a crucial first step.

Clever researchers will engineer questions that will trick specific models or families of models, but model development is progressing so quickly that these innovations are likely to be fleeting. Ensuring the continued validity of polling and social science research will require exploring and innovating research designs that are resilient to the challenges of an era defined by rapidly evolving artificial intelligence.

Materials and Methods

Design of the Autonomous Synthetic Respondent. This study utilized an autonomous synthetic respondent built in Python. The system features a two-layer architecture. The first layer is designed to interact with online survey platforms like Qualtrics. It parses questions (e.g., multiple-choice, sliders, text entry), extracts all question content including text from images and transcribes audio from videos, and takes and describes key-frames from video. It enters responses by simulating human-like behavior. This includes generating realistic mouse trajectories, applying reading times calibrated to an assigned persona's education level, and typing open-ended responses on a keystroke-by-keystroke basis, complete with plausible errors and corrections.

The second layer is a core reasoning engine powered by a large language model (LLM). For all primary experiments, OpenAI's 'o4-mini' model was used. The engine is initialized for each survey completion with a unique demographic persona (gender, race, age, education, partisan affiliation, income and state) and maintains a memory of its prior answers to ensure internal and longitudinal coherence. A single, general-purpose prompt of approximately 500 words instructs the agent on its core objective: to answer survey questions plausibly and consistently based on its assigned persona, without any question-specific guidance (see SI Section S1 for the full prompt).

Experimental Procedure and Data Generation. To systematically evaluate the synthetic respondent's capabilities, a series of experiments were conducted. For each experiment, the procedure was repeated 300 times, with a new, unique demographic persona assigned for each trial. Personas were generated via weighted random draws for partisanship, age, gender, race, education, household income, and U.S. state, with probabilities based on U.S. Census estimates. In total, the analysis comprises 43,800 distinct evaluations across 139 questions and 6,700 trials.

To speed the data collection process the reasoning engine was used to generate a full set of responses for each of the 300 personas without engaging the slower, human-mimicking front-end interface. An example of the full end-to-end process—including survey parsing, response generation, screenshots of the agent entering data into Qualtrics, and the corresponding Qualtrics output—is provided in SI Section S4.

Data availability. All code necessary to replicate the analyses and generate survey responses using the Large Language Model (LLM) is available on the Open Science Framework: https://osf.io/ektqr/?view_only=c75f58815f804164a6a7685bff7f1800. In accordance with ethical research principles and to prevent misuse, the scripts used to automate the final submission of these responses to Qualtrics have been withheld. These excluded scripts, whose sole function is to simulate human data entry, are not required to reproduce the study's analytical findings. Given that the findings of this study highlight the potential for such automation to undermine the integrity of online data collection, withholding this specific tool is the responsible course of action. Please contact the author for information about accessing the code.

ACKNOWLEDGMENTS. I thank Aral Cay and Derek Holliday for helpful research assistance and Dean Knox for pushing me to pursue the project.

