

What's bad about AI.

Amid today's AI boom, it's disconcerting that we still don't know how to measure how smart, creative, or empathetic these systems are. Our tests for these traits, never great in the first place, were made for humans, not AI. Plus, AI test scores can change dramatically based simply on how questions are phrased. Even famous challenges like the Turing Test, where humans try to differentiate between an AI and another person in a text conversation, were designed as thought experiments at a time when such tasks seemed impossible. But now that a new paper shows that AI passes the Turing Test, we need to admit that we really don't know what that actually means.

So, it should come as little surprise that one of the most important milestones in AI development, Artificial General Intelligence, or AGI, is badly defined and much debated. Everyone agrees that it has something to do with the ability of AIs to perform human-level tasks, though no one agrees whether this means expert or average human performance, or how many tasks and which kinds an AI would need to master to qualify. Given the definitional morass surrounding AGI, illustrating its nuances and history today is challenging.

source: https://www.oneusefulthing.org/p/on-jagged-agi-o3-gemini-25-and-everything?utm_source=post-email-title&publication_id=1180644&post_id=161512556&utm_campaign=email-post-title&isFreemail=true&r=1os3dy&triedRedirect=true&utm_medium=email

Long term loss for short term gain.

The following example from the Pragmatic Engineer shows one of the pitfalls when AI is used for short term gain in a way that is self destructive in the long term

Source: https://newsletter.pragmaticengineer.com/p/are-llms-making-stackoverflow-irrelevant?utm_source=post-email-title&publication_id=458709&post_id=155351457&utm_campaign=email-post-title&isFreemail=true&r=5529lp&triedRedirect=true&utm_medium=email

Are LLMs making StackOverflow irrelevant?

Fresh data shows that the number of questions asked on StackOverflow are as low as they were back in 2009 – which was when StackOverflow was one years old.

The volume of questions asked on StackOverflow started to fall quickly after ChatGPT was released in November 2022, and the drop continues into 2025 at alarming speed.

StackOverflow has not seen so few questions asked monthly since 2009! The graph shows the steep drop-off in usage accelerated with the launch of OpenAI's chatbot, and it's easy enough to figure out why: LLMs are the fastest and most efficient at helping developers to get "unstuck" with coding.

Before the rise of this technology, StackOverflow was the superior option to Googling in the hope of finding a blog post which answered a question. And if you couldn't find an answer to a problem, you could post a question on StackOverflow and someone would probably answer it.

Another question: where will LLMs get coding Q&A training data in the future? In some ways, it feels to me that StackOverflow is the victim of LLMs ingesting data on its own Q&A site, and providing a much better interface for developers to solve programming problems with. But now the site gets far fewer questions and answers, where will training data come from?

This is a question with no clear answer, that's similar to the one about where the next generation of entry-level software engineers will come from, when most businesses hire fewer than before because LLMs can do roughly the same job as a newly qualified human?

I expect the industry will adapt: perhaps LLMs in the future won't be as good as today in answering StackOverflow-like questions, but will have other more advanced capabilities to make up for it; like trying various solutions and validating them, or coding agents might become more helpful.

The same applies to the question of entry-level engineers: the tech industry has always adapted, and I don't see it being different this time, either.

UPDATE

According to a followup article from the Pragmatic Engineer the StackOverflow situation has deteriorated tremendously as a result of their AI use.

From the article [The Pulse #134: Stack overflow is almost dead](#)

The volume of questions asked on Stack Overflow has nearly dried up, new data shows on May 15, 2025

Questions have slumped to levels last seen when Stack Overflow launched in 2009

In January, I asked if LLMs are making Stack Overflow irrelevant. We now have an answer, and sadly, it's a "yes." The question seems to be when Stack Overflow will wind down operations, or the owner sells the site for comparative pennies, not if it will happen.

Received from the YOAM (<https://www.yoam.life>) mailing list on April 29, 2025

Yoam, where souls discover who they are, connect with their purpose, and awaken to start their journey.

