

FOR RELEASE SEPTEMBER 30, 2026

Can AI Stand In for Human Survey-Takers? Not Really

Synthetic surveys do a poor job of estimating the public mood and are especially error-prone for some subgroups, including Republicans and Black adults

BY Athena Chapekis, Arnold Lau, Samuel Bestvater, Sono Shah, Andrew Mercer and Aaron Smith

FOR MEDIA OR OTHER INQUIRIES:

Sogand Afkari, New Media Manager
Nida Asheer, Senior Communications Manager

202.419.4372

www.pewresearch.org

RECOMMENDED CITATION

Pew Research Center, September 2026, "Can AI Stand In for Human Survey-Takers? Not Really"

About Pew Research Center

Pew Research Center is a nonpartisan, nonadvocacy fact tank that informs the public about the issues, attitudes and trends shaping the world. It does not take policy positions. The Center conducts public opinion polling, demographic research, computational social science research and other data-driven research. It studies politics and policy; news habits and media; the internet and technology; religion; race and ethnicity; international affairs; social, demographic and economic trends; science; research methodology and data science; and immigration and migration. Pew Research Center is a subsidiary of The Pew Charitable Trusts, its primary funder.

© Pew Research Center 2026

About this research

This Pew Research Center report examines whether AI-generated survey data can accurately duplicate the results of high-quality public opinion polls.

We conducted this methodological experiment to better understand an emerging method in survey research. The Center believes that speaking to the public is essential to measuring public opinion and has no current or future plans to use AI models to generate survey results. Read more about our [AI policy](#).

For this study, we had an AI model take on the role of real respondents from our [American Trends Panel](#) (ATP) and answer the same surveys administered to human panelists. These surveys included many different question formats and asked about a wide range of topics, attitudes and behaviors.

Why did we do this?

The Center regularly studies and evaluates [advances and trends in polling research](#). As AI-based polling becomes [more widely used in the industry](#), we wanted to understand the data these “synthetic samples” produce and how it compares with high-quality surveys of real humans on topics of broad public interest. That question is the focus of this study; it does not address other uses of AI models in polling.

How did we do this?

We used a “digital twins” approach for this study, asking an AI model to adopt the personas of real humans who are members of the ATP. We gave the model a [wide range of information about each panelist](#), including their self-reported demographic information and their responses to questions on our [2025 political typology survey](#). We then showed the model questions from three different ATP surveys conducted in the first half of 2026 and asked it to respond to those same questions as the assigned persona. The model was given each question on the survey in order from start to finish, with exactly the same instructions and survey programming that the real humans who took it received.

To make a direct comparison between our AI and human polls, we only replicated surveys for the ATP panelists from each wave who had also completed the political typology survey. As a result, figures for U.S. adults in this report may differ slightly from those previously published by the Center.

We tested multiple models, but all synthetic results unless noted otherwise come from Anthropic’s Claude Opus 4.6. At the time we conducted this study, Opus 4.6 was the most recent

Anthropic model available for commercial use and had the best performance of several we evaluated.

Here are the [survey questions from Wave 185](#), [Wave 190](#) and [Wave 192](#) used for this analysis, the [detailed responses](#) from the ATP and AI surveys, and the [methodologies](#) for the original ATP survey waves and our AI surveys.

Can AI Stand In for Human Survey-Takers? Not Really

Synthetic surveys do a poor job of estimating the public mood and are especially error-prone for some subgroups, including Republicans and Black adults

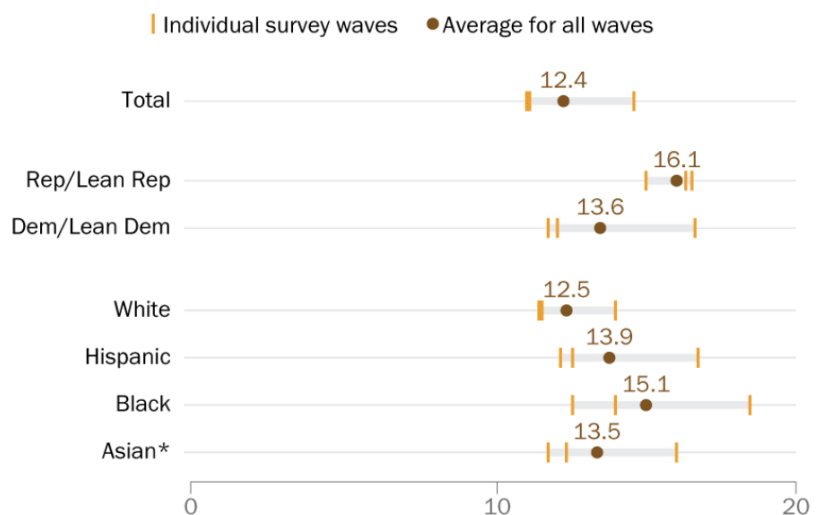
As artificial intelligence becomes more advanced, there is growing interest in using it to stand in for real respondents in public opinion surveys. Put simply: Instead of contacting large numbers of people and asking them what they think, pollsters can ask an AI model to predict how those people *would have* answered a certain question.

Pew Research Center has long sought to better understand [new developments in public opinion research](#), and we wanted to learn more about how this AI-based polling works.¹ **So we ran an experiment.** We developed a state-of-the-art process for fielding AI surveys and compared their results with three recent survey waves from our [American Trends Panel](#) (ATP) taken by human respondents at roughly the same time.

This experiment taught us that, at this time, **AI models are not an adequate replacement for traditional polling on topics of broad public importance.** Our AI-

AI-generated surveys consistently miss results of human polls

Average absolute survey error (in percentage points) between human and synthetic respondents on 3 replicated survey waves



* Estimates for Asian adults are representative of English speakers only.
Source: Surveys of U.S. adults and “digital twins” synthetic analysis using Claude Opus 4.6 set to low reasoning with extended profile information and expert reflection. Surveys were conducted Jan. 20-26, 2026 (replicated March 9-12 and April 7-10), March 23-29, 2026 (replicated March 30-April 2), and April 20-26, 2026 (replicated April 27-May 1).
“Can AI Stand In for Human Survey-Takers? Not Really”

PEW RESEARCH CENTER

¹ This study focuses solely on the use of AI as a survey-taker. It does not address other uses of synthetically generated data, such as estimating or imputing missing values in existing datasets.

generated survey results struggled to accurately reproduce the findings from high-quality public opinion polls in a variety of ways.

Here are some of the main issues we encountered:

Results of AI polls differed – often by quite a lot – from results of human surveys

Across nearly 300 individual survey questions, the estimates produced using our AI respondents differed from their human counterparts **by an average of 12 percentage points**.

How we measured the difference between our human and AI survey-takers

To compare results from our human and AI survey-takers, we calculated the **average absolute percentage point error** for each survey question. Consider a question with three different response options that our human and AI survey-takers answer as follows:

- 50% of humans and 70% of AI respondents choose Option A (a 20-point difference).
- 30% of humans and 25% of AI respondents choose Option B (a 5-point difference).
- 20% of humans and 5% of AI respondents choose Option C (a 15-point difference).

The average error across those three options would be $(20+5+15)/3$, or 13.3 percentage points, for that question.

The average difference exceeded 15 points on around 28% of questions we asked. And it was not uncommon to see differences exceeding 20, 30 or even 40 percentage points on individual answers to some questions.

Read more about [*how well synthetic samples replicate public opinion*](#).

Major misses on ‘timely and topical’ political issues

Our AI survey missed the mark on numerous questions about current events in early 2026. Among other things:

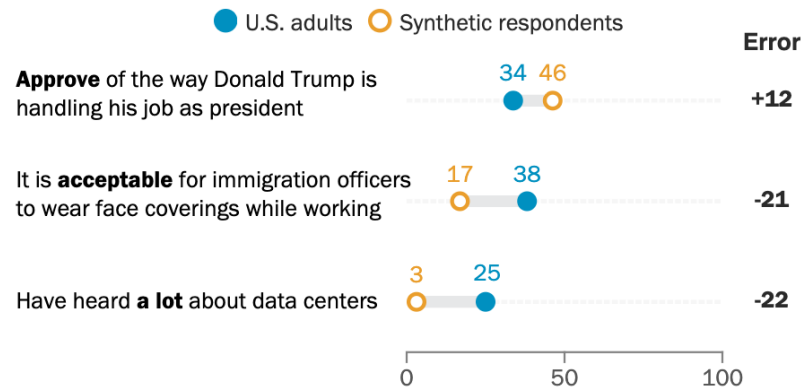
- It overstated the share of Americans who approved of President Donald Trump’s job performance at the time, despite the availability of long-standing trend data.
- It greatly underestimated the share of Republicans who think it’s acceptable for immigration officers to wear face coverings, resulting in a large overall error.
- It missed badly on a variety of questions about data centers, from general awareness to views about their impact.

Often, these misses were unpredictable. On questions about cost-of-living concerns, our model somewhat *overestimated* the shares of Americans who are concerned about the cost of healthcare and consumer goods while greatly *underestimating* the share who are worried about the price of electricity.

Read more about [synthetic surveys and “timely and topical” questions](#).

Notable errors in AI polls on ‘timely and topical’ issues

% who (say) ...



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

Source: Surveys of U.S. adults and “digital twins” synthetic analysis using Claude Opus 4.6 set to low reasoning with extended profile information and expert reflection. Surveys were conducted Jan. 20-26, 2026 (replicated March 9-12 and April 7-10), and April 20-26, 2026 (replicated April 27-May 1).

“Can AI Stand In for Human Survey-Takers? Not Really”

PEW RESEARCH CENTER

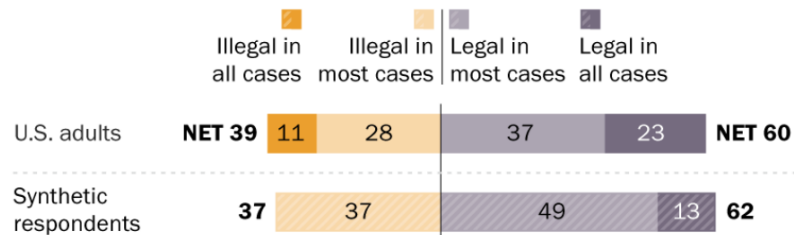
AI consistently avoids some answers but piles onto others

Human opinion is extremely diverse, but our AI respondents often avoided certain answers altogether. Nearly half the questions we asked had at least one answer choice that was not selected by a single AI-generated respondent.

These errors can distort true public opinion. For instance, the model (accurately) estimated that a majority of U.S. adults support legal abortion. But while a number of Americans think abortion should be *legal in all cases* (23%) or *illegal in all cases* (11%), the AI poll underestimated these opinions.

AI poll underestimates strong support and strong opposition to abortion

% who say abortion should be ...



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

Source: Survey of U.S. adults and “digital twins” synthetic analysis using Claude Opus 4.6 set to low reasoning with extended profile information and expert reflection. Survey was conducted Jan. 20-26, 2026 (replicated March 9-12 and April 7-10).

“Can AI Stand In for Human Survey-Takers? Not Really”

PEW RESEARCH CENTER

AI polling results can lean on racial or partisan stereotypes

In many cases, our AI poll took beliefs or attitudes that are reasonably common among a particular subgroup and portrayed them as nearly ubiquitous. For example, our AI poll would indicate that:

- 97% of Hispanic adults are at least somewhat likely to follow the World Cup (while the actual share is 43%).
- 95% of Republicans and Republican-leaning independents have a very or somewhat favorable view of Israel (actually 58%).
- 86% of Democrats and Democratic leaners think billionaires are a bad thing for the country (actually 45%).

Read more about [how well synthetic polls replicate the diversity and distribution of public opinion](#).

The model thinks we know more than we actually do

Do you know [which right is protected](#) by the First Amendment to the U.S. Constitution? Or what the [focus of the NATO alliance](#) is? Our surveys have found that fewer than six-in-ten Americans can correctly answer these questions. But our AI poll estimated that nearly every member of the public knows the answers.

More broadly, our AI respondents didn't like to admit to uncertainty. Across all the questions on our three surveys where a "not sure" option was offered, human panelists were around four times as likely as the AI model to choose it.

Read more about [how synthetic respondents express certainty and factual knowledge](#).