1. National Academies of Sciences, Engineering, and Medicine, *Fostering Integrity in Research*. (The National Academies Press, Washington, DC), (2017).
2. TB Ustun, S Chatterji, A Mechbal, CJL Murray, The world health surveys in *Health Systems Performance Assessment: Debates, Methods and Empiricism*, eds. CJL Murray, DB Evans. (World Health Organization, Geneva), pp. 797–808 (2003).
3. G Groth-Marnat, *Handbook of Psychological Assessment*. (John Wiley & Sons, Hoboken, NJ), 5th edition, (2009).
4. J Bricker, et al., Changes in us family finances from 2013 to 2016: Evidence from the survey of consumer finances. *Fed. Res. Bull.* **103**, 1 (2017).
5. R Inglehart, C Welzel, *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. (Cambridge University Press, New York), (2005).
6. M Buhrmester, T Kwang, SD Gosling, Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychol. Sci.* **6**, 3–5 (2011).
7. J Bohannon, Mechanical turk upends social sciences. *Science* **352**, 1263–1264 (2016).
8. R Grim, A Lacey, Pete buttigieg's campaign used notoriously low-paying gig-work platform for polling (2020).
9. JA Krosnick, Response strategies for coping with the cognitive demands of attitude measures in surveys. *Appl. cognitive psychology* **5**, 213–236 (1991).
10. ME Browning, SL Satterfield, EE Lloyd-Richardson, Mischievous responders: data quality lessons learned in mental health research. *Ethics & Behav.* **34**, 303–313 (2024).
11. Z Zhang, et al., Beyond Bot Detection: Combating Fraudulent Online Survey Takers in *Proceedings of the ACM Web Conference 2022 (WWW '22)*. (Association for Computing Machinery, Lyon, France), pp. 699–709 (2022).
12. Demystifying the fraud mirage, (Research Defender), Ror summary (2025).
13. S Zhang, J Xu, A Alvero, Generative ai meets open-ended survey responses: Research participant use of ai and homogenization. *Sociol. Methods & Res.* p. 00491241251327130 (2025).
14. DM Oppenheimer, T Meyvis, N Davidenko, Instructional manipulation checks: A simple method to improve data quality in web-based experiments. *J. Exp. Soc. Psychol.* **45**, 847–850 (2009).
15. LP Argyle, et al., Out of one, many: Using language models to simulate human samples. *Polit. Analysis* **31**, 337–351 (2023).
16. C Betts, N Power, D Lynott, How we learnt to battle the bots. *The Psychol.* (2024) Available online at https://www.bps.org.uk/psychologist/how-we-learnt-battle-bots (accessed 25 June 2025).
17. A Ashokkumar, L Hewitt, I Ghezae, R Willer, Predicting results of social science experiments using large language models. Working paper. Equal contribution, order randomized (2024).
18. JS Park, et al., Generative agent simulations of 1,000 people. Working paper (2025).
19. JT Cacioppo, RE Petty, C Feng Kao, The efficient assessment of need for cognition. *J. personality assessment* **48**, 306–307 (1984).
20. SD Gosling, PJ Rentfrow, WB Swann Jr, A very brief measure of the big-five personality domains. *J. Res. personality* **37**, 504–528 (2003).
21. MB Petersen, M Osmundsen, K Arceneaux, The "need for chaos" and motivations to share hostile political rumors. *Am. Polit. Sci. Rev.* **117**, 1486–1505 (2023).
22. AJ Berinsky, MF Margolis, MW Sances, Separating the shirkers from the workers? making sure respondents pay attention on self-administered surveys. *Am. journal political science* **58**, 739–753 (2014).
23. JV Kane, J Barabas, No harm in checking: Using factual manipulation checks to assess attentiveness in experiments. *Am. J. Polit. Sci.* **63**, 234–249 (2019).
24. SJ Westwood, J Grimmer, M Tyler, C Nall, Current research overstates american support for political violence. *Proc. Natl. Acad. Sci.* **119**, e2116870119 (2022).
25. J Martherus, E Cook, A Podkul, Are bots taking online surveys? in *Proceedings of the 80th Annual American Association for Public Opinion Research (AAPOR) Conference*. (St. Louis, MO), (2025) Presentation in the "Questionnaire Design and Interviewing" track, May 15, 2025.
26. G Gallup, SF Rae, *The Pulse of Democracy: The Public-Opinion Poll and How It Works*. (Simon and Schuster, New York), (1940).
27. JR Lax, JH Phillips, The democratic deficit in the states. *Am. J. Polit. Sci.* **56**, 148–166 (2012).

28. S Ansolabehere, P Jones, Constituents' responses to congressional roll-call voting. *Am. J. Polit. Sci.* **54**, 583–597 (2010).
29. B Woodward, *Veil: The Secret Wars of the CIA, 1981–1987*. (Simon and Schuster, New York), (1987).
30. J Mummolo, E Peterson, Demand effects in survey experiments: An empirical assessment. *Am. Polit. Sci. Rev.* **113**, 517–529 (2019).
31. MR Tomz, JL Weeks, Public opinion and the democratic peace. *Am. political science review* **107**, 849–865 (2013).
32. L Aarøe, MB Petersen, Crowding out culture: Scandinavians and americans agree on social welfare in the face of deservingness cues. *The J. Polit.* **76**, 684–697 (2014).