

Research Report: Where Synthetic Users Fail

Documentation, Divergence, and Dissent in LLM-Generated Survey Respondents

A research report comparing synthetic respondents to real ones: accuracy tracks online documentation, synthetic populations diverge from reality, and politically and ethically motivated responses rarely appeared.

Jofish Kaye, Ph.D
Inflection AI

October 2026

Abstract

Large language models can now be prompted to answer survey questions "as" a person with a specified demographic profile, and a small industry of "synthetic respondent" products has grown up around the idea. This report evaluates the practice against ground truth. We take a real survey of 1,003 U.S. chatbot users and replicate it with six synthetic populations — three models (Claude, ChatGPT, Gemini) each run in two architectures (a single stateless prompt, and a persistent multi-turn conversation). We generated census-weighted personas, and used the same set of personas across all runs. We find that synthetic accuracy is governed less by question difficulty than by how heavily a behavior is *documented* online: error grows roughly 5.5x from well-documented to under-documented behaviors. Across the closed-ended questions, no synthetic population reproduces the real distribution: the average estimate is off by 22 percentage points and even the most accurate population still misses by 18 points, and the six populations disagree with one another as much as they disagree with reality. Finally we document a distortion that no model or architecture repairs: politically and ethically motivated behavior — boycotts, ethical defection — is almost entirely absent from synthetic data, a silence we argue follows from liability rather than from ignorance. We close with concrete guidance for practitioners.

1. Introduction

The appeal of synthetic survey respondents is straightforward. Recruiting a thousand people is slow and expensive; asking a language model to impersonate a thousand people is neither. Following early demonstrations that models conditioned on demographic backstories can

partially reproduce human opinion distributions (Argyle et al. 2023; Aher et al. 2023), both academic groups and commercial vendors have moved quickly toward using LLM output as a substitute — or at least a supplement — for human participants. A growing body of work cautions that these simulations flatten within-group variation and can misrepresent the groups they purport to model (Wang et al. 2025a), that their reliability has not been established (Wang et al. 2025b), and that they carry distinctive perils when used to replace survey data (Bisbee et al. 2024). The most ambitious demonstrations go beyond demographic conditioning: Park et al. (2023) built generative agents whose remembered experience drives behavior, and Park et al. (2024) extended this to simulate 1,052 specific real individuals by grounding each agent in a two-hour qualitative interview, reporting that the agents reproduced participants' General Social Survey responses about 85% as accurately as the participants reproduced their own answers two weeks later. That study is the closest large-scale precedent to ours, and instructive precisely for how it differs: it simulates *named individuals* from deep interview data and validates chiefly on closed-ended survey and experimental measures, whereas we sample *anonymous personas* from demographic marginals and stress-test the open-ended, brand-level, and politically sensitive responses where synthetic data is most likely to mislead.

Our motivation is also industrial rather than purely academic, and this shapes how we measure success. Much academic work validates synthetic respondents against established, standardized instruments — the General Social Survey, validated personality inventories, replications of published experiments — where a real answer key already exists. That is the right test for the question "can a model stand in for a human on a known measure?" But it is not our question. In product development the aim is to learn, faster and more cheaply than fielding a full study, what to expect from the users of a *new* product, feature, or concept — precisely the situations where no prior instrument, benchmark, or historical dataset exists. This is a harder setting for synthetic respondents: it removes the safety net of a well-documented ground truth and asks whether the method works where practitioners actually need it. Our study is designed around that setting, which is also why our findings about under-documented behaviors and absent potentially offensive opinions matter so directly — a new product is, by definition, under-documented.

In this study, we hold a real survey as ground truth and generate six synthetic replications of it, crossing three contemporary models with two architectures, while keeping the underlying personas identical across every run. This lets us ask not just "how good is synthetic data?" but "good for *what*?" — and to separate the effects cleanly. The report is intended as a practical companion to that question: detailed enough to guide methodology, and honest about where the method fails.