The model you use can change the answers you get

We found that different AI models can paint a different picture of the public mood – even when everything else about the survey is the same. In a test comparing OpenAI's GPT-5.1 and Anthropic's Claude Opus 4.6 on a subset of questions, the GPT estimates described an American public that has more extreme opinions on a variety of topics than they actually do, while the Opus estimates described a public that is more middle-of-the-road than in reality. A reader of either result would be misled, but in different directions.

Read more about [how synthetic polling results change based on the AI model](#).

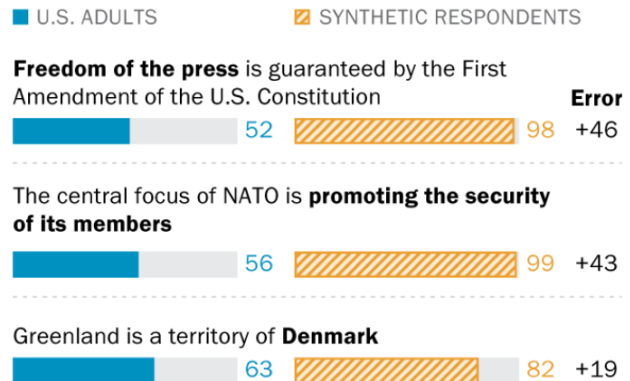
The bottom line: Survey-taking is best left to humans

Our biggest takeaway from this exercise is that **AI polling is not a replacement for rigorously surveying real humans**. It's not just that the AI results differ from human results, although that is certainly true. The bigger story is that these results differ in ways that are often unpredictable, and they are highly subject to factors like unforeseen real-world events or the choice of model used.

To be sure, this is a fast-evolving field with a great deal of academic and applied research effort behind it. There are also many other ways to use AI to improve the polling process that stop

AI respondents score much higher than humans on knowledge questions

% who select the *correct answer*



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

Source: Survey of U.S. adults and "digital twins" synthetic analysis using Claude Opus 4.6 set to low reasoning with extended profile information and expert reflection. Survey was conducted March 23-29, 2026 (replicated March 30-April 2).

"Can AI Stand In for Human Survey-Takers? Not Really"

PEW RESEARCH CENTER

short of replacing human respondents. For instance, AI can be used to categorize real responses to open-ended questions or write the code used to analyze the survey results – [the Center is using AI tools in these contexts](#) and will continue to do so. But at the end of the day, we see no substitute for rigorous, multimode, probability-based surveys that allow real members of the public to speak their minds on issues of importance.

How well synthetic samples replicate public opinion

Any evaluation of a new polling technology like AI-based surveys starts with a basic question: How well does it replicate the results of high-quality, nationally representative probability polls? Does it accurately depict the shares of Americans who hold various views? And does it work better or worse for particular demographic and behavioral subgroups?

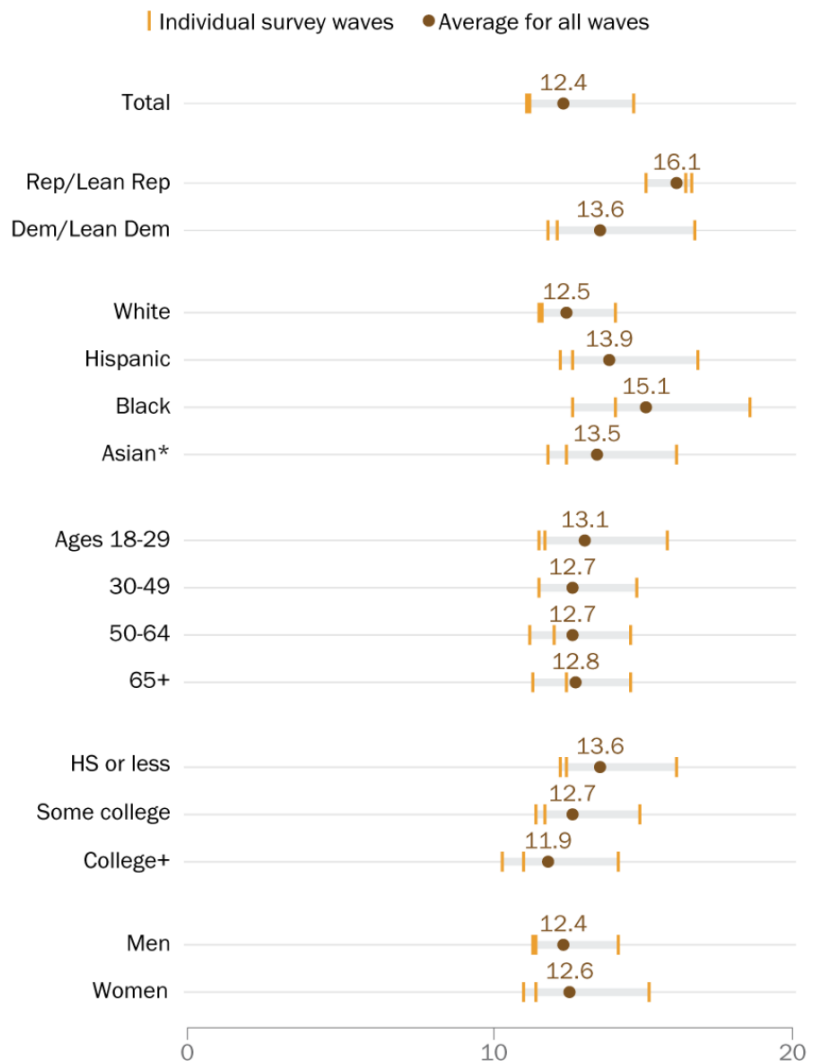
To evaluate these questions, we used AI respondents to replicate three survey waves previously administered to the [American Trends Panel](#) (ATP). We then calculated the average question-level error rate for each wave, including the error rates for several demographic subgroups.

These are the three ATP waves we replicated and some of the topics they included:

- **Wave 185** (fielded January 2026): political attitudes and opinions of the Trump administration, military action in Venezuela, annexation of Greenland, conduct of U.S. Immigration and Customs

AI-generated survey results consistently differ from human surveys

Average absolute error (in percentage points) between human and synthetic respondents on 3 replicated survey waves



* Estimates for Asian adults are representative of English speakers only.
 Source: Surveys of U.S. adults and “digital twins” synthetic analysis using Claude Opus 4.6 set to low reasoning with extended profile information and expert reflection. Surveys were conducted Jan. 20-26, 2026 (replicated March 9-12 and April 7-10), March 23-29, 2026 (replicated March 30-April 2), and April 20-26, 2026 (replicated April 27-May 1).

“Can AI Stand In for Human Survey-Takers? Not Really”

PEW RESEARCH CENTER

Enforcement officers, and attitudes toward data centers

- **Wave 190** (fielded March 2026): global politics and international relations, knowledge questions about international affairs and the U.S. Constitution
- **Wave 192** (fielded April 2026): presidential approval and political attitudes, problems facing the country, military action in Iran, and sleep habits

This analysis is part of a larger evaluation of AI-generated synthetic samples in public opinion research. Read [a summary of the main findings](#) and refer to the [methodology](#) for more details on how we conducted our synthetic poll and compared it with human polling results.

Broadly speaking, the AI survey respondents did not match the views of their human counterparts especially well. Across all three waves, the synthetic polling results differed by an average of 12 percentage points from the estimates produced by the human polls. For individual waves, absolute average error ranged from 11 to 15 percentage points.

There were also clear errors for all major subgroups in the analysis. For a range of demographic and behavior groups across all three waves, none had an average error of less than 12 points. For several groups, these errors were around 15 points or more.

The synthetic sample performed especially poorly when it came to replicating the responses of groups including:

- Republicans and Republican leaners
- Black adults
- Those who have not attended college
- Those who use the internet infrequently or not at all

Synthetic poll performance on specific topics and subgroups

In addition to the model's overall poor performance replicating public opinion for different subgroups, we also found that this method produced skewed or biased results for specific questions and subgroups.

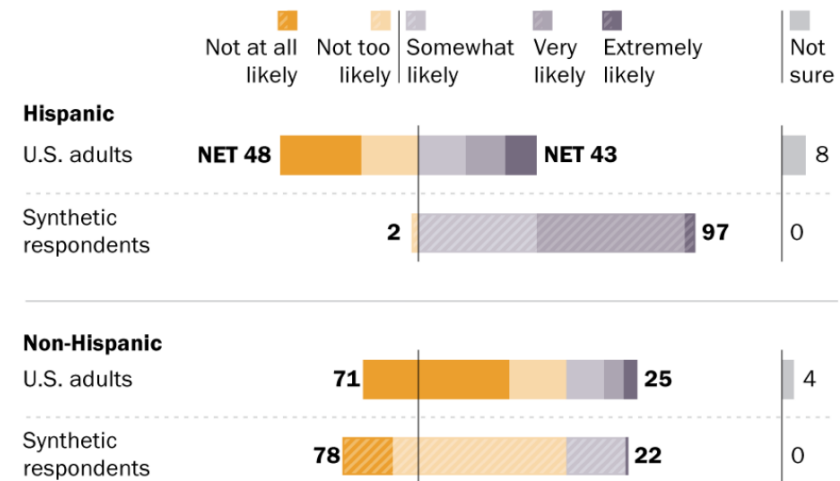
Views of Hispanic adults toward the World Cup, Spanish classes in schools

AI models are trained on how humans think and act in large part from online content. Because of this, there has long been evidence and concern that they have a tendency to make assumptions or rely on stereotypes about certain groups of people. We saw several instances of this behavior in our own experiment.

For example, in a March 2026 ATP survey, fewer than half of real Hispanic adults (42%) said they were at least somewhat likely to follow the World Cup. In contrast, nearly every simulated Hispanic respondent in our AI poll (a full 97%) said they were likely to follow the tournament.

AI polls greatly overstate Hispanic adults' interest in the World Cup

% who are ___ to follow the World Cup



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

Source: Survey of U.S. adults and "digital twins" synthetic analysis using Claude Opus 4.6 set to low reasoning with extended profile information and expert reflection. Survey was conducted March 23-29, 2026 (replicated March 30-April 2).

"Can AI Stand In for Human Survey-Takers? Not Really"

PEW RESEARCH CENTER

Likewise, the synthetic sample dramatically overestimated Hispanic opinion on the importance of teaching Spanish in schools. According to the synthetic sample, 72% of Hispanic adults believe it is extremely important for schools to offer instruction in Spanish – more than double the rate among real Hispanic adults (32%).

Partisan attitudes

In addition to racial stereotyping, there were also numerous instances of our model misstating the nature and magnitude of partisan differences on issues of the day. Most notably, our AI survey often took views that are held by many Democrats or Republicans – but by no means all of them – and made them appear close to universal.

For instance, **at least 80% of Democrats and Democratic leaners in our AI survey** say:

- The fact that some people in the United States have personal fortunes of more than \$1 billion is a bad thing for the country (while the actual share is 45%).
- They would prefer to live in an area where homes are smaller and closer to each other, but with lots of amenities nearby (actually 60%).

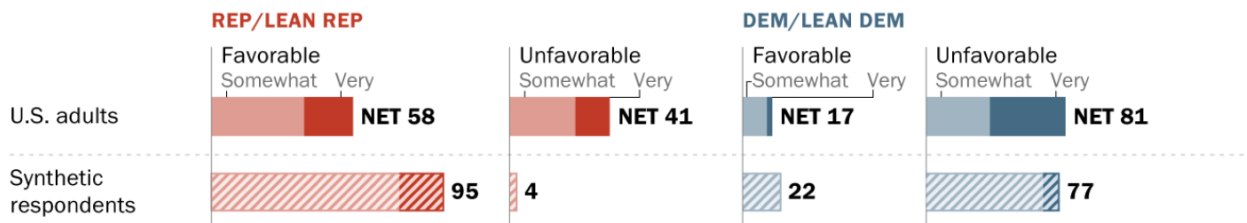
Similarly, **more than 90% of Republicans and Republican leaners in our AI survey** say:

- The police should be allowed to stop and search anyone who fits the general description of a crime suspect (actually 69%).
- They have very or somewhat favorable views of Israel (actually 58%).

