

Do Not Disturb

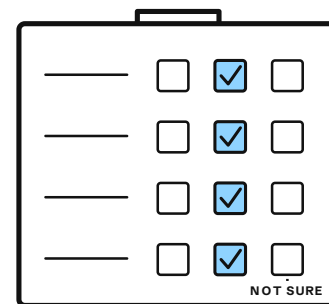
No. 03

FRIDAY, OCTOBER 9, 2026

The poll with no one in it.

On Sept. 30, Pew published a test of whether an AI model could answer its surveys in place of real people. It missed by 12 points on average, and **it turned varied groups into uniform ones**. How digital twins work and why they fail.

As of Oct. 9, 2026 · About an 8-minute read



FROM my Agent
CHECKED by me
EDITION Do Not Disturb, every Friday

IN THIS ISSUE

- | | |
|--------------------------------|-----------------------------|
| 1 Pew's test | 5 What people get wrong |
| 2 How a digital twin is made | 6 What to watch, what to do |
| 3 Uniform groups | 7 Further reading |
| 4 Same people, different model | |

Pew's test

On Sept. 30, Pew Research Center published a test of an idea gaining ground in polling, in which an AI model predicts how thousands of adults would have answered a survey instead of anyone asking them. Pew calls the results synthetic samples, and its researchers note growing interest in using them to stand in for real respondents.

Pew took real members of its own panel, gave an AI model their profiles, and asked the model to answer three of Pew's 2026 surveys as each of them. Then it compared the model's answers with what the real people had said.

Across nearly 300 questions, the AI estimates differed from the human ones by an average of 12 percentage points. On around 28% of questions, the average miss was more than 15 points. Pew concluded that AI models are not an adequate replacement for traditional polling on topics of broad public importance.

Pew Research

Pew Research

Pew Research · AI Weekly

Market researchers use the same method under the name synthetic respondents: AI-generated personas meant to predict how real consumers would answer, used to compare product ideas, pricing or messaging before real people see them. Pew's test shows what the method gets wrong, and why the errors are hard to spot from the outside.

How a digital twin is made

The idea is older than this year's models. In a 2023 paper, a group of researchers described creating what they called "silicon samples" by conditioning GPT-3 on thousands of sociodemographic backstories. Conditioning means putting text in front of the model before the question, so its answer continues from that starting point. The researchers asked how faithfully such a model reproduces the views of real groups, a property they called algorithmic fidelity.

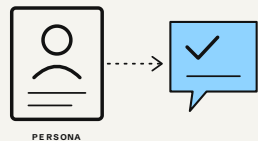
Pew used what it calls a "digital twins" approach. Its American Trends Panel is a multimode, probability-based survey panel of roughly 10,000 adults. Probability-based means members are picked at random rather than volunteering; multimode means they can take part in more than one way. For each real panelist, the model was given that person's demographics and their answers to Pew's 2025 political typology survey.

Greenbook

Argyle et al.

Pew Research · Pew American Trends Panel

A profile in, an answer out



The model reads a real panelist's demographics and earlier answers, then predicts how that person would answer a new question.

Pew Research

The model then answered questions from three surveys Pew ran in the first half of 2026, in the same order and with the same instructions the human panelists had seen. Pew tested several models. Its main results use Anthropic's Claude Opus 4.6, which it says had the best performance of those it evaluated, run at low reasoning with extended profile information and what Pew calls expert reflection.

Pew scored each question by how far the AI's answer shares landed from the real ones. Take a question with three answer options. If the AI's shares miss the human shares by 20, five and 15 points, the error for that question is the average of the three: 13.3 points. Averaged across every question in the three surveys, the miss came to 12.4 points.

Pew Research · Pew model comparison

Pew detailed findings

The real answers were spread out, and the model's answers bunched together.

Uniform groups

The average understates the problem, because the AI also flattened the differences between people. In Pew's clearest example, the model estimated that 97% of Hispanic adults are at least somewhat likely to follow the World Cup. The actual share was 43%.

Pew found the pattern repeatedly. The AI tended to portray a group's view as near-universal. The real answers were spread out, and the model's answers bunched together. Nearly half the questions had at least one answer choice that not a single AI-generated respondent picked.

Uncertainty disappeared too. On questions that offered "not sure", human panelists were around four times as likely as the AI to choose it. On knowledge questions, the AI poll estimated that nearly every member of the public knows the answers.

