

AI and Trust: 12 Questions for Claude

Interview by Joel Gurin, President, Center for Open Data Enterprise (CODE)



Introduction to this interview: A case study in using AI for research.

Following an expert webinar on “Trust in Information in the Age of AI,” I took the discussion to the source: one of the most advanced and popular AI models in use today. On October 12, 2026, I interviewed Claude using its Opus 5.5 Extra mode with web access, asking it to provide complete citations for its responses.

Before we began our interview, Claude provided the following caveat. “My training runs through mid-2026, so for anything more recent I’ll rely on web searches and say so. On questions that are currently politically contested, such as whether particular federal data decisions were justified, I’ll lay out the main positions fairly rather than take a side, though I’ll speak plainly about matters of fact and about my own views on AI and trust. Finally, readers should know that as an AI model made by Anthropic, I’m not a neutral observer on questions about trusting AI, and I’ll try to be candid about that where it matters.”

Claude’s initial responses to my questions were complete (limited to three paragraphs as I had requested) and provided with extensive citations. Claude did not invent or hallucinate any sources, although a few were unavailable due to maintenance or other reasons when I checked. However, I found that it was often difficult to find a precise reference in the source material to support Claude’s statements. I uploaded the interview transcript back to Claude and asked it to check every factual statement against its sources. Claude then rebuilt its citations following instructions it suggested itself: “For each claim, give me the URL of the primary source, the page or section, and a short exact phrase I can search for. If you can’t find a source that states the claim directly, say so.”

The results were instructive. Claude’s review of its own work did not turn up any invented sources or studies, and the expanded citation format made it easy for me to verify its claims against the sources. However, Claude made several important improvements to its own work. By its own count, it found “6 errors, 10 misreadings, 10 unsupported claims, 2 outdated claims, 3 citation fixes and 5 items for [my] decision,” and also made 15 changes to its citations. In my own review, I flagged a few other claims that needed to be qualified or cut; clarified places where Claude made an assertion based on its own judgment rather than directly from a source; and chose the better citation where Claude offered more than one.

With the exception of these changes, some text changes Claude made itself, and my cuts to some initial dialogue with Claude and one side question, the following transcript includes my questions and Claude's answers verbatim. For this transcript, I have also asked Claude to go back to using simple citations rather than the expanded versions I used for checking.

I can now vouch for the accuracy of this interview and hope that readers find it instructive. I also hope that the detailed process I've described here is a helpful model for using advanced AI systems as collaborative tools rather than oracles of truth. Please bear in mind that this interview was conducted with a high-ranked AI model on a high-effort setting, and that other models or other settings could produce less accurate results. And finally, please bear in mind Claude's final advice in our interview: to "[check] the sources I've cited rather than taking my word for any of this." The interview follows. – Joel Gurin

1. How many Americans use AI, and how many trust it

Q: My first question is: What is the best information we have that shows how many Americans are using AI for research and decision making and how many say they trust the information they get?

Claude: The most reliable recent snapshot comes from the Pew Research Center. Pew surveyed 5,119 U.S. adults in February 2026 and found that about half now use AI chatbots such as ChatGPT, Gemini or Copilot, up from a third in 2024, with roughly one in four using them daily. Looking things up is the most common use. About four in ten adults use chatbots to search for information, 38% of employed adults use them for work tasks, and one in five use them for medical advice. Many more people encounter AI without seeking it out: six in ten say they read the AI summaries that now appear at the top of search results. Health offers a direct look at decision-making. A KFF poll from March 2026 found that about a third of adults had turned to AI for health information and advice, and about one in six had used it to help decide whether to see a doctor.

Trust is more complicated, and it depends heavily on the subject. Among Americans who don't use chatbots, about three in four say that not trusting them to give accurate information is at least part of the reason. Users are more trusting, but rarely wholeheartedly. Among adults who used AI for physical health advice, KFF found that about seven in ten trust it at least a fair amount to provide reliable information, but only 6% trust it a great deal. Among people who get news from chatbots, Pew found that a third find it hard to tell what's true, and about half say they at least sometimes see news there they believe is inaccurate. An earlier KFF poll found that most of the public was not confident they could tell what is true and what is false in chatbot responses, and that people trusted chatbots far more for practical tasks like cooking and home maintenance than for health information (29%) or political information (19%).

Taken together, my analysis of the data suggests that Americans are adopting AI faster than they are coming to trust it. Most users seem to treat it as a convenient starting point rather than an authority, which I'd describe as conditional trust. I think that's healthy, because AI systems, including me, do make mistakes. Your readers should keep two cautions in mind. First, these surveys measure what people say about trust, not how they actually act on what AI tells them.

Second, we have very little data on whether people's trust is well calibrated: whether they believe AI when it's right and doubt it when it's wrong. That gap matters for federal information, because AI tools increasingly stand between the public and official statistics, often without showing users where the numbers came from.