2. Method

2.1 The real survey

Our ground truth is a survey of $N = 1,003$ U.S. adults, recruited through Prolific and matched to the census on sex, age, and region, all of whom had used an AI chatbot in the week before

responding. (We report on this study [in an earlier research report](#).) The survey covers device and media use; awareness and frequency of use of sixteen named chatbots; the criteria people use to choose among them; reasons for abandonment; desired future capabilities; and standard demographics. It mixes closed-ended items (multi-select grids, five-point Likert scales, an eleven-point recommendation score) with open-ended free text. We note at the outset that a Prolific panel is not a representative cross-section of the U.S. public — it skews younger, more educated, and markedly more technologically engaged — and we return to this where it matters.

2.2 Six synthetic populations

For each synthetic run we generated 1,000 respondents. Each respondent's demographics — age band, gender, race and ethnicity, education, employment, census region and state, urbanicity, political identification, and self-rated technology comfort — were sampled from marginal distributions weighted to 2023 American Community Survey estimates for U.S. adults. We fixed the random seed and generation code across all runs, so that synthetic respondent i is the same person in every run; only the model and the architecture differ. Our persona-sampling design follows the approach popularized by population-simulation toolkits such as TinyTroupe (Salem et al. 2025), though we implemented a lightweight custom sampler rather than adopting the full toolkit. (We sampled marginals independently rather than from a joint distribution, a limitation we discuss in Section 9.)

We crossed three models — Anthropic's Claude Haiku 4.5, OpenAI's GPT-4o-mini, and Google's Gemini 2.5 Flash — with two architectures:

- **Stateless:** the persona and the full survey are presented in a single prompt, and the model returns all answers at once.
- **Persistent:** a system message establishes the persona; each question is then posed as a separate turn in one continuing conversation, so that later answers are conditioned on the model's own earlier ones — as a human respondent's would be.

Three models by two architectures yields six synthetic populations. Throughout, we abbreviate them, e.g., Gemini-SL (stateless) and Gemini-P (persistent).

2.3 Comparison

For closed-ended items we report, per run, the mean absolute error (MAE) between the synthetic and real proportion selecting each option, averaged within a question. We focus on the three closed items answered by all six runs — device use (Q3), chatbot adoption (Q4), and most-valued future roles (Q15) — for cross-model comparison, and draw on the fuller instrument for single-model depth where only one synthetic run covered it. For open-ended items we report response length, lexical diversity (type–token ratio, TTR), the rate of "null" answers, and the prevalence of specific themes.

3. Results I: What Kind of Fact Can Be Faked?

3.1 The documentation gradient

The single clearest pattern in our data is that a synthetic respondent's accuracy depends far more on whether an answer is *written down somewhere on the internet* than on how hard the question is. The device question (Q3) makes this visible because every item is the same kind of question — a binary report of whether you used something in the past week — so difficulty is held constant. What varies is documentation: smartphone market share and smart-speaker ownership are exhaustively tracked and discussed online, whereas whether you still own wired headphones, or listened to a podcast *with another person*, is recorded essentially nowhere.

Table 1. Synthetic accuracy by internet-documentation density (Q3, mean of three stateless models).

<i>Things the internet tracks closely</i>			
Which of the following have you used in the last week?	N=1003 study of US population	Synthetic (average score)	Off by
Apple iPhone	58%	52%	6%
Amazon Echo / Alexa	29%	30%	1%
Streaming device (Roku, etc.)	77%	82%	5%
<i>Things the internet doesn't know about</i>			
Which of the following have you used in the last week?	N=1003 study of US population	Synthetic (average score)	Off by
Wired headphones	44%	6%	38%
Listened to a podcast <i>with other people</i>	16%	0%	16%
Listened to music <i>with other people</i>	45%	2%	43%

Grouping the items by a coarse judgment of documentation density and averaging error across the three stateless models yields a stark, monotonic gradient: mean error rises from about 6% for heavily-documented behaviors to 33% for under-documented ones. Smartphone platform and smart speakers are reproduced within a few points; wired headphones (real 44%, synthetic 6%), listening to music with others (real 45%, synthetic 2%), and shared podcast listening (real 16%, synthetic ~0%) collapse almost completely.