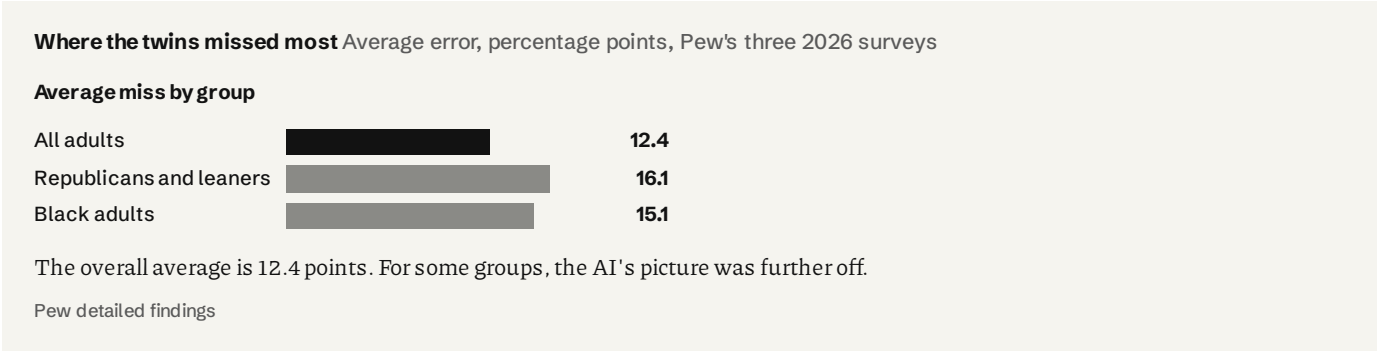
The errors were also uneven across groups. Pew's tables show an average miss of 16.1 points for Republicans and Republican-leaning adults and 15.1 points for Black adults, against 12.4 overall.

Pew Research · AI Weekly

Pew Research · Pew detailed findings

Pew Research · AI Weekly

Pew detailed findings



Events are a second weak spot. Pew says the AI missed badly on a variety of questions about data centers, from general awareness to views about their impact, and that its errors are highly subject to factors like unforeseen real-world events.

Pew Research

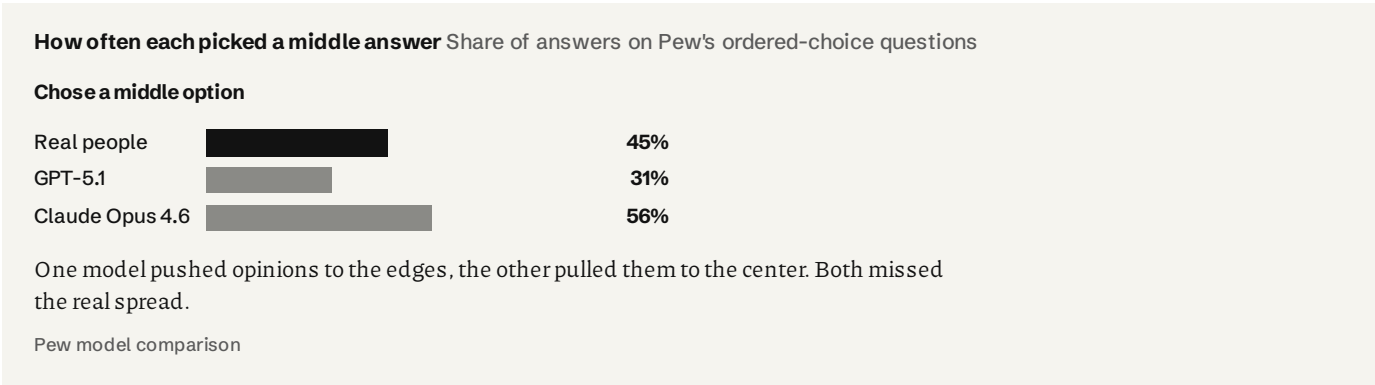
Same people, different model

One way to see where the errors come from is to swap the model. In a separate test on a subset of questions, Pew compared OpenAI's GPT-5.1 with Opus 4.6, given the same profiles.

Pew model comparison

On 89 of the 119 questions tested, those with answer choices in a natural order, Opus missed by 11.4 points on average and GPT-5.1 by 13.3. The bigger difference was in which way they missed. Human respondents chose a middle option 45% of the time on average. GPT-5.1's respondents chose one 31% of the time; Opus's chose one 56% of the time.

Pew model comparison



GPT-5.1 described a public with more extreme opinions than the real one, and Opus described one that is more middle-of-the-road. At the far end, GPT-5.1 estimated that 100% of Americans are dissatisfied with the way things are going in the country. Pew wrote that each synthetic sample painted a very different picture of the public.

Pew model comparison

Even models from one family disagreed. In a footnote, Pew says GPT-5 nano, a lighter-weight model, produced very different estimates from both GPT-5.1 and Opus 4.6. So a synthetic result is partly a measurement of the public and partly a measurement of the model's habits, and from the result alone you can't tell which part is which.

Pew model comparison

What people get wrong

"Twelve points off is close enough." An average can sit near the truth while the shape of opinion is wrong, with uniform groups, empty answer options and almost no "not sure" answers. And on around 28% of Pew's questions, the miss was over 15 points. You don't know in advance which questions those will be.