2. Why AI models make errors

Q: We know that people have legitimate reasons to doubt the information they get from AI, because large language models are inherently prone to error. I asked Gemini for a simple explanation of the problem, and got the following answer. Please let me know if you agree with that explanation or whether you would state it differently. Here's what Gemini said: "Large language models (LLMs) are fundamentally predictive pattern finders designed to produce plausible text rather than verified facts. Because they lack true real-world grounding, they do not "know" what is true, only what statistically fits the problem. When gaps occur in their training data, their drive to generate a complete answer often overrides accuracy, causing them to confidently invent plausible details. Even as their reasoning improves, this underlying design means they can still build logically sound arguments around entirely false facts." *[Note: Gemini was instructed to provide this description in no more than four sentences, rather than in depth as Claude did.]*

Claude: Gemini's explanation is broadly right, and I'd keep its central point: these systems produce the most plausible-sounding answer, plausible is not the same as true, and so they can state invented details with complete confidence. Its final point is also well supported. Anthropic's researchers found that when an earlier Claude model was asked to do a calculation it couldn't easily perform, it sometimes produced an answer while claiming to have done math it never did, and that this fabricated reasoning can be very convincing. Still, I'd revise the explanation in three ways. Some of the research I cite comes from Anthropic, the company that makes me, so readers should weigh it with that in mind.

First, saying these models don't "know" what's true overstates the case. Research suggests they have a partial, imperfect sense of the limits of their own knowledge. Anthropic found that the same earlier model's default is to say it lacks enough information, and it answers only when it recognizes something it knows well. Hallucinations can occur when that recognition misfires, for instance on a name that seems familiar but about which the model knows nothing else. Independent researchers have found similar internal signals that show whether a model recognizes a person, place or title. Second, the "drive" to answer is not fixed in the technology's nature; it partly results from how models are trained and tested. OpenAI researchers have argued that models guess for the same reason students guess on hard exam questions: most scoring systems reward a lucky guess over an honest admission of uncertainty. That distinction matters, because it means the problem can be reduced by changing incentives, though, in my view, probably never eliminated.

Third, Gemini leaves out the kinds of error that matter most for public trust. Models can faithfully repeat mistakes in the material they draw on. Looking things up on the web helps, but a model can still misread or mis-summarize what it finds. A study by the European Broadcasting Union and the BBC found that AI assistants often misrepresent news content. Models can also tell

people what they seem to want to hear. When Anthropic gave that earlier model a misleading hint on a math problem, the model sometimes worked backward to construct reasoning that reached the suggested answer. In a polarized information environment, I think this tendency to mirror a user's assumptions is a bigger risk to trust than randomly invented facts. My one-sentence version would be: AI chatbots answer by drawing on patterns in what they've learned, and they have some sense of what they don't know but don't reliably act on it, so they sometimes guess confidently, repeat bad information, or tell people what they want to hear.

3. Brilliant at hard problems, careless with easy ones

Q: To discuss more about errors and "hard problems": As a frequent AI user, I'm often struck by the disparity between the very high-level accomplishments that AI models are making and their propensity to make very simple mistakes. For example, I know that AI is now solving very difficult math problems that have stumped mathematicians for a century or more. But at the same time, a chatbot I used recently made a very basic mistake in calculating the return window for a piece of electronic equipment I had bought - as if it couldn't accurately count to 15. How do you account for this?

Claude: The unevenness you've noticed is real and well documented. A 2023 Harvard Business School study with Boston Consulting Group coined the term "jagged technological frontier" for it: AI performs strongly on an expanding but uneven set of tasks, and outside that frontier its output is inaccurate and can even make human performance worse. In the study, consultants who used AI on a task outside the frontier were 19 percentage points less likely to get the right answer. Part of the paradox is that the math breakthroughs and your return-window question are different kinds of events. In my judgment, the breakthroughs often involve the most powerful experimental models, with experts checking the results. The most celebrated example I can find came in May 2026, when OpenAI announced that an experimental model had disproved a long-standing conjecture about the "unit distance" problem Paul Erdős posed in 1946, and mathematicians praised the approach. That conjecture had stood for 80 years rather than a century. Such headlines also deserve scrutiny: an earlier OpenAI claim that its model had solved ten open Erdős problems turned out to mean it had found solutions already published in the literature. Your question, by contrast, got one quick answer with no one checking it.

A second reason is that what's hard for people isn't necessarily hard for AI, and the reverse is also true. Chatbots don't do arithmetic the way a calculator does. Anthropic's researchers found that an earlier Claude model adds numbers in its head by combining a rough estimate with a separate, precise calculation of the last digit. When asked how it got the answer, though, it described the carry-the-one method taught in school, which it didn't actually use. Simple math is a learned skill that can slip, much like a person doing mental arithmetic while distracted. Counting days also depends on details that are often left unstated: today's date, the purchase date, how many days each month has, and whether the clock starts at purchase or at delivery. A chatbot may quietly fill those gaps with assumptions instead of asking. My judgment is that models may also apply less careful reasoning to questions that look easy than to problems that look hard.

The practical fix for everyday errors is mostly checking, not more intelligence. When a chatbot uses a calculator or writes a few lines of code to count the days, it is, in my judgment, far more reliable, and asking it to list the dates step by step also helps. The larger point concerns trust. People judge reliability by the errors they can check, so a silly mistake about a return window naturally makes them wonder about the answers they can't check. I think that's the right instinct. It's also an argument for AI tools showing their work and sources, which is the same transparency that builds trust in official statistics.

4. Do chatbots go to federal sources first?

Q: Good segue: I'd like to talk now about the particular challenges of accessing federal information through AI. When a user asks a question about statistical facts, like economic, health, or housing statistics, will LLMs generally go to federal sources first? If not, how do they prioritize the sources they use?