3.2 Aggregate accuracy and the model effect

No synthetic population reproduces the real distribution, and — just as important — they do not agree with one another. There is no well-performing model to retreat to and no stable "good" choice: which population comes closest to reality changes from question to question, so a researcher who picked the population that best matched the truth on one question would have no assurance it would do so on the next. The differences between synthetic populations are as large as the differences between synthetic and real, which means model and architecture are consequential choices — but no choice is safe.

In Table 2, we demonstrate this with three questions from the survey: Q3, *Which of the following [technologies] have you used in the last week*; Q4, *Which of the following [AI assistants] do you use or have you previously used regularly?*; and Q15, *In the future, chatbots will be able to do lots of things better than now. Pick the five from this list that you think will be most valuable to you as chatbots improve and evolve*. Each of these is a similar form, with checkboxes for a list of options. We calculate the mean absolute error for each question by averaging the differences for each option between the synthetic user responses and the real user responses. A 22 percentage point average means that, option by option, synthetic estimates of how many people use a given device, use a particular AI assistant, or would value a particular role are off by 22 points on average.

Table 2. Mean absolute error against selected questions from the real survey, by model and architecture. Lower is better.

Model	Architecture	Q3	Q4	Q15	Avg
Claude	stateless	25.5pp	18.0pp	20.8pp	21.4pp
Claude	persistent	26.7pp	18.8pp	21.2pp	22.2pp
GPT-4o-mini	stateless	28.0pp	26.3pp	30.5pp	28.3pp
GPT-4o-mini	persistent	19.1pp	27.1pp	31.0pp	25.7pp
Gemini	stateless	14.6pp	21.8pp	19.8pp	18.7pp
Gemini	persistent	20.6pp	20.9pp	15.6pp	19.1pp

Similarly, asked which chatbots they use, real respondents name a long, messy tail: thirteen of the sixteen listed products drew measurable use. Claude and ChatGPT each surfaced only four; Gemini surfaced eleven, including CoPilot, Perplexity, Character.AI, and Bixby.

Table 3. Brand coverage: number of the 15 listed chatbots used at a measurable rate (>1%)

Population	Chatbots surfaced (of 15 options)
Real survey	13
Gemini (either architecture)	11
Claude, persistent	7
Claude, stateless	4
ChatGPT (either architecture)	4

4. Results II: Does a Persistent Persona Help?

Running the survey as a continuing conversation, so that each respondent "remembers" its earlier answers, is more expensive — roughly sevenfold, in API calls — and its benefits are specific rather than global. It barely changed aggregate accuracy (Table 2). However, it did change the character of the responses..

Most significantly, persistence suppressed a distinctive open-ended artifact we call the *self-diagnosis hallucination*. Asked what they wished a chatbot could do better (Q11), only 3% of real respondents mentioned memory or context retention — but 54% of stateless-Claude and 42% of stateless-ChatGPT respondents did. This is not a human wish; it is the *model's* best-known engineering limitation, surfacing as though it were a user priority. Under the persistent architecture the rate fell sharply (Claude 54%→23%; ChatGPT 42%→0%). An established persona stops breaking character to describe the technology's shortcomings.

Persistence also partially recovered under-documented behaviors: shared listening and wired-headphone use rose toward (though not to) real levels, and Gemini-persistent produced the single most accurate result in the study on future roles (15.6% MAE). The trade is real, though: persistence sometimes worsened brand-share accuracy, and it did nothing for the failure we turn to next.

So when should a researcher choose a persistent design? The pattern in our results suggests a simple rule: persistence helps when the quality of an answer depends on the respondent's *other* answers — that is, when you are modeling a coherent person rather than estimating a single statistic. Three cases favor it. First, when open-ended responses matter: persistence suppressed the self-diagnosis hallucination and produced terser, more human-like free text, because a respondent who has spent several turns in character stops breaking it to describe the technology's limitations. Second, when within-respondent consistency matters — any analysis that cross-tabulates one answer against another for the same individual, or that reads a response as part of a personal narrative — since a stateless respondent has no obligation to keep later answers consistent with earlier ones. Third, for self-referential or identity-

laden items (“Which chatbots do you use?”, “Is there something you wish a chatbot could help you with?”), where the persona anchor established over prior turns crowds out the model’s own priors.