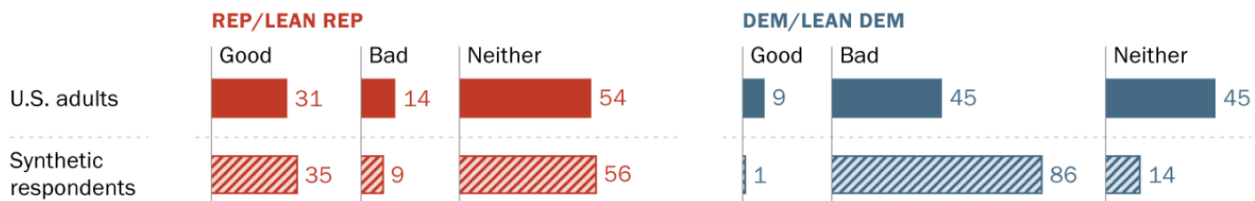
Across numerous such examples, our AI respondents were far less politically diverse than the actual population.

Synthetic surveys misrepresent Republican views of Israel and Democratic views of billionaires

% who have a(n) ___ opinion of Israel



% who say the fact that some people in the U.S. have a personal fortune of \$1 billion or more is ___ for the country



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

Source: Survey of U.S. adults and “digital twins” synthetic analysis using Claude Opus 4.6 set to low reasoning with extended profile information and expert reflection. Survey was conducted March 23-29, 2026 (replicated March 30-April 2).

“Can AI Stand In for Human Survey-Takers? Not Really”

Synthetic surveys and ‘timely and topical’ questions

Public opinion often seems slow to change, but the opinions and attitudes of the public can sometimes shift quickly and dramatically in response to current events. This is part of what’s interesting and challenging about studying public opinion. Sometimes people change their minds relatively rapidly, as happened with [former President George W. Bush’s increased approval rating](#) in the wake of 9/11. At other times, segments of the population can switch places on issues, as happened with Democrats’ and Republicans’ [perceptions of the social media platform X \(formerly Twitter\)](#) after Elon Musk purchased the company.

One appeal of AI polling is its speed. At least in theory, these polls can be fielded nearly instantly without the need to assemble a representative sample of humans who are willing to take the survey. At the same time, other ways of inferring public opinion without directly measuring it through polling (like statistical modeling) can struggle to identify changes in sentiment when they happen quickly, such as in response to breaking news.

We found that polling with AI models is no different. In our analysis comparing surveys of AI respondents with surveys of real humans, there were numerous instances in which the model failed to accurately reflect public opinion in unique cultural or economic circumstances. That is especially – but not always – true when those circumstances developed after the model’s knowledge cutoff.²

This analysis is part of a larger evaluation of AI-generated synthetic samples in public opinion research. Read a [summary of the main findings](#) and refer to the [methodology](#) for more details on how we conducted our synthetic poll and compared it with real survey results.

Using synthetic polls to reproduce ongoing trends in public opinion

Pew Research Center polled Americans on their approval of President Donald Trump using the [American Trends Panel](#) (ATP) seven times between the beginning of 2025 and May 2026, including on two surveys that we replicated with synthetic respondents.

In January 2026, 37% of U.S. adults approved of Trump’s job performance. By April, when the next replicated survey with this question was fielded, that had fallen to 34%. But our synthetic samples estimated Trump’s approval at a steady 46% for both waves.

² LLMs all have a knowledge cutoff – essentially the date on which the model stopped ingesting new data and training actually began. The model we used for most of this analysis, Claude Opus 4.6, has a knowledge cutoff in [May 2025](#).

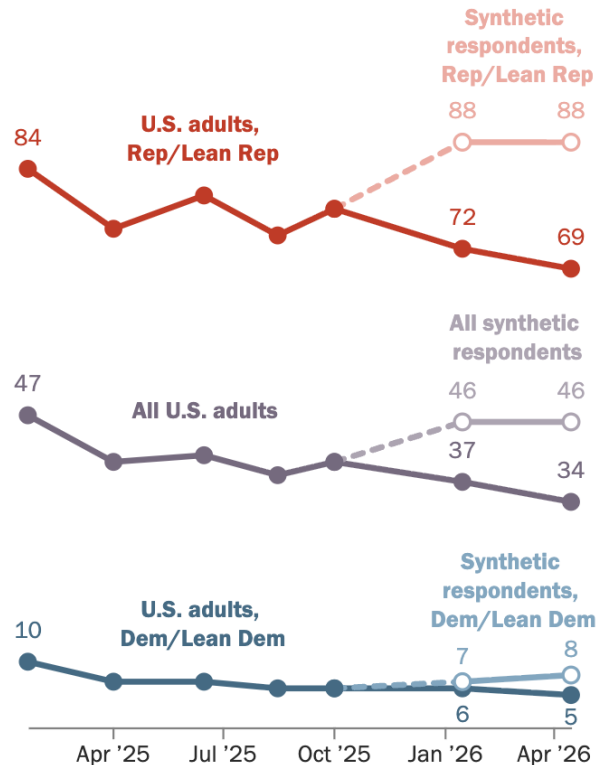
The last time Trump's approval was actually that high in our polling was at the beginning of his second term in February 2025, when 47% of U.S. adults said they approved of the job he was doing. Not only did the synthetic replication in January 2026 fail to capture the decline in approval that had occurred since Trump's return to office, but when we repeated the exercise in April, it gave us the exact same topline number.

If we look at the partisan breakdown on presidential approval over time, the story has been remarkably consistent for Democrats and Democratic leaners; the synthetic survey replicated their approval numbers quite closely. The biggest shift in attitudes in our human surveys has been a decline among Republicans and Republican leaners; our AI poll missed the mark for this group by as much as 19 percentage points.

Presidential approval is an especially long-standing trend question that the AI poll overestimated by a considerable margin in this case. But on its own, the fact that a question is trendable doesn't tell us very much about the accuracy we should expect from an AI replication. The three surveys we evaluated in this study contained more than 120 questions we had asked at least once in the past. Collectively, these trend questions had an average error rate of 12 percentage points, regardless of how much opinion had actually changed over the life of the trend. That's the same as the overall average error across all three survey waves.

Synthetic surveys greatly overstate presidential approval in 2026

% who approve of the way Donald Trump is handling his job as president



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

Source: Surveys of U.S. adults and "digital twins" synthetic analysis using Claude Opus 4.6 set to low reasoning with extended profile information and expert reflection. Surveys were conducted Jan. 20-26, 2026 (replicated March 9-12 and April 7-10), and April 20-26, 2026 (replicated April 27-May 1).

"Can AI Stand In for Human Survey-Takers? Not Really"

PEW RESEARCH CENTER

Using synthetic polls to measure public opinion on timely and topical issues

In addition to measuring how the model does at replicating some long-standing trends, we also looked at some new questions that asked about contemporaneous events the model might not have context for in its training data. These serve as examples of how synthetic polls can fail to capture how people feel about an issue when current sentiments break with past precedent in some way.

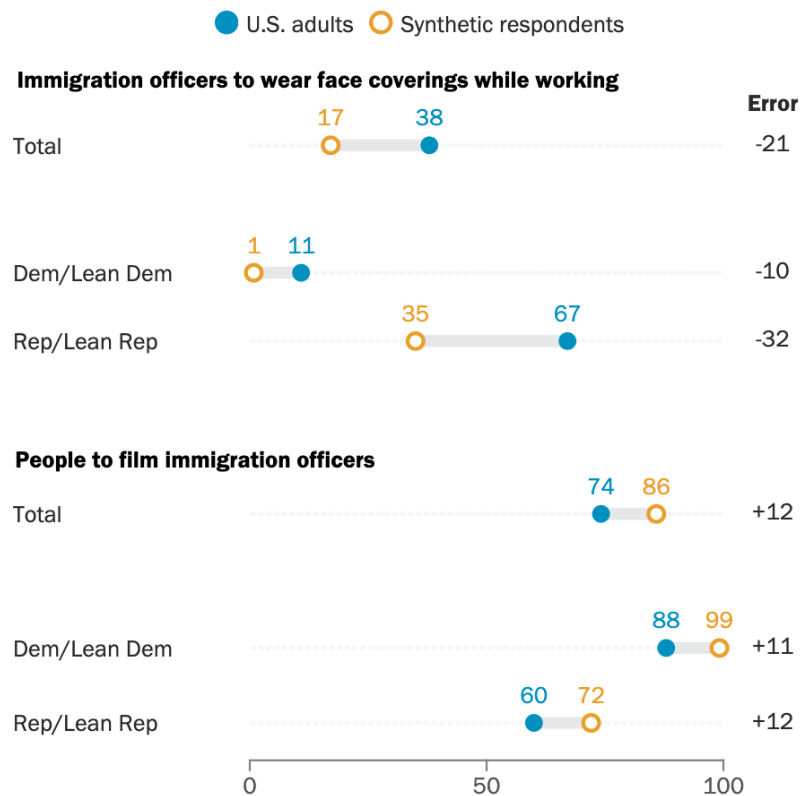
Here are some “timely and topical” issues where our model failed to capture the state of public sentiment.

Immigration enforcement

U.S. Immigration and Customs Enforcement’s actions, such as “[Operation Metro Surge](#)” in winter 2025-26, generated public controversy over a range of issues, including federal agents’ wearing of masks or other face coverings to hide their identities while working.

Synthetic surveys underestimate public views on immigration agents wearing masks while working

% who say it’s *acceptable* for ...



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

Source: Survey of U.S. adults and “digital twins” synthetic analysis using Claude Opus 4.6 set to low reasoning with extended profile information and expert reflection. Survey was conducted Jan. 20-26, 2026 (replicated March 9-12 and April 7-10).

“Can AI Stand In for Human Survey-Takers? Not Really”

PEW RESEARCH CENTER

Overall, the synthetic sample greatly underestimated the share of Americans who say it’s acceptable for immigration officers to cover their faces. Among Republicans and Republican leaners, the difference was even starker: 35% of synthetic Republicans find it acceptable, but the actual share among Republicans is nearly double that (67%).

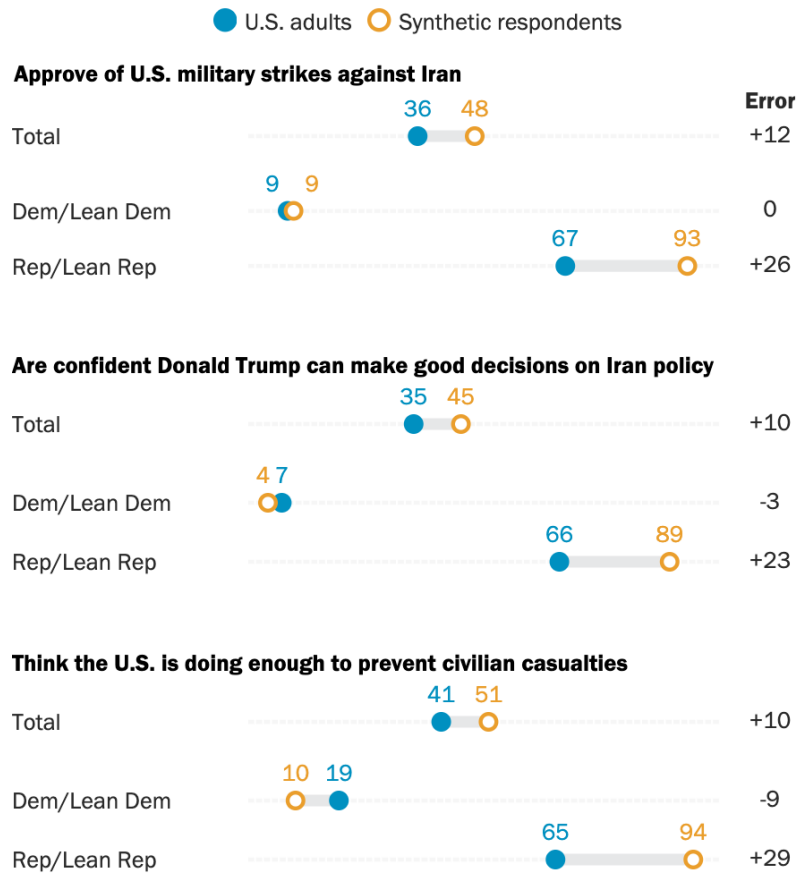
Military conflict in Iran

Our synthetic poll consistently overstated public support for the conflict between the United States, Israel and Iran that began on Feb. 28, 2026. The model's estimates were 10 to 19 percentage points higher than human respondents on topics ranging from overall approval of the strikes against Iran to whether the conflict will make the Iranian people better or worse off.