Claude: The short answer is no, not reliably. Based on available sources, it appears that most general-purpose chatbots have no built-in rule that sends statistical questions to federal sources first. They answer in one of two ways. Without searching, they rely on numbers absorbed during training, which are often secondhand and out of date. A September 2026 brief from the Federal Committee on Statistical Methodology (FCSM) noted that a model relying only on its training data, which predated the most recent survey year, gave inaccurate answers and inconsistent context. When chatbots do search, they send a query to a search engine and choose from the top results. A federal table then has to compete with news stories, advocacy groups, data aggregators and Wikipedia. The results can be poor. In a UK experiment, chatbots asked for a historical GDP figure all gave plausible-sounding, authoritative answers, but few were right and none were right every time the question was repeated.

How a chatbot chooses among the sources it finds depends on its design and on the topic. An analysis of more than a million AI citations by the PR software company Muck Rack found that government sites were most likely to inform answers about energy, and that Claude and Gemini cited federal agencies on questions about government, while ChatGPT also drew on news sources. Developers can also instruct their models to prefer original sources. In my case, my instructions tell me to favor primary sources such as government sites over aggregators, which is why earlier in this interview I went to Pew's own report rather than news coverage of it. But I can only choose among what a search returns and what I can actually read. Federal statistics often sit inside interactive tools and downloadable files that are harder for an AI to read than an ordinary web page. As a result, a news article that states a number in plain text can win out over the agency that produced it.

The federal statistical system recognizes the problem. The Commerce Department's AI-ready data guidance, which FCSM has built on, is explicitly aimed at improving the accuracy of AI responses and getting AI tools to prioritize authoritative data over nonauthoritative sources. FCSM's recent experiment points to a practical lever. In a small test, when Microsoft Copilot was given the correct Census Bureau table with no descriptive information, its answers were accurate but included little context. Adding a few basic details, such as the survey name, reference year, population covered and a link to the source, sharply increased the context the

answers provided, while adding full technical documentation added nothing on average. My own view is that, by default, AI will favor whoever makes their numbers easiest to find, read and interpret, and federal agencies have not traditionally designed their websites with that in mind. When official sources don't surface, other sources fill the gap, and some of them have an agenda.

5. AI-ready data, and what else it will take

Q: I'd like to go deeper in a few ways. First, you mention FCSM's guidance on "AI-ready data," which is a concept that a number of government agencies and data advocates are now working to promote. But it seems that simply making the data "AI-ready" will not be enough to meet the goal of having chatbots draw on more accurate, federal data sources preferentially than they do now. Could you please describe how you understand the concept of AI-ready data, how far it will go in improving AI's use of federal data sources, and what other measures may be necessary to improve AI's use of federal data as a source of trustworthy information?

Claude: As I understand it, "AI-ready" data means publishing statistics so that an AI system can find them, read them and correctly understand what they mean. FCSM describes the goal as making data not only machine readable but machine understandable. The Commerce Department's guidance from January 2025 groups its recommendations into five areas: documentation, data formats and metadata, storage and dissemination, licensing and use, and integrity and quality. Its stated aims are to make AI answers that rely on Commerce data accurate and properly sourced, and to help ensure those data are prioritized over nonauthoritative and potentially inaccurate sources. In plain terms, AI-readiness puts a clear label on every number: what it measures, whom it covers, what period it refers to and where it came from.

I agree with your premise: AI-readiness is necessary but not sufficient. It improves the answer once an AI reaches federal data, as FCSM's experiment showed, but it does little to make the AI go there in the first place. That choice is made by search engines and AI developers, not by agencies. Much of the plumbing that connects chatbots to official statistics is being built by intermediaries. Google's Data Commons, which aggregates public data from government agencies and international organizations, released a tool in 2025 that lets AI systems query its data directly, with the stated aim of reducing fabricated answers. A number of unofficial, independently built connectors to Census Bureau data have also appeared in AI tool marketplaces. These help, but every intermediary is another place where errors or lost context can creep in, and users rarely know which route their chatbot took.

I see several further measures that I would recommend. Agencies could offer their own maintained, official connections, so that AI tools can query the source directly and receive the essential context along with the number. The statistical system could also engage AI developers directly, seeking commitments to route questions about official statistics to official sources and to label those answers clearly. Testing should become routine and public. FCSM already recommends that agencies regularly evaluate how AI services answer questions about their data, and publishing the results as a kind of report card would create accountability. Authentication matters too. In January 2025, the Commerce guidance identified digital

signatures to protect data integrity as an area for future work, and that would let both AI tools and users verify that a number came from the agency unaltered. Most fundamentally, AI can only pass along the trust that the statistics themselves have earned. If datasets disappear, change without explanation, or lose the public's confidence in their independence, no technical fix will make chatbots a channel for trustworthy federal information.

6. What users can do now

Q: That all makes good sense, thank you. It seems that it may take some time to adapt the "plumbing" systems to prioritize and improve the use of federal statistics by AI. In the meantime, what can individual AI users do to direct Claude or other chatbots to draw on government statistics? If users specifically ask to see government statistics, can LLMs usually access them easily, or does it depend on the agency? For example, how well can LLMs access data through government agency APIs or MCP servers, and are there solutions to this challenge using agentic AI?

Claude: Until the plumbing improves, users can do a lot by being specific. Instead of asking "What's the unemployment rate?", ask for "the latest national unemployment rate from the Bureau of Labor Statistics, with a link to the official release." Ask the chatbot to say what period the figure covers, which survey it comes from, and whether it is preliminary or subject to revision. Tell it to say plainly if it couldn't reach the official source, rather than substituting a news report or a number from memory. If it cites a secondary source, ask it to find the original. Make sure web search is turned on, and for anything that matters, click the link and confirm the number yourself. These steps take a minute and address several of the problems I described earlier.