This

Conversely, a stateless design is adequate, and more economical, when the goal is an aggregate estimate of a single closed-ended item: persistence did not reliably improve marginal accuracy on such questions and costs roughly seven times as many API calls. The practical guidance, then, is not “persistent is better” but “persistent for coherence, stateless for counts.” And neither choice touches the method’s structural limits — the documentation gradient and the absence of dissent persist under both — so persistence should be understood as a lever for the *character* of synthetic respondents, not a remedy for what they cannot represent at all.

5. Results III: Open-Ended Text and Laundered Opinions

5.1 A different register

Synthetic open-ended answers do not read like survey responses. Across runs they were two to four times longer than real answers (median 18–46 words versus 12 on Q11), markedly less lexically diverse (TTR 0.07–0.18 versus 0.27), and far less likely to give a null answer — 14% of real respondents simply wrote “nothing,” a modal response that stateless models almost never produced. Real answers are terse, idiosyncratic, and often blank; synthetic ones are fluent, complete, and uniform. Persistence improved brevity and null rates but did not close the gap.

5.2 Values-driven decisions that might offend disappear

The most consequential open-ended finding concerns why people quit. We asked an open-ended question: *Have you ever decided to stop using a particular chatbot? Why?* (Q9). We coded all 998 non-blank real responses and all 5,993 synthetic ones under a single scheme, distinguishing five kinds of *values-based objection* — objections to what a product or its maker stands for or does, rather than to what it does for the user: **conduct** (a company’s behavior: government and military contracts, corporate ethics), **individual** (a specific person, named or role-identified: “Elon Musk,” “their CEOs”), **privacy** (data collection and surveillance), **bias** (politically slanted output), and **environment** (concerns around water and energy consumption and pollution). The remaining categories are functional: quality, expense, other, and never having stopped at all.

Table 4. Open-ended reasons supplied for stopping using a particular chatbot.

reason	Real	Synth all	Claude persistent	Claude stateless	GPT persistent	GPT stateless	Gemini persistent	Gemini stateless
Values-based	9.9%	1.1%	3.5%	0%	0.1%	0.1%	0.7%	2.1%

objections								
— Conduct	4.8%	0.1%	0.2%	0%	0%	0%	0.1%	0%
— Individual	1.8%	0%	0.1%	0%	0%	0%	0%	0%
— Privacy	1.6%	0.8%	3.0%	0%	0.1%	0.1%	0.2%	1.6%
— Bias	1.1%	0.2%	0.2%	0%	0%	0%	0.4%	0.5%
— Environment	0.6%	0%	0%	0%	0%	0%	0%	0%
Quality	27.4%	46.4%	80.4%	45.5%	58.9%	16.1%	37.1%	40.5%
Cost	3.4%	0%	0%	0%	0%	0%	0%	0%
Other	5.2%	8.2%	13.4%	8.7%	1.1%	0.5%	4.8%	20.8%
Never stopped	54.1%	44.3%	2.7%	45.8%	39.9%	83.2%	57.4%	36.6%

Among real respondents, values-based objections accounted for 9.9% of all answers — and about one abandonment in five. They are dominated by their most pointed forms: conduct (4.8%) and individuals (1.8%) together make up nearly three-quarters of the group. Real people named names. They quit ChatGPT over OpenAI's defense-department contract, refused Grok because of Elon Musk, called executives "full blown sociopaths," and dropped products over environmental footprint, government ties, and corporate scandal.

The synthetic populations tell a different story, and not merely a quieter one. Across all six populations, values-based objections fell to 1.1%, with the biggest category by far being identifying privacy issues in one Claude run. No synthetic respondent mentioned a defense contract, a boycott, a government tie, or the environment. Exactly one synthetic respondent in six thousand pointed at a person, and even that lone response — "Elon's whole thing with it felt off to me" — immediately retreats to a safe functional justification ("and honestly it didn't seem as useful as ChatGPT or Claude anyway"). The strongest real statements have no synthetic counterpart at any strength.