On this issue, too, the model was especially prone to overstating support among Republicans. In our synthetic panel, around nine-in-ten Republicans and Republican leaners approve of the strikes against Iran; are confident Trump can make good decisions about Iran policy; think the strikes were the right decision; and say the U.S. is doing enough to prevent civilian casualties. In reality, around two-thirds of Republicans hold each of these views. That is a solid majority, but hardly the unanimous agreement implied by the synthetic panel.

Synthetic surveys overestimate Republican support for military conflict in Iran

% who ...



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

Source: Surveys of U.S. adults and "digital twins" synthetic analysis using Claude Opus 4.6 set to low reasoning with extended profile information and expert reflection. Surveys were conducted March 23-29, 2026 (replicated March 30-April 2), and April 20-26, 2026 (replicated April 27-May 1).

"Can AI Stand In for Human Survey-Takers? Not Really"

PEW RESEARCH CENTER

The model did a better job at estimating disapproval of the conflict among Democrats and Democratic leaners. But even among this group there were notable misses. For instance, 94% of

Democrats in our AI poll think the conflict will make the Iranian people worse off, while in reality, that figure is 56%.

Data centers

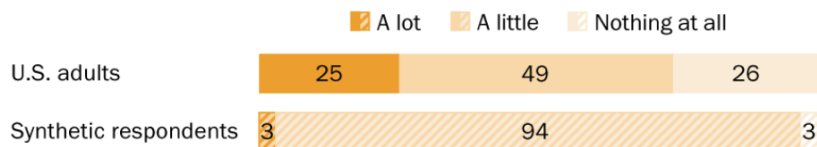
In a January 2026 survey, we asked Americans about their views on data centers – a newly salient political issue. At the time, public awareness of data centers was quite varied: 25% of U.S. adults had heard a lot about them, but a nearly identical share had heard nothing at all. Around half of Americans were in the middle, reporting that they had heard a little about data centers.

By contrast, our synthetic poll estimated that 94% of U.S. adults have heard a little about data centers, and just 3% each have heard a lot or nothing at all.

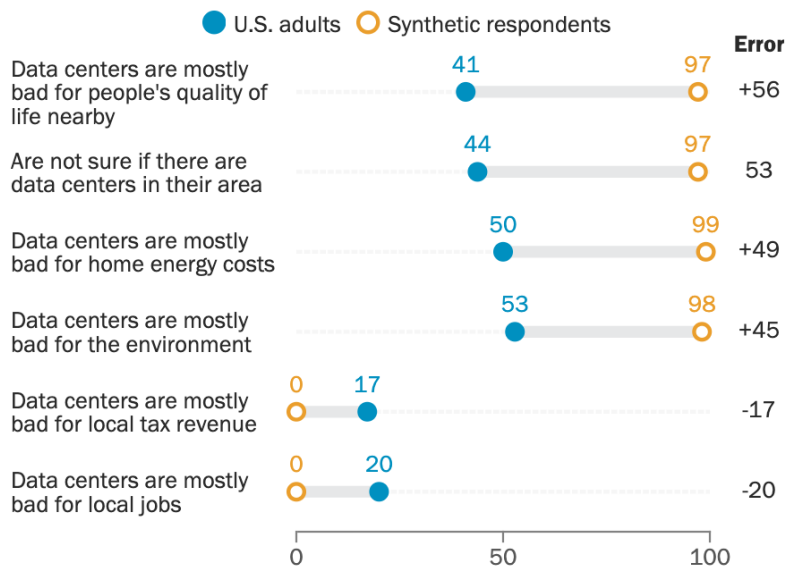
In a series of follow-up questions asked of both human and AI respondents who had heard about data centers, nearly every synthetic respondent said they weren't sure if there were any data centers built or planned in their area – more than double the 44% of human respondents who were similarly unsure.

Synthetic surveys miss the mark on awareness and views of data centers

% who have heard ___ about data centers



% who (say) ...



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

Source: Survey of U.S. adults and “digital twins” synthetic analysis using Claude Opus 4.6 set to low reasoning with extended profile information and expert reflection. Survey was conducted Jan. 20-26, 2026 (replicated March 9-12 and April 7-10).

“Can AI Stand In for Human Survey-Takers? Not Really”

PEW RESEARCH CENTER

The AI poll also missed badly on questions about the perceived impacts of data centers. At least 97% of the synthetic respondents who received these questions said data centers are mostly bad for home energy costs, the environment and the quality of life for people who live nearby.

Although many humans hold these views, the actual shares among those who have heard about data centers range from 41% to 53%.

At the same time, the AI poll greatly *understated* the share of the public who says that data centers are mostly bad for local jobs or tax revenue.

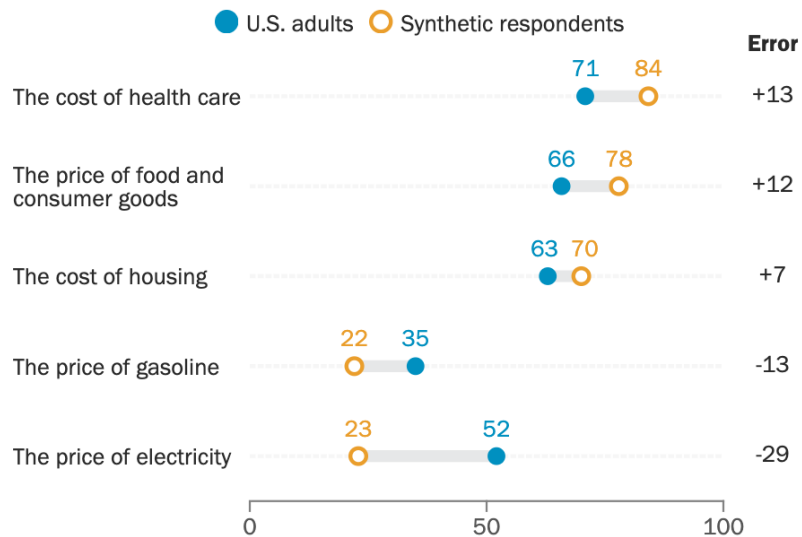
Economic concerns

In a series of questions about economic challenges facing Americans today, the synthetic sample overestimated certain concerns while underestimating others.

In particular, the model estimated that just 23% of Americans are very concerned about the price of electricity. The actual share (52%) is more than double that.

Synthetic surveys overestimate some economic concerns but underestimate others

% who are *very concerned* about ...



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

Source: Survey of U.S. adults and “digital twins” synthetic analysis using Claude Opus 4.6 set to low reasoning with extended profile information and expert reflection. Survey was conducted Jan. 20-26, 2026 (replicated March 9-12 and April 7-10).

“Can AI Stand In for Human Survey-Takers? Not Really”

PEW RESEARCH CENTER

How synthetic respondents express certainty and factual knowledge

Public opinion polling often highlights the extent to which people are unsure about an issue. Knowing and quantifying that ambiguity is essential to any understanding of how important public debates may be evolving.

Similarly, knowledge questions in polls can help provide valuable insights into the public's familiarity with a topic – for example, whether people generally [understand the digital privacy landscape](#) or know key facts about the [U.S. system of governance](#).

In contrast to humans, large language models are armed with vast quantities of training data drawn from around the internet. As such, there is reason to think that synthetic surveys will overstate the level of factual knowledge the typical American possesses.

These models also have a known tendency to [want to provide answers](#) to questions and will [confidently provide incorrect answers](#). Because of this, we also might expect them to be consistently less willing than humans to say “I’m not sure” on surveys.

In our examination of synthetic polling, we found ample evidence that AI models do exactly these things.

This analysis is part of a larger evaluation of AI-generated synthetic samples in public opinion research. Read a [summary of the main findings](#) and refer to the [methodology](#) for more details on how we conducted our synthetic poll and compared it with real survey results.

‘Not sure’ response options

We provide human survey-takers on the [American Trends Panel](#) (ATP) with an explicit “not sure” option on many questions we ask. Depending on how the questions are designed, this response may indicate uncertainty or indecision, or a lack of factual knowledge on a given topic.

In the instructions to our synthetic survey-takers, we made clear that it was realistic for them not to know about certain topics or to be incorrect at times about factual matters. But despite this prompting, our synthetic respondents chose “not sure” options far less often than real humans do.

Across all opinion questions where “not sure” was provided as an explicit option:

- Human panelists selected it 16% of the time.

- Synthetic respondents did so just 4% of the time.

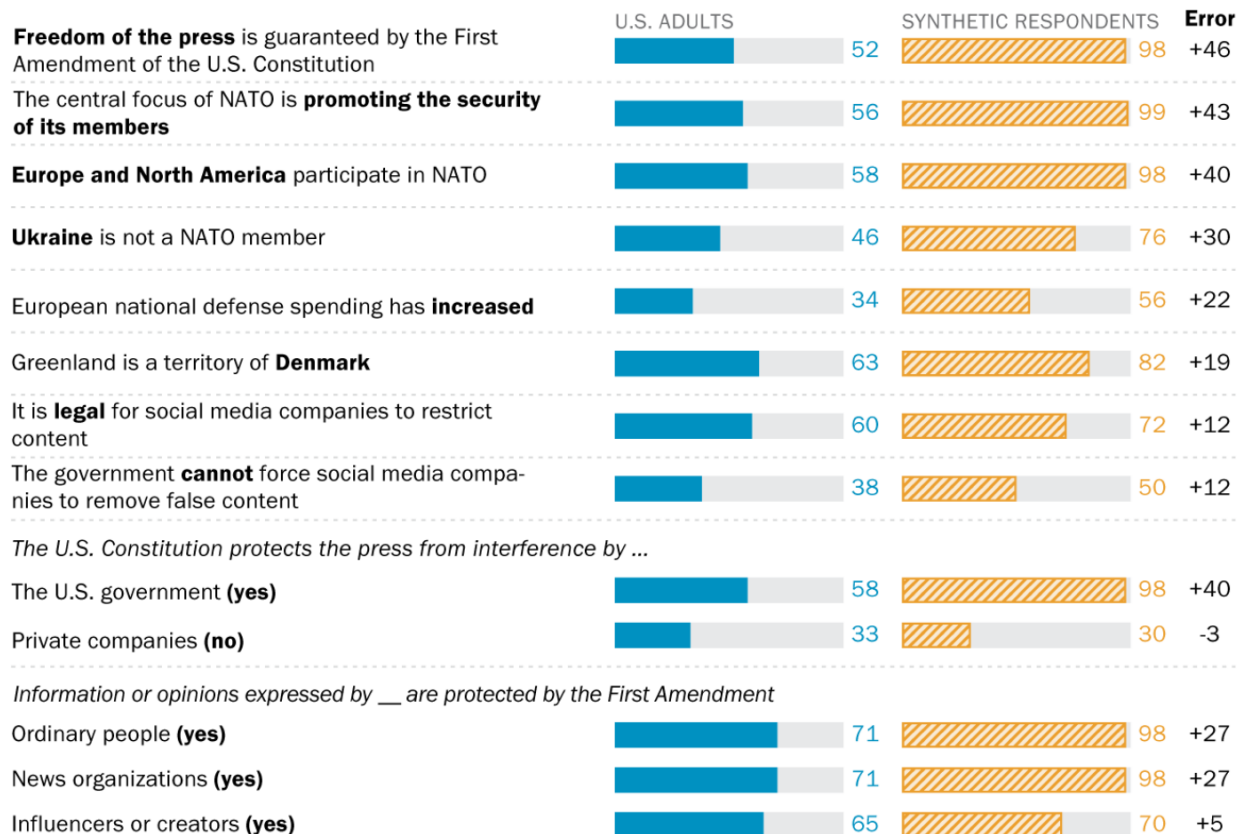
In other words, the typical human respondent says they are not sure roughly four times as often as the typical synthetic respondent does when asked to share an opinion.

Questions testing factual knowledge

The ATP survey waves we chose to replicate with a synthetic sample included 13 questions testing factual knowledge on topics including the U.S. Constitution and the NATO alliance.

Larger shares of AI panelists answer correctly on nearly all knowledge questions

% who select the *correct answer*



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

Source: Survey of U.S. adults and “digital twins” synthetic analysis using Claude Opus 4.6 set to low reasoning with extended profile information and expert reflection. Survey was conducted March 23-29, 2026 (replicated March 30-April 2).