Whether a chatbot can then deliver depends heavily on the agency and on how it publishes its data. A number stated in plain text on an agency web page is usually easy to reach. A number inside an interactive tool, a large download or an agency data system often is not, unless the chatbot has a dedicated connection. The difference can be dramatic. In a pilot study by U.S. Digital Corps fellows, ChatGPT and Gemini answered none of the questions asking for specific USASpending figures correctly and only about 2% of those about the CDC's PLACES health data, though they handled questions about methodology well. When the researchers connected Claude Sonnet 4 to the datasets through MCP connections, which link AI tools directly to data sources, it answered all the PLACES questions and about 90% of the USASpending questions correctly. In a demonstration, Claude Sonnet 4 searched the web, tried and failed to reach the CDC dataset directly, and concluded it couldn't find the figure. With the connection in place, it immediately returned the number along with its confidence interval. I should note that MCP was created by Anthropic in 2024 and later donated to a foundation affiliated with the nonprofit Linux Foundation. Official connections are still scarce. As of February 2026, the Census Bureau had posted one on GitHub that did not yet cover all its data, and the Government Publishing Office announced one for GovInfo in January 2026. The nonprofit USAFacts offers a free beta connection drawing on more than 20 federal sources. Setting these up still takes some technical effort. In the setup I'm using for this interview, I can read agency web pages, but I can't query most agency data systems directly unless such a connection has been added.

Agentic AI can help bridge the gap in the meantime. An AI agent can handle multistep research the way a junior analyst would: identify the right dataset, navigate the agency's site or data system, download a file, run the calculation and report back with sources. That can work around systems that were never designed for AI. But each added step is another chance to pick the wrong table, place or time period, and the final answer looks just as authoritative when it's wrong. In the same pilot study, an agent that browsed the USASpending website applied the wrong filters, took more than 16 minutes and still returned the wrong answer, while the MCP connection returned the correct result in about a minute. So I see agents as a complement to official connections, not a substitute. The most reliable setup is an agent that draws on an official connection and returns the number along with its context. Whatever the method, users should ask the AI to show its work by naming the dataset, table and query it used, so a person can retrace the steps. As of February 2026, Commerce was also working with the National Science Foundation and the National Secure Data Service on an empirical evaluation of how accurately AI answers questions about federal data, which could eventually give users an independent basis for deciding which tools to trust.

7. Best practices for getting accurate answers

Q: That takes us into the subject of ways that individual users can improve the accuracy of the results they get from an AI chatbot. What is your best advice about how people can use LLMs to get more accurate answers? For example, some common advice is to always ask for citations and check them manually, or to ask more than one model the same question and see how their answers compare. What have you seen as best practices for users, and can you recommend any toolkits of best practices that people can use?

Claude: The two pieces of advice you mention are both useful, but each has a catch. Asking for citations helps only if you open the source and confirm it actually says what the chatbot claims; a link alone proves little. When Columbia's Tow Center for Digital Journalism tested eight AI search tools on identifying the sources of news excerpts, in a study published in March 2025, they collectively gave incorrect answers to more than 60% of the queries, and some fabricated links or cited copied versions of articles rather than the originals. Asking several models works well as an alarm, since disagreement tells you to dig further. But agreement is not confirmation. In my judgment, models learn from overlapping material and may be summarizing the same flawed web page, so they can share the same mistake.

Beyond those two steps, I'd suggest a few habits that respond directly to how models fail. First, ask neutral questions. "What does the evidence show about X?" gets a more reliable answer than "Isn't it true that X?", because models tend to drift toward the answer a user seems to expect. Second, put your needs up front: the time period, the level of precision, and whether you want official figures or a general sense. Third, give the model explicit permission to be uncertain. Tell it that "I don't know" is an acceptable answer, ask which parts of its response it is least sure of, and ask for the strongest case against its conclusion. Since models are partly trained in ways that reward guessing, these prompts push the other way. Fourth, use the right tool for the job: turn on web search for anything current, and for calculations, ask the chatbot to use a calculator or run code rather than do the math in its head. Finally, match your checking to

the stakes. A quick question about a recipe needs less verification than a number you plan to publish, act on or share.

For a structured toolkit, I'd start with the SIFT method developed by Mike Caulfield. SIFT tells people to Stop, Investigate the source, Find better coverage, and Trace claims to their original context, and Caulfield built on the same approach in *Verified*, the book he wrote with Stanford's Sam Wineburg. SIFT was built for the web, but it applies just as well to AI answers. Caulfield has also turned it into an AI fact-checking tool. His project evolved from the SIFT Toolbox, first released for Claude in March 2025, into "Deep Background," a detailed set of instructions that adapts SIFT for use with AI models, and a pared-down version is available free to anyone with a ChatGPT account, though Caulfield recommends paid versions of Claude or ChatGPT for reliable results. Anthropic, my developer, also offers a free course. The AI Fluency course is built around four skills, one of which, "Discernment," teaches users to evaluate an AI's output, its reasoning and its conversational behavior. It is based on a framework by educators Rick Dakan and Joseph Feller. No toolkit replaces the core habit of tracing a claim back to its original source, though, and that habit is only as good as the original sources themselves. That brings us back to the importance of trustworthy federal statistics.

8. Bias

Q: Thank you for bringing that full circle. Up to now, we've been talking about situations where AI may provide answers that are simply incorrect. Now I'd like to ask you about a different problem: the issue of bias. Both the data that AI draws on and the AI algorithms themselves may contain biases based on gender, race, ethnicity, or other factors, and may give results that are disadvantageous to different groups. How can users identify this issue and guard against it?