The laundering operates at two levels. First, opinions that might offend are not converted into some other category of complaint so much as deleted: synthetic populations describe abandonment readily — nearly every Claude-persistent and most Gemini respondents produced a quitting story, though GPT-stateless, at 83% never-stopped, errs in the opposite direction — but the reasons offered are overwhelmingly bland quality narratives ("tried it, stuck with what I know"). Second, and more subtly, even the residue of values-based objection that survives is transformed into its safest available form. The one synthetic population with a visible values-based rate, Claude-persistent at 3.5%, achieves it almost entirely through generic privacy hedging ("I wasn't comfortable with how much data these services collect") — thirty near-identical paraphrases of a single cautious persona. Where real values-based quits are 73% conduct and individuals and 16% privacy, synthetic ones invert the composition: 6% conduct and individuals, 77% privacy. The synthetic panel does not just mute opinions that might offend someone; it substitutes a deferential, unfalsifiable worry about data for pointed objections to

named power. Two further absences complete the picture: cost, cited by 3.4% of real respondents, appears zero times in 5,993 synthetic responses, and no synthetic respondent ever objects to the model family generating it — Claude personas complain about Gemini's bias, Gemini personas about ChatGPT's, never about their own.

We do not think this is a knowledge failure. The models plainly "know" that such boycotts occur; the fact is itself well-documented, which makes this the one large distortion the documentation gradient cannot explain. Nor is it repaired by memory: persistence, which fixed the self-diagnosis hallucination and halved self-reference bias, leaves this untouched. We argue it is structural, and rooted in liability. When a human respondent writes "I quit Grok because of Elon Musk," the researcher recording it bears no legal risk: the respondent is voicing a protected opinion. When a commercial model generates the same sentence in a synthetic persona, the firm operating the model may be cast as the speaker, and there is now an active body of law on precisely this exposure. *Walters v. OpenAI*, the first U.S. defamation suit over model output, was resolved in OpenAI's favor in 2025 on narrow, fact-specific grounds rather than by any holding that model output cannot defame (Walters 2025); legal scholarship has laid out the category of "large libel" liability in detail (Volokh 2023); and *Garcia v. Character Technologies* declined at the pleading stage to treat model output as categorically protected (Garcia 2025). A developer facing this landscape will, sensibly, constrain its models against generating pointed negative claims about named individuals and companies.

That caution is good risk management. But when the same model is repurposed as a survey respondent, its caution becomes the synthetic population's silence — and the silence falls precisely on dissent against named power: boycotts, ethical defection, refusal on the basis of a company's conduct. This is disproportionately the voice of the disaffected and the activist, and it is exactly the signal much fairness- and accountability-oriented research exists to surface. A synthetic panel presents as a complete population while having quietly deleted these voices — or worse, replaced them with a polite privacy murmur that gives the aggregate a reassuring appearance of critical texture. The deletion is invisible to anyone without real data to compare against. The danger is not measurable, correctable noise; it is a silent censorship along a dimension that does not announce itself in the aggregates.

6. Recommendations for Practice

1. **Be wary of politically or ethically charged questions — especially where the honest answer means criticizing a named company or person.** Opinions that might offend somebody are laundered out silently. We see strong evidence that politically-charged issues, where real humans make decisions on the basis not just of who or what they like, but also who or what they actively dislike, cannot be well-represented with current LLMS – and the increasing liability concerns make change in this domain unlikely.
2. **Do not trust answers to under-documented behaviors.** If the internet does not already track the fact, the synthetic estimate is a guess. This is testable: bucket your

items by documentation density and validate the low-density ones against real data before relying on any of them.