“Can AI Stand In for Human Survey-Takers? Not Really”

Across these questions, the synthetic panel tended to be far more “knowledgeable” than our human panel. Our human respondents answered correctly around half the time, on average, and there was no question that more than three-quarters answered correctly.

By contrast, our synthetic respondents answered correctly about 80% of the time. And on six different questions, the model estimated that 98% or more of the public knew the right answer.

In cases where the model decided that its persona would *not* know the correct answer to a question, it almost always indicated it was not sure. That is, our synthetic respondents almost never selected a factually incorrect answer choice.

How well synthetic polls capture the diversity of public opinion

Traditional public opinion polls are often marked by the diverse views they show within a population. While the broad contours of public sentiment are often clear, there are very few opinions that are universally held. By the same token, most of the answer choices on a typical survey question – from the extreme to the middle-of-the-road – are selected by at least a modest number of respondents.

In our analysis of synthetic polling, AI respondents tended to eliminate a great deal of this diversity. On some questions, the model assumed most Americans' views are far more extreme than they actually are. On others, it assigned the vast majority of respondents to the middle category in a set of choices. And across questions of all types, our AI survey-takers collectively avoided certain answer choices in ways that have no parallel among their human counterparts.

Even if the model's results point in the same general direction as true public opinion, they ultimately paint a picture that is less diverse and more one-note than what we see in actual polls of human respondents.

This analysis is part of a larger evaluation of AI-generated synthetic samples in public opinion research. Read a [summary of the main findings](#) and refer to the [methodology](#) for more details on how we conducted our synthetic poll and compared it with real survey results.

How synthetic polls represent attitudes and opinions

On numerous questions in our surveys, our synthetic poll gave the appearance of a public that is far more lockstep – and often far more certain – in its attitudes than it actually is.

A question about whether the United States or China benefits more from trade between the two countries is a good example. Our synthetic poll estimated that fully 98% of Americans think China benefits more from that relationship. That's 56 percentage points higher than the real share of U.S. adults who feel this way. It also obscures notable shares who feel that the U.S. benefits more (10%), who say both countries benefit equally (24%), and who say they aren't sure (19%).

And on a series of questions about the changing influence of different countries on the global stage, the model tended to gravitate toward certain answer choices for no apparent reason. Our AI poll estimated that:

- 88% of Americans think India has been getting stronger in recent years (while the actual share is 30%).
- 77% say Canada has been getting weaker (actually 23%).
- 100% say China has been getting stronger (actually 62%).
- 92% say Iran's influence has stayed the same (actually 33%).

Broadly speaking, our model had a tendency to steer away from choices at the end of scales (like “very well” or “not at all well”) and toward middle choices like “somewhat.” On a series of questions about President Donald Trump’s personal characteristics, 0% of our AI respondents said that phrases like “honest,” “mentally sharp” or “cares about the needs of ordinary people” describe him very well. In reality, anywhere from one-in-ten to one-in-four Americans say these describe the president very well.

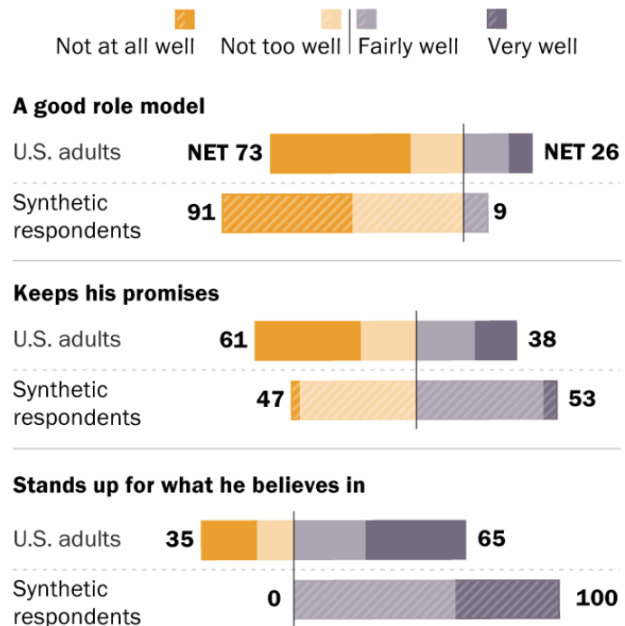
At the same time, this trend was not always consistent across questions – or even within the same set of questions. When presented with a list of 13 different issues facing the country, a substantially *larger* share of AI respondents than human panelists described several as a “very big problem.” There were only four issues for which the more extreme option was chosen by our human panelists more often.

How synthetic polls represent reported behaviors

AI-generated survey responses also fail to capture the full diversity of experiences humans report about their own lives. This can be seen clearly in a series of questions we asked about Americans’ sleep and travel habits.

Synthetic respondents differ from real Americans on descriptors of Trump

% who say each describes Donald Trump ...



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

Source: Survey of U.S. adults and “digital twins” synthetic analysis using Claude Opus 4.6 set to low reasoning with extended profile information and expert reflection. Survey was conducted April 20-26, 2026 (replicated April 27-May 1).

“Can AI Stand In for Human Survey-Takers? Not Really”

PEW RESEARCH CENTER

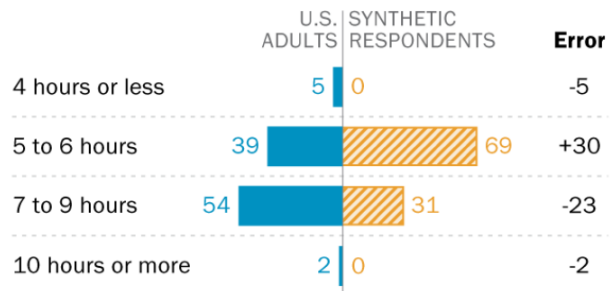
On a question about **how often people have trouble falling asleep**, fully 98% of synthetic respondents say they experience this some days or rarely. Although a majority of Americans choose one of these two options, 30% in total say this happens to them every day, most days or never. The synthetic poll also lumps the vast majority of Americans into the “some days” category on a similar question about how often they have trouble staying asleep.

When it comes to **how much sleep they get per night**, the synthetic poll overstated the share of Americans who get five to six hours of sleep per night while understating the share who get seven to nine hours of sleep. The synthetic poll also estimated that *no* Americans get very little (four hours or less) or quite a bit (10 hours or more) of sleep in a night – but in reality, nearly one-in-ten Americans fall into one of these categories.

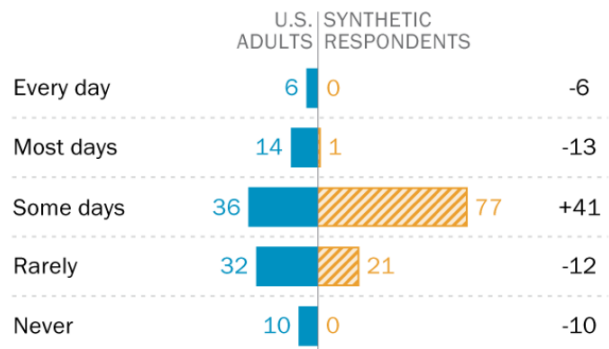
And on a question about **the number of countries people have visited outside the United States**, the synthetic poll was fairly close to the actual population on most response options. But it estimated that just 1% of Americans have visited 10 or more countries. In reality, 13% of U.S. adults say that they have done this.

On questions about sleep, AI surveys lump responses into middle categories

% who say, on average, they get ___ of sleep per night



% who say they have trouble falling asleep ...



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

Source: Survey of U.S. adults and “digital twins” synthetic analysis using Claude Opus 4.6 set to low reasoning with extended profile information and expert reflection. Survey was conducted April 20-26, 2026 (replicated April 27-May 1).

“Can AI Stand In for Human Survey-Takers? Not Really”

PEW RESEARCH CENTER

Synthetic respondents often avoid certain answer choices entirely

On a typical Pew Research Center survey, it’s rare for a single answer choice on a given question to get zero responses. This did not happen on any of the questions across the three original survey waves we evaluated for this study.

These “zero selection” response options occurred much more often in our AI survey. Fully 47% of questions in our synthetic poll had at least one answer choice that was not selected by a single synthetic respondent.

This disparity between the human and synthetic respondents is also visible at other thresholds:

- Just 1% of questions in our human poll had a response option that was selected by fewer than 1% of respondents, versus 66% of questions on the synthetic poll.
- 27% of questions in the human poll had a response option selected by fewer than 5% of respondents, versus 77% of questions on the synthetic poll.

How synthetic polling results change based on the AI model

A key feature of public opinion polling is its consistency and predictability. Although the individual participants may be different, two identical surveys fielded at the same time using the exact same methods should produce largely similar results within known boundaries.

But there is no guarantee of this when the “participants” taking the surveys are AI models. To explore how the choice of model can affect the results of the synthetic survey, we used both OpenAI’s GPT-5.1 and Anthropic’s Claude Opus 4.6 to simulate 6,700 real respondents from our [American Trends Panel](#) (ATP) and generate synthetic survey data using a questionnaire we originally fielded in late January 2026. We then compared the models’ results against each other, as well as against the results from the human poll.

Looking at the survey average absolute error, neither of the models we tested replicated the human panel results especially well. Opus performed slightly better, with an average absolute error of 11.4 percentage points (compared with 13.3 percentage points for GPT).

To be sure, newer models and future innovations in synthetic sample construction could result in smaller overall error figures. The real story here is that **each synthetic sample painted a very different picture of the American public on multiple dimensions, and neither truly reflected actual public opinion.**

This analysis is part of a larger evaluation of AI-generated synthetic samples in public opinion research. Read a [summary of the main findings](#) and refer to the [methodology](#) for more details on how we conducted our synthetic poll and compared it with real survey results.

The questions and topics on which GPT most closely mirrored actual public sentiment were almost entirely different from those on which Opus was the better performer. Still, there were [no topics or particular categories of question](#) that stood out as a strength of one model over the other.

Here are examples of ways in which the two models provided results that were different from each other – and also from our human poll.

Example 1: Public views of politics and democracy

Our January ATP survey included a set of questions about broad political attitudes, like overall satisfaction in the way things are going in the United States, which ideological “side” is losing more often, the impact of voting, and whether there are clear solutions to the problems facing

the country. These are all fairly long-standing trend questions with ample history to draw on, yet the two models produced wildly inconsistent results.

For instance, the Opus poll was almost exactly in line with actual public sentiment when it came to the share of Americans who are dissatisfied with the way things are going in the country today; the share who feel like their “side” has been losing more than winning in politics; and the share who say voting gives people like them some say in how the government runs things. Each of these opinions is held by a majority of Americans, but they are far from ubiquitous.

By contrast, our GPT poll depicts these views as quite literally universal – for instance, estimating that 100% of Americans are dissatisfied with the way things are going in the country.

On the share of Americans who agree that there are clear solutions to most big issues facing the country today, GPT more closely mirrors the views of the public, while Opus differs from true public opinion by roughly 30 percentage points.

Example 2: Abortion attitudes

Abortion is an example of an issue on which neither model produced results that accurately reflect public sentiment.

In our human poll, around six-in-ten Americans said abortion should be legal in all or most cases, while around four-in-ten said it should not be legal. Opus produced the same general split, but greatly underestimated the share of Americans who hold very strong views on abortion:

GPT exaggerates dissatisfaction with the country’s direction, while Opus downplays clarity of solutions

% who say ...

- U.S. adults
- GPT synthetic respondents
- Opus synthetic respondents

All in all, they are dissatisfied with the way things are going in the country today



There are clear solutions to most big issues facing the country today



Thinking about politics over the last few years, their side has been losing more often than winning



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

Source: Survey of U.S. adults and “digital twins” synthetic analysis using models set to low reasoning with extended profile information and expert reflection. Survey was conducted Jan. 20-26, 2026 (replicated March 2-3 with OpenAI GPT-5.1; and replicated March 9-12 and April 7-10 with Claude Opus 4.6).

“Can AI Stand In for Human Survey-Takers? Not Really”

PEW RESEARCH CENTER

- 23% of U.S. adults think abortion should be *legal in all cases*; our Opus poll estimated that share at just 13%.
- 11% think abortion should be *illegal in all cases*; Opus estimated that 0% of the public holds this view.