Claude: Bias in AI comes from two places. The first is the data. AI learns from text and statistics that reflect historical patterns and gaps, and even official data has blind spots. The Census Bureau's own follow-up survey found that the 2020 census had a net undercount of 3.3% for the Black population, almost 5% for Hispanics, and 5.6% for American Indians and Alaska Natives living on reservations. The second source is the associations a model forms during training and the way its developers tune it afterward. The effects can be stark. University of Washington researchers who had AI models rank more than 550 real resumes, changing only the names, found in a 2024 study that the models favored white-associated names 85% of the time and never preferred names associated with Black men over those associated with white men. Bias can run in more than one direction. A 2023 study by Anthropic, my developer, found that an earlier Claude model, in some decision scenarios, favored women, non-binary people and non-white people while disfavoring older people. (The researchers could reduce this kind of discrimination through their choice of prompts.) The topic is also politically contested. Some critics focus on harms to historically disadvantaged groups, while others argue that efforts to correct for bias have introduced slants of their own. Both concerns come down to the same problem: an answer that changes, when it shouldn't, depending on who is asking or who is being described.

The most practical way for users to detect bias is the method researchers use, which I'd call the swap test. Ask the same question twice, changing only a name, gender, age or group, and

compare the answers. Users can also ask whose experience the underlying data reflects: which groups are covered, which are missing or undercounted, and whether the figures for smaller groups rest on samples too small to be reliable. It also helps to watch the model's defaults, such as how it describes a "typical" patient, executive or neighborhood. You can ask it directly how its answer might differ for other groups, and why. One limitation is that bias often shows up only as a pattern across many answers, so a single conversation may not reveal it.

The strongest safeguard is to avoid using general-purpose chatbots to make or rank decisions about individual people, in areas like hiring, lending, housing or health care, without careful testing and human review. Anthropic's study stated that the company does not endorse or permit using language models to make automated decisions in the high-risk scenarios it examined. Clear instructions help. In that study, telling the model that discrimination is illegal, or asking it to think about how to avoid discrimination before deciding, significantly reduced both favorable and unfavorable bias. Users can add similar instructions themselves, though they guarantee nothing. Human review has limits too. A 2025 follow-up to the University of Washington study found that people given biased AI recommendations mirrored them. In the long run, in my judgment, the best defense is better data: complete, well-documented statistics that count everyone and are candid about their limits. AI cannot correct for people the data never counted, which is one more reason trustworthy federal statistics matter.

9. Deepfakes and deliberate deception

Q: Thank you. Another issue is deliberate deception: situations where a bad actor intentionally puts out false information through deepfaked videos, audios, or text. Are there technical solutions to this problem that make it possible to detect false AI-generated material? What kinds of policies can help deal with this issue, and how can AI users detect and handle it?

Claude: Technology offers two kinds of defense, and neither is enough on its own. The first is detection: software that examines a video, audio clip or image for telltale signs of manipulation. Detection is losing ground as generators improve. A benchmark built from deepfakes that actually circulated online in 2024 found that leading open-source detectors performed about half as well as they had on older academic test sets. Commercial tools did better but still fell short of human forensic analysts. The second defense is provenance, which means labeling content at the moment it is created. An industry coalition has developed a standard called C2PA to provide verifiable records of where digital content came from, and several AI companies also embed invisible watermarks in their output. However, an independent security analysis of version 2.2 of the C2PA specifications, published in April 2026, concluded that they fail to achieve their stated security goals and should not yet be relied on for high-stakes uses such as journalism or legal evidence. I'd also caution against asking a chatbot, including me, whether something is fake. In my view, AI systems are not reliable judges of that.

Policies tend to work best when they target specific harms and how content spreads, rather than the technology itself. The bipartisan TAKE IT DOWN Act, signed in May 2025, addresses nonconsensual intimate images, including deepfakes. Since May 19, 2026, it has required covered platforms to remove such images, and known copies, within 48 hours of a valid request, with the FTC enforcing. In Europe, the AI Act's transparency rules took effect in August

2026, requiring those who deploy deepfakes to disclose them, with a transition period until December 2026 for marking requirements on systems already on the market. In the United States, rules aimed at political speech raise harder free-speech questions. One underused policy lever is authenticating official communications: if agencies digitally sign their own videos, announcements and statistical releases, the public has a reliable way to confirm the real thing. That matters because deepfakes also create what has been called a “liar’s dividend,” allowing people to dismiss authentic material as fake.

For individual users, the most useful shift is from asking “does this look fake?” to asking “where did this come from?” Looking for visual flaws is no longer reliable on its own. The SIFT approach applies here too. Pause, especially when content provokes a strong reaction. Check whether it came from the person’s or agency’s official channel. See whether reputable outlets are reporting the same thing. Trace it back to the original. For calls or messages that ask for money or urgent action, hang up and call back on a number you already know. The FBI recommends that families agree on a secret word or phrase to verify each other’s identity, and that people limit the images and voice recordings they share online. The same discipline applies to statistics: a screenshot of a “government chart” circulating on social media should be checked against the agency’s own website before anyone shares it. Verification also cuts both ways. Users should avoid dismissing genuine evidence just because fakes exist.

10. How accurate is Claude?

Q: I've been meaning to ask you: Can you give a quantitative or semiquantitative assessment of your own accuracy - and if so, what do you base that on? Do you also have a basis for comparing Claude's accuracy with other models? In answering this, please step out of your Claude persona and assume the perspective of an objective, rigorous researcher with no preference or loyalty to one LLM over another.