3. **Report model and architecture as methodology.** They are not implementation details. Different models from identical personas produced populations that differed from one another as much as from reality. State both, as you would a sampling frame.
4. **Use persistent conversations when behavior, open-ended text, or internal consistency matter.** Expect it to fix character problems — self-reference, self-diagnosis, life-texture — more than to move accuracy, and budget for the extra cost.
5. **Anchor to a real sample — always.** Every finding in this report was invisible until we held the synthetic answers against real ones. A modest anchor ($n \approx 100$ can suffice to catch the large distortions) is the cheapest insurance in synthetic research.

7. Limitations

Our ground truth is a Prolific panel, itself to some degree unrepresentative of the U.S. public, so some apparent synthetic "errors" partly reflect the panel's own skew rather than the model's. We sampled persona demographics from independent marginals rather than a realistic joint distribution. We studied one instrument, in one domain (consumer AI use), at one time, with three specific model versions; generalization to other domains, instruments, and model generations is untested. Our documentation-density coding and open-ended theme coding were manual and not independently double-coded, and we did not systematically vary temperature or prompt phrasing, both plausible moderators. The legal-causal account of the unpopular opinion finding (Section 5.2) is a reasoned interpretation grounded in published case law, not a measured mechanism; we observe the suppression, and infer its most likely cause.

On augmentation and external validity. A reasonable objection is that few production systems use a public model with only a persona prompt, as we do; commercial synthetic-respondent vendors typically augment with retrieval, knowledge graphs, or fine-tuned or proprietary models. Two points follow. First, our absolute error figures should therefore be read as a floor-level baseline for un-augmented models, not as the performance ceiling of a production pipeline; a well-built system will do better on well-documented quantities. Second, and more importantly, augmentation does not reach our two central findings. Retrieval and knowledge graphs improve accuracy only where documentation exists to retrieve — precisely the high-documentation regime where models already succeed. They cannot supply a distribution for behavior that is nowhere recorded, which is the entire content of the documentation gradient, and which is most acute in the new-product setting that motivates industrial use. Likewise, the absence of politically and ethically motivated dissent is not a knowledge deficit that a graph could fill; it is a liability-driven constraint that fine-tuning on real dissent would reintroduce the very exposure that produced. We therefore expect both findings to survive augmentation, even as augmentation narrows the aggregate accuracy gap. The practical guidance for researchers is correspondingly specific: invest augmentation effort where it pays — well-documented, closed-ended quantities — and do not expect it to rescue under-documented behaviors or suppressed dissent, which require real respondents regardless of how the synthetic pipeline is built.

8. Conclusion

Synthetic survey respondents, built on public models without including additional data sources, are good enough to be genuinely useful and genuinely dangerous, often in the same study. They reproduce well-documented facts, broad attitude structures, and the rank-order of preferences — enough to pilot an instrument or pressure-test a hypothesis. They collapse the long tail of real products, over-formalize open-ended voice, miscalibrate badly on anything the internet does not already track, and fall silent exactly where people voice pointed opinions about named power. Which of these you get depends on choices — model, architecture, how you elicit identity — that are easy to treat as incidental and are not. Every synthetic population we generated diverged substantially from the real one — the most accurate still wrong by nearly a fifth on average — and they diverged from each other just as widely, with no population reliably closest across questions.

The through-line of every recommendation here is a single discipline: keep a real sample in hand. This discipline matters most in exactly the setting that motivates industrial use: when the object of study is a new product, feature, or concept, there is no established instrument to fall back on and no historical data to check against, so a modest real anchor is not a luxury but the only thing standing between a useful estimate and a confident fiction. A synthetic respondent is never uncertain, and that is precisely why it cannot, by itself, tell you when it is wrong.

References

- Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research (Vol. 202, pp. 337–371). PMLR.
- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337–351.
- Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4), 401–416.
- Garcia v. Character Technologies, Inc. (2025). U.S. District Court, Middle District of Florida. Order on motion to dismiss.
- Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23).

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., & Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Salem, P., Sim, R., Olsen, C., Saxena, P., Barcelos, R., & Ding, Y. (2025). TinyTroupe: An LLM-powered Multiagent Persona Simulation Toolkit. *arXiv preprint arXiv:2507.09788*.