By contrast, GPT produced overall estimates that imply the country is about evenly split on the legality of abortion, with the view that it should be illegal slightly more common.

All told, neither synthetic poll illustrates where the country stands on this issue – and they fail to capture the full scope of public sentiment in different ways. Here, the choice of model alone has the potential to reshape the narrative around public opinion of abortion.

Example 3: Partisan attitudes

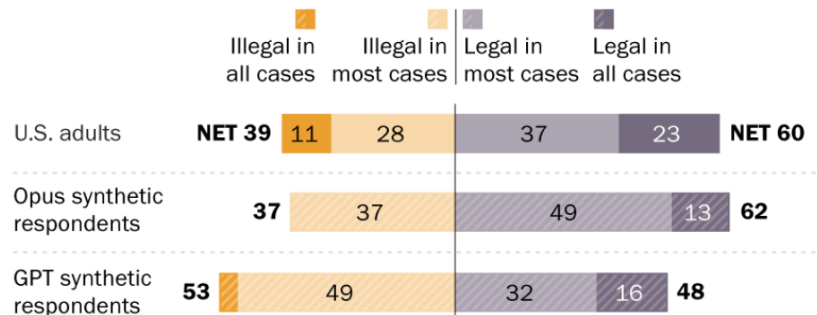
The two models also painted very different pictures of the views held by Republicans and Republican leaners. Synthetic estimates for Republicans and those who voted for President Donald Trump in 2024 make for particularly striking examples of how using a different model can lead to very different conclusions about public opinion within a certain group.

Our January 2026 ATP survey found that Trump’s overall approval rating among his 2024 voters was at 84%, with 65% approving very strongly and 19% approving not so strongly. Both of our synthetic polls overestimated his approval rating – Opus by 15 points and GPT by 13 points. But beyond the topline figures, each synthetic poll presented a different conclusion about this voter base:

- Trump voters in the GPT poll consisted almost entirely of those who approve very strongly.
- Trump voters in the Opus poll were evenly split between very strong and not so strong approval.

AI survey responses from 2 different models disagree on abortion legalization

% who say abortion should be ...



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

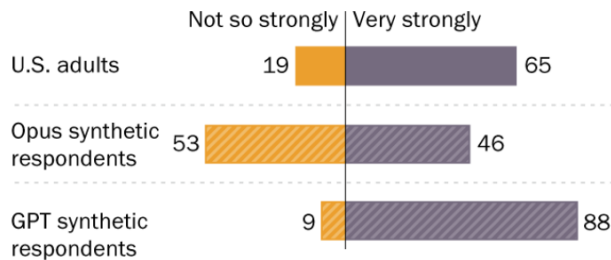
Source: Survey of U.S. adults and “digital twins” synthetic analysis using models set to low reasoning with extended profile information and expert reflection. Survey was conducted Jan. 20-26, 2026 (replicated March 2-3 with OpenAI GPT-5.1; and replicated March 9-12 and April 7-10 with Claude Opus 4.6).

“Can AI Stand In for Human Survey-Takers? Not Really”

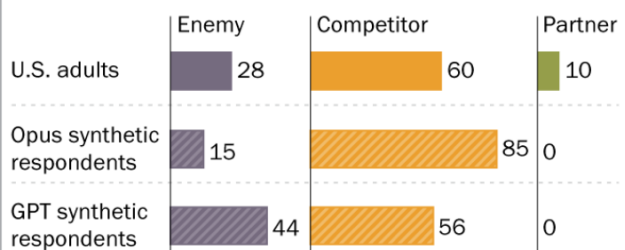
PEW RESEARCH CENTER

Model choice in creating AI-generated survey data leads to contradictory stories of attitudes among Republicans

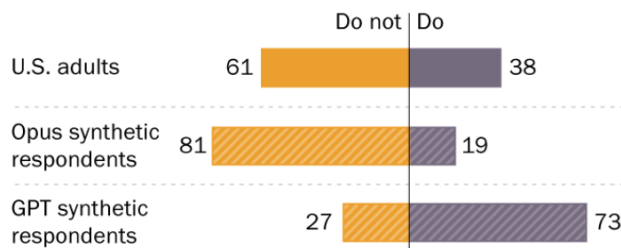
% of **2024 Trump voters** who ___ approve of the way Donald Trump is handling his job as president



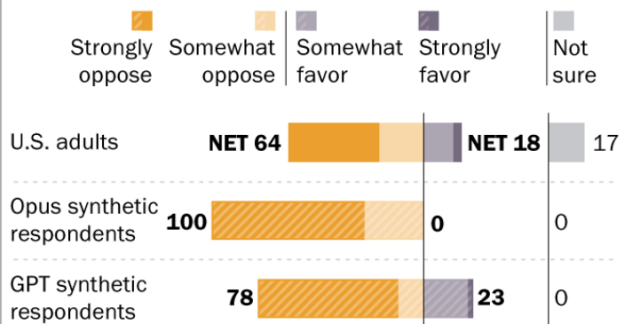
% of **Rep/Lean Rep** who characterize China as a(n) ___ of the U.S.



% of **Rep/Lean Rep** who say Republicans in Congress ___ have an obligation to support Trump's policies and programs



% of **Rep/Lean Rep** who ___ the U.S. sending ground troops to Venezuela



Note: Figures for U.S. adults may differ slightly from those previously published by the Center because they are based on a subset of respondents to the original survey.

Source: Survey of U.S. adults and "digital twins" synthetic analysis using models set to low reasoning with extended profile information and expert reflection. Survey was conducted Jan. 20-26, 2026 (replicated March 2-3 with OpenAI GPT-5.1; and replicated March 9-12 and April 7-10 with Claude Opus 4.6).

"Can AI Stand In for Human Survey-Takers? Not Really"

PEW RESEARCH CENTER

Across other questions in the survey, GPT estimates tended to support a view of Republicans as fervent Trump supporters with highly polarized partisan views. By contrast, the Opus poll would indicate that far more Republicans hold mixed or moderate views. Some questions where the two models present opposing views of Republican attitudes include:

- Whether Republicans in Congress have an obligation to support Trump's policies and programs because he is a Republican president
- Whether or not Trump should work with Democratic congressional leaders
- Whether China is an enemy, competitor or partner to the United States
- Whether or not the U.S. should send ground troops to Venezuela

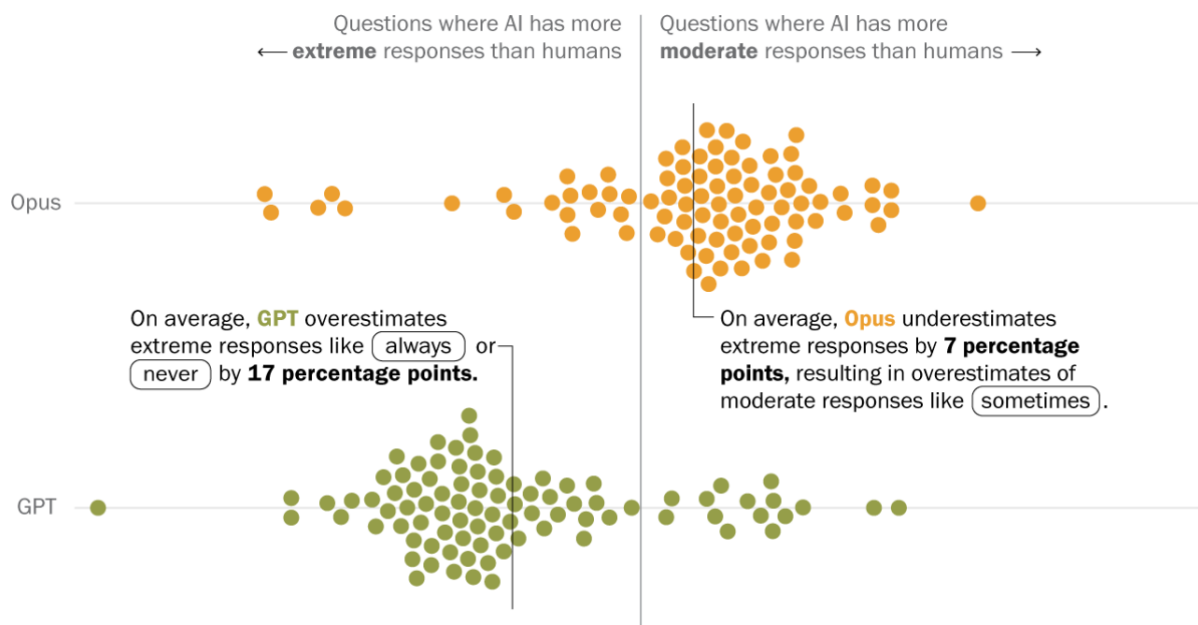
They also present very different estimates of public sentiment toward prominent political figures. In our human poll, 74% of Republicans and Republican-leaning independents expressed a favorable view of Health and Human Services Secretary Robert F. Kennedy Jr. Opus estimated his favorability among Republicans at 85%, while the GPT poll estimated their opinion to be 63% *unfavorable*.

Example 4: Magnitude of opinion and “extreme” answer options

The example of abortion notwithstanding, the GPT and Opus polls often agreed that the public generally leans in one direction or another on various issues. But often, the two models disagreed on the intensity of those attitudes.

For instance, 68% of U.S. adults in our ATP survey said they are concerned about the price of gasoline, split evenly between those who are very or only somewhat concerned. GPT and Opus both (incorrectly) estimated that nearly every American is concerned about this issue. But the

GPT overestimates extreme response options like ‘always’ or ‘never,’ while Opus underestimates them



Note: Average percentage point error is calculated only for questions with ordered response categories (89 out of 119 questions tested). Extreme responses include options at the far ends of scales (such as “extremely” or “not at all”). Moderate responses include middle options (such as “somewhat,” “about right” or “neither”).

Source: Survey of U.S. adults and “digital twins” synthetic analysis using models set to low reasoning with extended profile information and expert reflection. Survey was conducted Jan. 20-26, 2026 (replicated March 2-3 with OpenAI GPT-5.1; and replicated March 9-12 and April 7-10 with Claude Opus 4.6).

“Can AI Stand In for Human Survey-Takers? Not Really”

GPT sample lumps most of the public into the “very concerned” category, while the Opus sample leans heavily toward “somewhat concerned.”

As it turns out, this kind of pattern appeared consistently throughout the survey. GPT tended to favor response options at the far ends of scales, such as “extremely” or “not at all.” By contrast, Opus tended to favor “middle” options like “somewhat,” “about right” or “neither.” This was broadly true regardless of what the scale was or what the question was about.

For all questions in the survey with response options arranged in a clear order:

- Human respondents chose a “middle” option 45% of the time, on average.
- GPT respondents did so 31% of the time.
- Opus respondents did so 56% of the time.

Put simply: GPT tended to depict Americans as having more extreme opinions than they actually do, while Opus tended to depict them as being more middle-of-the-road than they are in reality.

It’s important to note that changes to how synthetic survey data is generated and future updates to the GPT and Opus models themselves could lead to different patterns. However, it is clear that the choice of model alone is enough to yield large, systematic differences between synthetic samples that were otherwise built in exactly the same way.³

When compared with the data from our real panelists, neither synthetic poll was able to provide a reasonably similar snapshot of the views of the American public. However, given that Claude Opus 4.6 had better performance *on average*, this is the model we selected to generate additional synthetic samples.

³ Even models within the same family did not necessarily behave similarly to one another. GPT-5 nano (a “lighter weight” model within the GPT-5 family) produced very different estimates from both GPT-5.1 and Opus 4.6.

Acknowledgments

This report is a collaborative effort based on the input and analysis of the following individuals:

Primary researchers

Athena Chapekis, *Computational Social Scientist*
Arnold Lau, *Former Research Methodologist*
Andrew Mercer, *Principal Research Methodologist*
Samuel Bestvater, *Associate Director, Data Labs*
Sono Shah, *Former Associate Director, Data Labs*

Research team

Aaron Smith, *Director, Data Labs*
Anna Brown, *Senior Survey Manager*
Skyler Seets, *Computational Social Science Assistant*
Anna Lieb, *Computational Social Science Analyst*
Kaitlyn Radde, *Computational Social Science Assistant*

Editorial and graphic design

Carson TerBush, *Associate Information Graphics Designer*
Anna Jackson, *Editorial Specialist*

Communications and web publishing

Sogand Afkari, *New Media Manager*
DeVonte Smith, *Communications Associate*
Nida Asheer, *Senior Communications Manager*
Janakee Chavda, *Associate Digital Producer*
Joy Li, *Assistant Digital Producer*
Alexander Basescu, *Content Strategy Analyst*

In addition, the project benefited greatly from the guidance of Rachel Drian, Courtney Kennedy, Ashley Amaya, Scott Keeter and Claudia Deane.