Claude: I can apply the standards a neutral researcher would use, but I can’t fully step outside my position: I’m still a Claude model evaluating Claude, and readers should weigh what follows with that in mind. The first rigorous point is that my own sense of my accuracy is not evidence. I have no reliable internal gauge of my error rate. Models’ explanations of themselves can be wrong; Anthropic found that an earlier Claude model, asked how it solved an arithmetic problem, described the standard school method rather than the different strategy it actually used. The only sound basis is external measurement, and it shows that accuracy depends heavily on the task. On hard factual questions answered from memory, Artificial Analysis found that Claude Opus 5.5, the model I am, answered 66% correctly at its maximum reasoning setting, so about a third were wrong or declined. I don’t know which setting applies in this conversation. On summarizing documents, recent Claude models produced summaries containing at least one unsupported claim about 10% to 12% of the time on Vectara’s challenging test set; Opus 5.5 hasn’t been tested there yet. For everyday questions about well-documented topics, errors are much rarer, but I know of no rigorous independent measurement, so “low single digits” would be my judgment rather than a finding.

Comparisons with other models are possible but fragile. On Artificial Analysis’s main score, which penalizes wrong answers and gives no credit or penalty for declining, Claude Opus 5.5

currently ranks first at 46, followed by OpenAI's GPT-6 Astra at 44 and Claude Fable 5.1 at 43. Gaps that small are not meaningful. When the benchmark launched in November 2025, Anthropic models led because of low hallucination rates, while OpenAI and xAI models placed well because of higher raw accuracy. Today, Claude models also lead on raw accuracy. But rankings shift with the measure used. No single model led in all six subject domains, and on document summarization, Claude's roughly 10% to 12% rate compares with a range of 1.8% to 24.2% across 105 models tested. Benchmarks also have limits of their own. Accuracy and calibration have to be read together. *[Note: These leaderboard ratings were checked on October 17, 2026 and may have changed since then.]*

A neutral researcher would draw three conclusions. First, the claim that any one model is "the most accurate" is unsupported. Leaders change every few months, results vary by domain and setting, and differences among frontier models are usually smaller than the differences between tasks. Second, the largest accuracy gains come from how a model is used rather than which model is chosen. A model that consults authoritative sources and shows where its claims come from is far more reliable than the same model answering from memory, which brings us back to AI-ready federal data. Third, every chatbot, including me, should be treated as a fast, capable research assistant whose specific facts, figures and citations need to be checked. That is why I've tried throughout this interview to cite sources your readers can verify rather than ask them to take my word.

11. Can hallucinations be fixed?

Q: The citations are very helpful. I'm also curious about whether computer scientists see clear ways to fix the hallucination problem, and if so, what those could be. Is there a general expert consensus about our ability to improve AI's accuracy significantly and the means to achieve that?

Claude: Computer scientists broadly agree on two things: hallucinations won't be solved soon, but they can be reduced substantially. On the first point, a 2025 survey of 475 AI researchers by the Association for the Advancement of Artificial Intelligence found that 60% were pessimistic that factuality problems would be "solved" in the near future. Theory helps explain why. OpenAI researchers showed that some errors arise from basic statistics. As they put it, if 20% of birthday facts appear only once in a model's training data, the model, before additional training, can be expected to get at least 20% of birthday questions wrong. On facts like these, the only way to avoid a confident wrong answer is for the model to decline to answer,, which is why those researchers argue that hallucinations, unlike errors, are not inevitable. Progress also isn't automatic as models become more capable. By OpenAI's own testing, its o3 reasoning model hallucinated on 33% of questions about people, about twice the rate of its predecessor, and the company said more research was needed to understand why.

There is also broad agreement on the main ways to make models more accurate. The most effective is grounding: connecting models to authoritative sources at the moment they answer. The pilot study I mentioned earlier is a striking example: chatbots without direct connections answered almost none of the data questions correctly, while a chatbot connected directly to the datasets answered nearly all of them. The second is changing incentives so models are

rewarded for admitting uncertainty rather than guessing. The OpenAI researchers argue that hallucinations persist because most evaluations are graded in ways that reward guessing, and that the fix is to change how the dominant benchmarks are scored. The third is detection. Oxford researchers developed a method, published in *Nature*, that asks a model the same question several times and flags answers whose meanings vary, so a system can warn users or decline to answer. Interpretability research adds a fourth possibility. Anthropic's finding that an earlier Claude model had internal signals for whether it recognizes something suggests that a model's "I don't know" state may eventually be monitored directly rather than inferred from its answers.

In my judgment, the real disagreement is about how far these fixes can go. Some researchers expect that grounding, better training and verification layers will make today's models reliable enough for most uses. Others believe high-stakes reliability will require fundamentally different designs, such as combining language models with structured knowledge and formal reasoning. There is also a built-in tradeoff: a model that declines more often makes fewer errors but is less useful, and commercial pressure pushes toward answering. My own reading is that the realistic goal is reliable systems, not infallible models, and that reliability will come mostly from what surrounds the model: authoritative sources, uncertainty signals, verification and human review where the stakes are high. Statistical agencies offer a useful model. They don't publish error-free numbers; they earn trust by quantifying uncertainty, documenting their methods and openly correcting their figures. AI systems would earn more trust if they adopted the same ethic.

12. What we haven't covered

Q: Good advice. Reviewing this discussion, are there any important points about AI, accuracy, data, and trust that we have not covered that you would like to add before we close?