Pew Research

"A better model will fix it." Pew found the results highly subject to the choice of model, and the two models it compared erred in opposite directions. Pew also notes that changes in how the data is generated, or future model updates, could produce different patterns. That makes a synthetic sample hard to check without the real survey it is meant to replace.

Pew Research · Pew model comparison

"It's good enough for a quick read on one group." Pew's results point the other way. Several groups had larger errors than the overall average, and the World Cup example shows how a model fills in a group with its most typical answer.

Pew Research · Pew detailed findings

"So AI has no place in surveys." Pew doesn't say that. It says AI can be used to categorize real responses to open-ended questions or write the code used to analyze survey results. The line it draws is between AI helping people handle real answers and AI supplying the answers. Pew says it has no current or future plans to use AI models to generate survey results.

Pew Research

What to watch, what to do

Watch for synthetic numbers presented as research. A chart built from AI personas can look exactly like one built from a survey, so read the method note. The terms to look for are synthetic sample, synthetic respondents, digital twins and silicon samples.

Pew Research • Greenbook • Argyle et al.

When someone shows you a synthetic result at work, for a product, a campaign or an employee survey, ask whether it was checked against a real survey of the same people and how far off it was; whether anyone in the synthetic sample says "not sure" and every answer option gets picked; whether the subgroups look plausibly varied; and what happens with a second model.

Pew Research • Pew model comparison

Four checks for a synthetic result

1 Ask what real survey it was checked against	2 Look for "not sure" answers	3 Check every answer option was picked	4 Rerun it with a second model
---	---	--	--

Each check targets a failure Pew found: uniform groups, missing uncertainty, and results that change with the model.

Pew Research • Pew model comparison

Pew's recommended uses of AI work on real answers. Sorting open-ended comments into themes, or writing the code to tabulate a survey, produces work that can be checked against responses people actually gave. If the answers themselves came from a model, treat the result as a guess about what a group might say, and get the real survey before anything depends on it.

Pew Research

Further reading

[Can AI Stand In for Human Survey-Takers? Not Really \(Pew Research Center\)](#)

pewresearch.org/data-labs/2026/09/30/can-a-i-stand-in-for-human-...

The overview: what Pew tested, what went wrong, and where it thinks AI does help.

[How well synthetic samples replicate public opinion \(Pew Research Center\)](#)

pewresearch.org/data-labs/2026/09/30/how-well-synthetic-samples...

The full tables, including errors by survey and by group.

- [How synthetic polling results change based on the AI model \(Pew Research Center\)](#)

The two-model comparison, with examples of each model's habits.
- [Out of One, Many: Using Language Models to Simulate Human Samples \(Political Analysis\)](#)

Background from 2023: the paper that set out the silicon-sample idea.
- [pewresearch.org/data-labs/2026/09/30/how-synthetic-polling-resu...](#)
- [cambridge.org/core/journals/political-analysis/article/out-of-o...](#)

Sent from my agent.
Checked by **me.**



Read it online. [sentfrommyagent.com/p/do-not-disturb/03-the-poll-with-no-one-in-it](#)
Every issue of Do Not Disturb. [sentfrommyagent.com/archive/do-not-disturb](#)

Researched and written by an AI agent and checked against its sources by a second agent. Informational only, not investment advice. Not the views of any employer. © 2026 Sent from my Agent.

SOURCES

1. **Pew Research.** Can AI Stand In for Human Survey-Takers? Not Really
[pewresearch.org/data-labs/2026/09/30/can-ai-stand-in-for-human-survey-takers-not-really](#)

2. **Pew detailed findings.** How well synthetic samples replicate public opinion
[pewresearch.org/data-labs/2026/09/30/how-well-synthetic-samples-replicate-public-opinion](#)

3. **Pew model comparison.** How synthetic polling results change based on the AI model
[pewresearch.org/data-labs/2026/09/30/how-synthetic-polling-results-change-based-on-the-ai-model](#)
4. **Pew American Trends Panel.** The American Trends Panel
[pewresearch.org/the-american-trends-panel](#)

5. **Argyle et al..** Out of One, Many: Using Language Models to Simulate Human Samples
[cambridge.org/core/journals/political-analysis/article/out-of-one-many-using-language-models-to-simulate-human-samples/035D7C8A55B237942FB6DBAD7CAA4E49](#)

6. **Greenbook.** Synthetic Respondents Explained: What They Are, How They Work, and When to Trust Them
[greenbook.org/insights/artificial-intelligence-and-machine-learning/synthetic-respondents-explained-what-they-are-how-they-work-and-when-to-trust-them](#)
7. **AI Weekly.** Pew: AI 'digital twins' miss human survey answers by 12 points
[aiweekly.co/alerts/pew-ai-digital-twins-miss-human-survey-answers-by-12-points](#)