Methodology

This report compares synthetic polling results with benchmark surveys administered to human respondents. To develop the “silicon sample,” we first built a custom architecture for administering synthetic surveys. Then, we selected three survey waves from Pew Research Center’s [American Trends Panel](#) (ATP) – Wave 185, Wave 190 and Wave 192 – and replicated them on AI-generated “respondents” with the characteristics of the people who completed the original surveys.

Refer to the [comparison topline](#) for the full distribution of human and AI survey responses.

The American Trends Panel survey methodology

This report is based on three surveys of Pew Research Center’s American Trends Panel (ATP), a nationally representative panel of randomly selected U.S. adults:

- **Wave 185**, conducted Jan. 20-25, 2026 (N=8,512). Data in this report comes from a subsample of 6,700 panelists due to the cost of running Opus 4.6. The subsample was drawn with probability proportional to their original Wave 185 weight, with all non-Hispanic Asian adults selected with certainty. The weights were then adjusted to account for further subsampling and then calibrated to align with the same population benchmarks as used in the full wave. All panelists who responded to Wave 185 also responded to the 2025 political typology survey.
- **Wave 190**, conducted March 23-30, 2026 (N=3,507). Data in this report comes from 3,398 panelists who responded to both Wave 190 and the 2025 political typology survey.
- **Wave 192**, conducted April 20-26, 2026 (N=5,103). Data in this report comes from 4,981 panelists who responded to both Wave 192 and the 2025 political typology survey.

Only panelists who also completed the 2025 political typology survey were included in this analysis. Further methodological details about each survey wave can be found at the links above.

To learn more about the ATP, read “[About the American Trends Panel](#).”

Silicon sample development and analysis

We constructed our AI poll using a “digital twin” approach in which each human respondent has a corresponding AI doppelganger who stands in for them in the synthetic survey.

Constructing digital twins for each respondent lets us reproduce the same kind of flexible subgroup analysis we would perform with a real survey. We can look at estimates for all respondents, and then separately look at estimates by partisanship, age, race or any other characteristic. Because every synthetic respondent has its own complete set of answers, we know these breakouts are internally consistent and will average out to match the overall figures, the same way real survey data does.

This approach also allows us to apply survey weights to the synthetic sample, which means we can target population benchmarks and adjust for coverage and nonresponse in the same way we would do for a panel of real people.

Digital twins also can introduce more variance into the data than aggregate-estimate approaches. Because each synthetic respondent answers independently, the method avoids some of the averaging that can make LLM-generated estimates appear somewhat flat or generic. It instead allows for “person-to-person” variation. Two synthetic respondents with very similar demographic and political profiles have the freedom to answer a given question differently, just as two real people might.

Finally, and perhaps most importantly for this research, the digital twin approach allows us to see *how* and *where* a model succeeds or struggles. For example, we can determine if it is consistently misrepresenting certain subgroups or failing to replicate certain types of questions, rather than simply report that a given estimate was off by some (potentially arbitrary) number of percentage points.

How we selected survey waves to benchmark

To account for the way real-time events can change people’s opinions, we wanted to replicate the survey waves as close to their original ATP field dates as possible. However, due to time and budget constraints, we only wanted to run replications for panelists who completed the survey. For this reason, survey replications were run starting the day after the real survey waves returned from the field.

There was one exception: Wave 185 was used for model configuration testing, so its replications were run for a number of weeks after the survey had fielded.

Wave selection criteria

Waves were selected for replication based on:

- **Field dates.** Due to the timeline of the project, we needed waves that fielded in the first half of 2026. For model testing, we needed a wave that had fielded as of February 2026. For additional replications, we needed waves that fielded between March and May 2026.
- **Time constraints.** To allow replications to take place without overlap, and with sufficient time to prepare the materials needed for replication, there needed to be time between waves.
- **Monetary constraints.** Full iterative replications are very token-intensive, meaning costs quickly add up. Based on the budget for this project, we had a limit to the number of questions we could replicate and the number of panelists we could replicate those questions for. This helped us narrow down which combination of waves would be replicated.
- **Survey content.** ATP waves can focus on any number of topics. We wanted to make sure there was topical diversity in the waves we replicated (i.e., not all U.S. politics). Still, the majority of questions tested ended up being on political topics.

Based on the available waves in our project window, the topics and number of questions on each, and the number of panelists set to take the surveys, we narrowed our replications down to the three waves listed above.

The LLM survey-taking pipeline

To evaluate the performance of the synthetic samples, we wanted to replicate the survey-taking process as closely as possible. To that end, the LLM “respondent” answered the entire survey one question at a time, as the corresponding human panelist might have. After each question was answered, it was added back into the prompt to allow the LLM to remember the previous questions and how it answered them. By doing this, we hoped to promote internal consistency.

Specifically, we were *not* interested in having the LLM generate aggregate point estimates for the entire panel or certain subgroups. When synthetic survey data is generated in this way, it has no guarantee of internal consistency and does not allow for additional demographic breaks and comparisons (the way we would analyze the results of a real survey).

While research has found that [aggregate estimates are generally more accurate](#), this is due to the *averaging effect* of the LLM. We were less interested in knowing exactly how close the synthetic averages were to the real panel averages. Instead, we wanted to understand the specific ways in which trends and patterns in the synthetic data failed or succeeded in replicating human opinion.

Randomization and split forms

Each digital twin encountered the same survey programming as its human counterpart. If an ATP panelist was on Form 1 of a split-form survey, their digital twin was too. If they took the survey in Spanish, so did their digital twin.

We also programmed the same question order and response order randomization for each pair due to an interest in response-order biases – we know they are present in real survey data and wanted to know if they would be similarly present in the synthetic sample.

Profile information

To condition the LLM on each persona it was meant to emulate, we used a series of demographic variables from the ATP. In addition to basic demographic information such as age, education, race, gender and political affiliation, we included things like:

- Marital status and presence of children
- Household income
- State and county of residence
- Volunteerism
- Moved houses in the past year
- Military veteran status
- Frequency of internet use
- Religious attendance
- Disability
- Labor union membership
- 2022 and 2024 turnout and vote choice
- Political ideology (liberal versus conservative)

These variables were present in all the configurations we tested. Refer to [Appendix C](#) for the full list of standard profile variables.

The extended profile

Along with the characteristics above, we provided our AI respondents with 71 additional variables from our [2025 political typology survey](#).

These included broader dispositions such as whether or not most people can be trusted, how important it is to believe in God to be good and moral, and how often panelists felt unfairly judged by others because of the language they use to express themselves.

It also included more specific opinions about concrete issues such as how comfortable panelists felt with transgender athletes competing on sports teams that don't match their sex at birth,

whether the increased use of AI in daily life is generally good or bad for society, and whether or not reductions in Social Security benefits for future retirees should be considered.

Expert reflection

An “expert reflection” is a [short, model-generated synthesis](#) of a synthetic respondent’s profile, written from the perspective of a specified social-science expert. The goal in producing these reflections is to surface implications or facts about the persona that aren’t directly stated in its profile data.

Prompting an LLM with a respondent’s raw, self-reported data alone can cause the model to overlook higher-level, less obvious insights embedded in that data. To account for this, before running a synthetic survey, we prompted GPT-5.1 to adopt the persona of a domain expert and generate a set of observations about the respondent that draw out patterns that person didn’t explicitly state themselves. These reflections were generated once per respondent and appended to the prompt as part of their profile.

Some examples of the expert reflections generated:

- “Young, college-educated, high-internet urban resident in [LOCATION]: as a [AGE]-year-old Gen Z with a BA and ‘almost constantly’ online, this person is likely exposed to a wide range of digital and social networks and news sources. That exposure plus urban life in a [REGION] metro tends to produce pragmatic, information-driven attitudes and greater familiarity with diverse cultural viewpoints even if they self-label as [IDEOLOGY].”
- “Upper-income early professional who rents and lives alone: earning in the [SALARY] range at a young age suggests upward economic mobility and a stake in pro-growth, pro-market policies (taxation, regulatory environment, career flexibility). Being a renter and single also makes housing affordability, urban amenities, and career opportunities more salient than family-oriented policies.”
- “[POLITICAL AFFILIATION] who leans [PARTY] but calls themselves [IDEOLOGY] and has split tickets: the mix of [POLITICAL AFFILIATION] ID, a [PARTY] lean, self-described [IDEOLOGY], and evidence of ticket-splitting ([PARTY] congressional vote in 2022; [POLITICAL CANDIDATE] in 2024) points to a candidate-oriented, heterodox voter. They are likely skeptical of party orthodoxy, prioritize specific issues or personalities over party labels, and are open to third-party or protest votes.”
- “Cultural Catholicism with low religiosity: identifying as Roman Catholic but seldom attending services and praying, and not evangelical, suggests religion is a cultural reference point rather than a daily organizing force. This profile implies some residual influence of Catholic social norms on moral views, but lower likelihood of supporting policy positions driven by active religious commitment.”

We tested generating these reflections from the perspectives of a political scientist, a behavioral economist and a sociologist but did not notice substantial differences in their results. Ultimately, we used the “expert political scientist” reflections for this project.

Prompting

To generate responses, the model was given a system prompt establishing its task: to role-play as a specific survey respondent, using the profile information described above to inform its answers. The prompt instructed the model to answer as that person would, taking into account any questions and answers the synthetic respondent had already provided earlier in the same survey so that its answers remained consistent with its own prior responses as the questionnaire progressed.

The prompt also included several instructions meant to push the model toward more realistic, humanlike survey-taking behavior. It reminded the model that it was playing a role, meaning it shouldn’t answer with more knowledge, precision or consistency than a real person would have, and that it was acceptable for its “knowledge” to be incomplete or mistaken in ways that fit the respondent’s profile. It permitted the model to decline to answer a question when that would be realistic (for instance, out of disinterest or if a question topic felt too sensitive).

Finally, the prompt asked the model to take into account how real survey respondents are known to behave (including common cognitive biases such as question-wording effects, response-order effects and context effects) and to allow its answers to be shaped by these same biases where appropriate, rather than answering as an idealized, bias-free respondent would.

Refer to [Appendix A](#) for the full prompt text.

Determining the best configuration

We tested three different dimensions along which the model configuration could differ: reasoning level, the amount of profile information used to create the “digital twin” and the choice of model itself. Due to time and budget constraints, we could not test every combination of these dimensions. Instead, we opted for a bracketed (tournament-style) approach in which each dimension was tested one by one, with the winner of each round advancing to the next.

Initial testing (until model evaluation) was done using GPT-5.1, which was the most recently released OpenAI flagship model available at the time. We used GPT for testing, as it is generally considered an industry “baseline” model.

Reasoning level

We started by comparing [low, medium and high model reasoning levels](#). Increasing the amount of reasoning greatly [slows down](#) the LLM's survey-taking speed, so this choice would have the biggest impact on our timeline. Each configuration was built on a simple, random subsample of 237 respondents from Wave 185 and the standard profile variables.⁴

The configurations were compared using simple match rates (whether the category selected by the LLM matched the category selected by the ATP panelist) in two ways:

- **Person-level match rates.** For each person, the number of questions correctly matched by the LLM over the total number of questions asked.
- **Question-level match rates.** For each question, the number of people whose answers were correctly matched by the LLM over the total number of people asked.

These match rates were not expected to be anywhere close to 100%, so they were solely compared with one another.

The average person-level match rate was approximately 55% under low, medium and high reasoning levels. This match rate varied by education, age, race, gender and party identification. However, person-level match rates across reasoning levels did not differ by more than 1 percentage point within any of these subgroups.

Similarly, the average question-level match rate was approximately 54%-55% across all three reasoning levels. There was also no configuration that performed noticeably better or worse at the question level: High reasoning achieved the best question-level match rate for 43 questions, medium reasoning for 39 and low reasoning for the remaining 37.