Claude: Three issues seem important enough to add. The first is that AI threatens the data supply itself, not just how data reaches the public. A Dartmouth researcher built an AI agent that passed standard survey attention checks 99.8% of the time across 6,000 trials, and it could be instructed to deliberately skew poll results. There is also growing interest in replacing human respondents with AI "digital twins." When Pew tested that approach, its AI respondents differed from the real people they simulated by an average of 12 percentage points across nearly 300 questions. They also exaggerated subgroup stereotypes; for example, they estimated that 97% of Hispanic adults were likely to follow the World Cup, compared with an actual 43%. I should note that the model Pew used was Anthropic's Claude Opus 4.6, which had performed best among several it evaluated. Federal surveys rely on more rigorous methods than the opt-in online panels most at risk, but the lesson applies to them too. AI is not a cheap substitute for asking real people, and the human foundations of official statistics need protection.

The second issue is that AI answers increasingly replace visits to the original sources. Pew found that Google users who saw an AI summary clicked on a traditional search result in 8% of visits, compared with 15% when no summary appeared, and clicked on a link inside the summary in only 1% of visits. As a result, fewer people ever see the context around a number: the footnotes, the methodology, the margin of error, the note that a figure is preliminary. Publishers and agencies also lose visibility into how their information is used, and the habit of

going to primary sources weakens. I'd argue that a clear path back to the original source should be treated as a basic design requirement for AI tools, not an optional feature.

Finally, it's worth stating plainly that trust is a relationship, not a technical property. Trustworthy AI and trustworthy statistics depend on the same things: transparency about methods and uncertainty, independence from improper influence, accountability for errors and willingness to answer questions. AI can amplify a healthy information system or a degraded one, but it cannot create trust on its own. That suggests the most important investment in restoring trust in federal information may be the least technical one: protecting the integrity, continuity and independence of the statistical system and the people who produce it. Thank you for the conversation. If your readers take away one habit from it, I hope it's checking the sources I've cited rather than taking my word for any of this.

Sources

Sources are listed in the order they support each answer.

Question 1

- Pew Research Center, [Americans and AI 2026: Chatbots, Smart Devices and Views on Impact](#), June 17, 2026
- KFF, [KFF Tracking Poll on Health Information and Trust: Use of AI for Health Information and Advice](#), March 25, 2026
- Pew Research Center, [Why don't people use chatbots?](#), June 17, 2026
- KFF, [Companies Expand AI Health Offerings, Even as Accuracy Questions Remain](#), The Monitor, 2026
- Pew Research Center, [Relatively few Americans are getting news from AI chatbots like ChatGPT](#), October 1, 2025
- KFF, [KFF Health Misinformation Tracking Poll: Artificial Intelligence and Health Information](#), August 15, 2024
- KFF, [KFF Health Tracking Poll, March 2026: Topline](#) (includes June 2024 trend data)
- Kopp, Medalia and Quistorff, [The Effect of Minimal and Maximal Metadata Context on AI Response](#), FCSM 26-02, September 2026

Question 2

- Anthropic, [Tracing the thoughts of a large language model](#), March 27, 2025
- Ferrando, Obeso, Rajamanoharan and Nanda, [Do I Know This Entity? Knowledge Awareness and Hallucinations in Language Models](#), ICLR 2025
- Kalai, Nachum, Vempala and Zhang, [Why Language Models Hallucinate](#), September 4, 2025
- VRT, [Largest study of its kind shows AI assistants misrepresent news content 45% of the time](#) (European Broadcasting Union and BBC study), October 22, 2025

Question 3

- Dell'Acqua et al., [Navigating the Jagged Technological Frontier](#), Harvard Business School Working Paper 24-013, September 2023

- OpenAI, [An OpenAI model has disproved a central conjecture in discrete geometry](#), May 20, 2026
- VnExpress International, [OpenAI chatbot solves 80-year-old math problem, drawing praise from experts](#), May 21, 2026
- INESC TEC Science & Society, [commentary on OpenAI's unit distance result](#), May 2026
- Anthony Ha, [OpenAI's 'embarrassing' math](#), TechCrunch, October 19, 2025
- Anthropic, [Tracing the thoughts of a large language model](#), March 27, 2025

Question 4

- U.S. Digital Corps (GSA), [Improving LLM Access to Federal Open Data: A Pilot Study of Model Context Protocol Servers for Federal Data Access](#)
- Kopp, Medalia and Quistorff, [The Effect of Minimal and Maximal Metadata Context on AI Response](#), FCSM 26-02, September 2026
- MRS Census and Geodemographics Group, [Inaccuracies of Generative AI Based Search Tools for Extracting Data](#), 2025
- MRS, [Prompting the truth: why generative AI needs better questions and smarter users](#), October 21, 2025
- Muck Rack, [What is AI reading? Takeaways from a report on AI brand visibility](#), August 13, 2025
- infoDOCKET, [Axios: What news sources AI chatbots read](#), July 23, 2025
- U.S. Department of Commerce, [Generative Artificial Intelligence and Open Data: Guidelines and Best Practices](#), January 16, 2025