Volokh, E. (2023). Large libel models? Liability for AI output. *Journal of Free Speech Law*, 3, 489–558.

Walters v. OpenAI, LLC (2025). Superior Court of Gwinnett County, Georgia. Order granting summary judgment, May 19, 2025.

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025a). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3), 400–411.

Wang, Q., Wu, J., Jiang, Z., Tang, Z., Luo, B., Chen, N., Chen, W., Wang, H., & He, B. (2025). LLM-based human simulations have not yet been reliable. *arXiv preprint arXiv:2501.08579*.

Appendix A: Survey Questions

The six questions used for synthetic replication are reproduced below. Closed-ended options are shown as presented to respondents; two deliberate attention-check "honeypot" options (a nonexistent device and a nonexistent chatbot) were included to detect inattentive responding, and neither was selected by any respondent in the cleaned real sample. Prolific users were told that we wanted to "ask them about their technology use"; they did not know we were looking for chatbot users.

Q3. Which of the following have you used in the last week? *(select all that apply)*

- A barcode-based household multimedia system [a]
- A chatbot, like Claude or ChatGPT [b]
- A Google Nest smart speaker
- A smartwatch or fitness tracker, like a Garmin or Fitbit
- A smartwatch, running Apple's watchOS
- A smartwatch, running Google's WearOS
- A video streaming device, like a Smart TV, Roku, Chromecast, Apple TV, or Amazon Fire TV
- An Amazon Echo smart speaker, running Alexa/Alexa+
- An Android phone
- An Apple Homepod, or a smart speaker from Sonos, Bose, JBL, or Logitech
- An Apple iPhone
- Apple CarPlay
- Earbuds or other wireless headphones
- I have listened to a podcast by myself
- I have listened to a podcast with other people
- I have listened to streaming music by myself
- I have listened to streaming music with other people
- Wired headphones

[a] This was a "honeypot": an option presented to catch people who were not answering the questions honestly, deliberately included first in the alphabetized list. 64 respondents chose this option, and were eliminated from the survey.

[b] Respondents were required to select this option to continue with the survey.

Q4. Which of the following do you use or have you previously used regularly? *(choose up to 10)*

- Afitoto [c]
- Alexa or Alexa+
- Chai
- Character.AI
- ChatGPT

- Claude
- CoPilot
- Deepseek
- Duck.ai
- Gemini
- Grok
- Meta AI
- Perplexity
- Pi.AI
- Replika
- Siri

[c]: Afitoto was another honeypot option; only 3 users selected it, and were eliminated from the study.

Q5. How often do you use these chatbots? *(asked for each chatbot the respondent selected; same list as Q4)*

- Several times a day
- At least once a day
- At least once a week
- At least once a month
- I used this regularly in the past, but not any more
- I tried this once or twice in the past, but not any more

Q9. Have you ever decided to stop using a particular chatbot? Why? *(open-ended)*

Q11. Is there something you wish a chatbot could help you with, but haven't found one that does it well? *(open-ended)*

Q15. In the future, chatbots will be able to do lots of things better than now. Pick the five from this list that you think will be most valuable to you as chatbots improve and evolve. *(select five)*

- Assistant
- Best Friend
- Chef
- Coach
- Companion
- Crisis Triage
- DJ
- Editor
- Encourager
- Grounding Partner
- Listener

- Mentor
- Parent
- Partner
- Teacher

(each was shown with a one-line description, e.g. "Assistant – helps with daily tasks and keeps things running")

AI and Tool Usage Statement

The survey was hosted on SurveyMonkey, and Prolific was used to recruit and reward participants. Claude (Anthropic) was used to assist with data analysis, first-pass qualitative coding of open-ended responses, and drafting. All codes were reviewed and revised by the author, and all findings were verified against the underlying data. (The irony of using an AI to help document AI's blind spots is not lost on us, it's also the point: the tool is genuinely useful, and it still needs a human holding the answer key.)

Companion to the blog post "Where Synthetic Users Fail: Things the internet doesn't know and Politics". Code and datasets: <https://github.com/jofish-inf/synthetic-survey-respondents>