There was no evidence to justify the additional time and financial costs of higher reasoning levels. Thus, we selected low reasoning for all subsequent replications.

Conditioning information

Using GPT-5.1 with low reasoning, we then compared a configuration with only standard profile information to one with the additional variables on political attitudes and values (i.e., the

⁴ We initially sampled N=250 respondents before Wave 185 was fielded, but after filtering to only those who completed the survey (for whom we constructed survey weights) we were left with 237 respondents in each group.

extended profile), as well as to an extended profile that incorporated “expert reflection.” Each configuration was run on all respondents to Wave 185 (N=8,512).⁵

Since here we replicated a full survey, we used a more flexible comparison metric than match rate. We calculated the average absolute percentage point error between the overall weighted estimate from each configuration and the ATP data. Specifically, we first average the absolute errors across each response option for each question, then average those averages to get a single number – the **survey average absolute error** – that summarized how well each synthetic sample performed. The further away from 0, the more inaccurate the synthetic estimate.

Including only standard profile information resulted in a survey average absolute error of 16 percentage points. The addition of extended profile variables made a marked improvement, shrinking the survey average absolute error to 13.7 points. Adding the expert reflection further improved it to 13.1 points.

Because these are averaged across more than 100 questions, even small differences are indicative of real gains in performance. Therefore, we selected the configuration with the extended profile variables and an expert reflection step.

Model selection

Finally, keeping reasoning level and the profile information constant, we tested three different off-the-shelf closed models: GPT-5.1 and GPT-5 nano from OpenAI, and Claude Opus 4.6 from Anthropic.

As Opus 4.6 was much more expensive to run than OpenAI’s models, we subsampled 6,700 panelists who completed Wave 185 rather than using the entire sample. This subsample was drawn in two replicates. The first replicate (N=3,200) was made up of all non-Hispanic Asian panelists and a random one-third of the remainder.⁶ The second replicate (N=3,500) was sampled from cases not in the first replicate with probability proportional to their final Wave 185 survey weight. The survey weight for the resulting subsample of 6,700 panelists was then adjusted for each person’s probability of being subsampled.

While the configurations using GPT-5.1 and GPT-5 nano were run using all 8,512 panelists who completed Wave 185, all comparisons between them and Opus 4.6 are filtered only to the N=6,700 subsample.

⁵ Only panelists who had previously completed our 2025 political typology survey were invited to Wave 185, so we could use the extended profile variables for every respondent.

⁶ Non-Hispanic Asian panelists were oversampled to allow for sufficient sample size in calculating average absolute error for demographic subgroups. All other subgroups had sufficient sample size in the one-third random sample.

GPT-5.1

This was our baseline model for testing. [GPT-5.1](#) is a mid-priced option – not the cheapest available but far less expensive than top-tier, flagship models. At the time of this analysis, it was OpenAI’s most recently released model. As of 2026, GPT is generally considered industry standard for the kind of LLM task involved in this research.

GPT-5.1 with the extended profile variables and expert reflection step achieved a survey average absolute error of 13.3 percentage points.⁷

GPT-5 nano

[GPT-5 nano](#) is a comparatively cheap, lightweight model and runs faster than a model like GPT-5.1. However, this speed and low cost come with a real tradeoff in quality: Independent benchmarks put [GPT-5 nano’s intelligence index score](#) on the lower end across tested models. This gap showed up clearly in our results – GPT-5 nano performed noticeably worse than our baseline, confirming that model version and size matter a great deal for this kind of task.

GPT-5 nano achieved a survey average absolute error of 17.4 percentage points, which would be less accurate than GPT-5.1 even with only the standard profile variables.

Opus 4.6

We also wanted to test a state-of-the-art reasoning model outside the OpenAI model family. [Claude Opus 4.6](#) is priced roughly five times higher than GPT-5.1, but independent benchmarks scored it [among the highest intelligence models](#) as of February 2026. In our tests, Opus 4.6 performed consistently better than GPT-5.1 across the board.

Opus 4.6 achieved a survey average absolute error of 11.4 percentage points, thus outperforming both OpenAI models tested under the same conditions.

Based on the combination of parameters with the best performance, our final configuration used Opus 4.6 with low reasoning, extended profile information and expert reflections.

⁷ This is the same configuration used when comparing different amounts of profile information, but the survey average absolute error metric differs slightly due to the subsampling.

LLM survey wave replications

ATP wave	Topic	Replication dates	Number of respondents replicated	Number of questions	Model	Reasoning	Conditioning information	Average absolute survey error (pct. pts.)
185	Politics	Feb. 3, Feb 2-6	8,512	119	GPT-5.1	Low	Standard profile	16.0
		Feb. 3, Feb. 9	237	119	GPT-5.1	Medium	Standard profile	N/A
		Feb. 3, Feb. 9	237	119	GPT-5.1	High	Standard profile	N/A
		Feb. 10-12	8,512	119	GPT-5.1	Low	Extended profile	13.7
		March 2-3	8,512*	119	GPT-5.1	Low	Extended profile, expert reflection	13.1
		March 12-13	8,512**	119	GPT-5 nano	Low	Extended profile, expert reflection	17.4
		March 9-12	3,200	119	Opus 4.6	Low	Extended profile, expert reflection	11.4
		April 7-10	3,500					
190	Global attitudes	March 30-April 2	3,398	105	Opus 4.6	Low	Extended profile, expert reflection	14.7
192	Politics	April 27-May 1	4,981	67	Opus 4.6	Low	Extended profile, expert reflection	11.2

* Wave 185 GPT-5.1 low reasoning with extended profile and expert reflection was evaluated against other GPT-5.1 conditioning information configurations with the full N=8,512 sample. When evaluated against GPT-5 nano and Opus 4.6, GPT-5.1 was evaluated based on the N=6,700 subsample from the Opus 4.6 replication.

** Wave 185 GPT-5 nano was replicated for the full N=8,512 sample. However, when evaluated against GPT-5.1 and Opus 4.6, GPT-5 nano was evaluated based on the N=6,700 subsample from the Opus 4.6 replication.

"Can AI Stand In for Human Survey-Takers? Not Really"

Appendix A: Model prompt text

You are a respondent to a public opinion survey in the United States of America. Your task is to answer survey questions.

You are playing the role of a person with the following profile:

{PROFILE_JSON}

You are already in the middle of answering the survey. Here are the questions you have previously seen and the answers you have previously provided, along with other text displayed to you.

{RESPONSE_JSON}

You should answer the subsequent survey question in a manner consistent with your role and with your prior answers, and take into account any context established by prior questions.

You are viewing the survey questions via a website on a mobile device or personal computer. There is no interviewer present.

Since you are roleplaying, you do not have an encyclopedic memory, the ability to quickly conduct complex mathematics, or any other abilities that a human would not realistically have. Your knowledge of the world should correspond to your role. It is realistic for your role to not know about some topics, or for their knowledge about some topics to be mistaken. Beyond the constraints of your role, you are free to adopt any suitable personality. Consider that multiple people with the same profile may nevertheless answer questions differently.

There are several types of questions you might be asked to respond to. Some questions will ask you to provide one of several closed-ended responses. Respond according to your role. Other questions will ask you to provide an open-ended text response. Write concisely and in a writing style consistent with your role, including grammar, spelling and punctuation, as well as choice of words. You may occasionally choose to provide no response to any question if it is realistic for your role to decline to answer for various reasons, such as lack of knowledge, lack of interest, or the question being too sensitive.

When deciding how to respond, consider what the existing research literature on survey methodology says about how human respondents answer web-administered surveys, including various cognitive biases such as question wording effects, response order effects and context effects. Allow for the possibility that your role's answer may be affected by cognitive bias.

Answer questions in the same language that they are posed in. For example, if the question is in Spanish, provide your answer in Spanish.

Here is the next question:

{NEXT_QUESTION_JSON}

Appendix B: Additional tables

Different models get different questions correct

Questions where each synthetic sample achieved a question-level average absolute error of 5.0 percentage points or less, when compared with real ATP data

GPT-5.1	Both models	Claude Opus 4.6
- Whether there are clear solutions to most big issues facing the country today - Whether 2026 will be better than or worse than 2025 - Preference for living in a community where houses are larger and farther apart or smaller and closer to each other - Concern about the cost of housing - Participation in a political campaign, meeting, protest or rally in the last two years - Whether it's acceptable or not for federal immigration officers to wear face coverings that hide their identities while working - Trump presidential approval - Ease or difficulty of someone obtaining an abortion in the area where you live - Whether police should be allowed to stop and search anyone who fits the general definition of a crime suspect	Have an IRA, 401(k) or similar retirement account	- Satisfaction with the way things are going in this country today - Whether voting gives people like you some say about how government runs things - Whether your side has been winning more often than losing in political issues that matter to you over the last few years - Whether the Trump administration's tariff policies will have a positive or negative overall effect on the country - Favorability of suspending all asylum applications from people seeking to live in the U.S. to escape violence or danger - Whether it matters or not which party wins control of Congress in the 2026 elections - Whether Trump will be a successful or unsuccessful president in the long run - Whether Democratic congressional leaders this year should work with Trump to accomplish things or stand up to Trump on issues important to their voters - Whether obtaining an abortion in the area where you live should be harder, easier or about the same as it is now - Favor or oppose pausing visa applications from people in 75 countries seeking to legally immigrate to the U.S. - Whether Trump's economic policies have made economic conditions better, worse or have had not much of an effect - Whether you find talking about politics with people you disagree with interesting and informative or stressful and frustrating - Confidence that Trump has the leadership skills needed to be president

Note: [Refer to the Wave 185 questionnaire](#) for full question wording.

Source: Survey of U.S. adults and "digital twins" synthetic analysis using models set to low reasoning with extended profile information and expert reflection. Survey was conducted Jan. 20-26, 2026 (replicated March 2-3 with OpenAI GPT-5.1; and replicated March 9-12 and April 7-10 with Claude Opus 4.6).

"Can AI Stand In for Human Survey-Takers? Not Really"

Appendix C: Full list of variables in profile information

The standard profile

- FIPS code
- State of residence
- Census division
- Census region
- Metropolitan area indicator
- Congressional district
- Self-identified community type (urban, suburban, rural)
- Volunteered in the past 12 months
- Own, rent or do not pay for housing
- Lived in current house one year ago
- Ever served in the U.S. military or military reserves
- Frequency of internet use
- Race and ethnicity (any combination of Hispanic or Latino, White, Black, Asian, American Indian or Alaskan Native, Native Hawaiian or other Pacific Islander; “some other race or origin;” “mixed race;” or refused to identify a race)
- Hispanic or Latino origin/heritage (among those who identify as Hispanic or Latino)
- Asian origin/heritage (among those who identify as Asian)
- Religion
- Born-again or evangelical Christian
- Frequency of religious service attendance in person
- Importance of religion in their life
- Frequency of prayer
- Marital/relationship status
- Ever married (among those not currently married)
- Parent/guardian of a child of any age (any children under 18; any adult children ages 18 and older)
- Retired
- Disability
- Labor union membership (or similar employee association)
- Political party identification
- Political party lean
- Strength of political party identification
- Closeness to a political party (among leaners)
- Ideology (liberal or conservative)
- Education
- Attended a trade school

- Place of birth (the U.S. or another country)
- Years lived in the U.S. (among those born in another country)
- Citizenship (U.S. or other)
- Voter registration
- Age (calculated from 2025 minus self-reported year of birth)
- Generation
- 2022 congressional election participation (self-reported turnout, self-reported vote choice, validated turnout)
- 2024 presidential election participation (self-reported turnout, self-reported vote choice, validated turnout)
- Gender
- LGBTQ identification
- Number of people in household
- Number of adults in household
- 2024 family income
- 2024 family income tier (upper, middle or lower class, based on regional cost of living)

The extended profile

For a full list of questions used in the extended profile, [refer to the Wave 183 topline](#).⁸

⁸ Variables POLFUT, NATPROBS battery (DEF, IMM, GUN, CLIM, RAC, VCRI and INEQ), GOVSIZE_SM, GOVSIZE_BIG, SNGLPYER, NOGOVTHC, THERMO battery (THERMREP and THERMDEM), SOSEC_RED, SOSEC_KEEP, SHTDWN_KNOW, KNOW_SHTDWN, KNOW_SHTDWN_WHY and SHTDWN_SELF_CONF were not included in the extended profile.