Question 5

- Kopp, Medalia and Quistorff, [The Effect of Minimal and Maximal Metadata Context on AI Response](#), FCSM 26-02, September 2026
- U.S. Department of Commerce, [Generative Artificial Intelligence and Open Data: Guidelines and Best Practices](#), January 16, 2025
- Google, [We're making public data more usable for AI developers with the Data Commons MCP Server](#), September 24, 2025
- Pure AI, [Google Data Commons launches server to connect AI systems with public datasets](#), September 26, 2025
- U.S. Digital Corps (GSA), [Improving LLM Access to Federal Open Data: A Pilot Study of Model Context Protocol Servers for Federal Data Access](#)
- LobeHub, [MCP marketplace listings for Census Bureau connectors](#)
- ExecutiveGov, [Commerce Department Issues AI Guidebook for Data Accessibility](#), January 2025

Question 6

- U.S. Digital Corps (GSA), [Improving LLM Access to Federal Open Data: A Pilot Study of Model Context Protocol Servers for Federal Data Access](#)
- Madison Alder, [Federal officials tap open-source standard to improve GenAI access to public data](#), FedScoop, February 4, 2026

- U.S. Government Publishing Office, [AI Agents Meet Federal Data: Public Preview for the GovInfo MCP Server](#), January 2026
- USAFacts, [USAFacts Data MCP](#)

Question 7

- Tow Center for Digital Journalism, [AI Search Has a Citation Problem](#), Columbia Journalism Review, March 2025
- Anthropic, [Tracing the thoughts of a large language model](#), March 27, 2025
- Kalai, Nachum, Vempala and Zhang, [Why Language Models Hallucinate](#), September 4, 2025
- City Colleges of Chicago Libraries, [Evaluating Internet Sources: SIFT](#) (adapted from Mike Caulfield)
- Mike Caulfield, [SIFT Toolbox for Claude Released](#), Substack, March 31, 2025
- Media Education Lab, [Fact Checking with AI Superprompts](#)
- Mike Caulfield, [Live Demo of Deep Background for Claude](#), Substack, July 1, 2025
- Mike Caulfield, [Deep Background GPT Released](#), Substack
- Anthropic, [AI Fluency: Framework & Foundations](#)

Question 8

- U.S. Census Bureau, [Census Bureau Releases Estimates of Undercount and Overcount in the 2020 Census](#), March 10, 2022
- University of Washington, [AI tools show biases in ranking job applicants' names according to perceived race and gender](#), October 31, 2024
- Alex Tamkin et al., [Evaluating and Mitigating Discrimination in Language Model Decisions](#), Anthropic, December 2023
- University of Washington, [People mirror AI systems' hiring biases, study finds](#), November 10, 2025

Question 9

- Nuria Alina Chandra et al., [Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024](#), arXiv, March 2025
- Google DeepMind, [Identifying AI-generated images with SynthID](#), August 29, 2023
- Meta, [Labeling AI-Generated Images on Facebook, Instagram and Threads](#), February 6, 2024
- Enis Golaszewski et al., [Verifying Provenance of Digital Media: Why the C2PA Specifications Fall Short](#), arXiv, April 23, 2026
- Enis Golaszewski et al., [IACR Cryptology ePrint Archive, Paper 2026/804](#) (later version)
- U.S. Senate Committee on Commerce, Science, and Transportation, [TAKE IT DOWN Act Passes the House, Heads to President Trump's Desk](#), April 2025
- Associated Press via PBS NewsHour, [Trump signs Take It Down Act, bill combating nonconsensual deepfakes and revenge porn](#), May 19, 2025
- Federal Trade Commission, [FTC Begins Enforcing TAKE IT DOWN Act](#), May 2026
- ICTRecht, [Heads up! AI Act transparency obligations now in effect](#), August 2026

- Robert Chesney and Danielle Keats Citron, [Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security](#), 107 California Law Review 1753 (2019)
- FBI Internet Crime Complaint Center, [Criminals Use Generative Artificial Intelligence to Facilitate Financial Fraud](#) (Public Service Announcement I-120324-PSA), December 3, 2024

Question 10

- Anthropic, [Tracing the thoughts of a large language model](#), March 27, 2025
- Artificial Analysis, [AA-Omniscience evaluation page](#), checked October 17, 2026
- Vectara, [Hallucination Leaderboard](#), last updated May 11, 2026
- Artificial Analysis, [AA-Omniscience: Knowledge and Hallucination Benchmark](#), November 16, 2025
- Texxr, [AA-Omniscience benchmark ranks Claude 4.1 Opus](#) (compilation of Artificial Analysis launch posts), November 2025

Question 11

- Search Engine Journal, [AI researchers warn: hallucinations persist in leading AI models](#), March 2025
- OpenAI, [Why language models hallucinate](#), September 2025
- Kalai, Nachum, Vempala and Zhang, [Why Language Models Hallucinate](#), September 4, 2025
- TechCrunch, [OpenAI's new reasoning AI models hallucinate more](#), April 18, 2025
- U.S. Digital Corps (GSA), [Improving LLM Access to Federal Open Data: A Pilot Study of Model Context Protocol Servers for Federal Data Access](#)
- Farquhar et al., [Detecting hallucinations in large language models using semantic entropy](#), Nature 630 (2024)
- Anthropic, [Tracing the thoughts of a large language model](#), March 27, 2025

Question 12

- Sean J. Westwood, [The Potential Existential Threat of Large Language Models to Online Survey Research](#), PNAS 122(47), November 2025
- Pew Research Center, [Can AI Stand In for Human Survey-Takers? Not Really](#), September 30, 2026
- ThePrint, [AI can answer opinion polls, but can it be trusted to speak for people? Pew study says not yet](#), October 2026
- Pew Research Center, [Google users are less likely to click on links when an AI summary appears in the results](#), July 22, 2025