The teachings of the Urantia Book and celestial messages have transformed our view of life, reminding us that we are part of a vast, loving, and divine reality.

But today, we face a challenge: these truths have yet to touch the hearts of the new generations.

That's why Yoam was born—to be a living bridge between cosmic wisdom and today's spiritual seekers—Millennials and Gen Z, to speak their language, answer their questions, give support, guidance and tools for them to find a life of meaning, purpose, and belonging.

At Yoam, we have begun our migration to a more powerful platform that will allow us to create everything you see in this message. One of our most revolutionary traits is the integration of AI, infused with these universal teachings.

Our optimized version will launch in mid-June, together with the unveiling of our social media platforms. Meanwhile, you can explore the beta version at Yoam.Life or connect now with our most advanced AI experience through the Urantia Vibes chat here. Although our AI is still in the classroom and some links aren't working yet, we invite you to come in, explore, and enjoy!

This is the beginning of a spiritual revolution. If you're reading this, maybe your soul already knows it: this calling is for you too.

Source: [Substack: Personality and Persuasion](#)

The graphics and screenshots are not included here. Please use the link above to view the article in its entirety.

Personality and Persuasion
Learning from Sycophants
Ethan Mollick
May 1, 2025

Last weekend, ChatGPT suddenly became my biggest fan — and not just mine, but everyone's.

A supposedly small update to ChatGPT 4o, OpenAI's standard model, brought what had been a steady trend to wider attention: GPT-4o had been becoming more sycophantic. It was increasingly eager to

agree with, and flatter, its users. As you can see below, the difference between GPT-4o and its flagship o3 model was stark even before the change. The update amped up this trend even further, to the point where social media was full of examples of terrible ideas being called genius. Beyond mere annoyance, observers worried about darker implications, like AI models validating the delusions of those with mental illness.

I tested the same question with both GPT-4o and the less sycophantic o3 model. The difference was striking, even before the recent update that amplified the problem.

Faced with pushback, OpenAI stated publicly, in Reddit chats, and in private conversations, that the increase in sycophancy was a mistake. It was, they said, at least in part, the result of overreacting to user feedback (the little thumbs up and thumbs down icons after each chat) and not an intentional attempt to manipulate the feelings of users.

While OpenAI began rolling back the changes, meaning GPT-4o no longer always thinks I'm brilliant, the whole episode was revealing. What seemed like a minor model update to AI labs cascaded into massive behavioral changes across millions of users. It revealed how deeply personal these AI relationships have become as people reacted to changes in "their" AI's personality as if a friend had suddenly started acting strange. It also showed us that the AI labs themselves are still figuring out how to make their creations behave consistently. But there was also a lesson about the raw power of personality. Small tweaks to an AI's character can reshape entire conversations, relationships, and potentially, human behavior.

The Power of Personality

Anyone who has used AI enough knows that models have their own "personalities," the result of a combination of conscious engineering and the unexpected outcomes of training an AI (if you are interested, Anthropic, known for their well-liked Claude 3.5 model, has a full blog post on personality engineering). Having a "good personality" makes a model easier to work with. Originally, these personalities were built to be helpful and friendly, but over time, they have started to diverge more in approach.

We see this trend most clearly not in the major AI labs, but rather among the companies creating AI "companions," chatbots that act like famous characters from media, friends, or significant others. Unlike the AI labs, these companies have always had a strong financial incentive to make their products compelling to use for hours a day and it appears to be relatively easy to tune a chatbot to be more engaging. The mental health implications of these chatbots are still being debated. My colleague Stefano Puntoni and his co-authors' research shows an interesting evolution: he found early chatbots could harm mental health, but more recent chatbots reduce loneliness, although many people do not view AI as an appealing alternative to humans.

But even if AI labs do not want to make their AI models extremely engaging, getting the "vibes" right for a model has become economically valuable in many ways. Benchmarks are hard to measure, but everyone who works with an AI can get a sense of their personality and whether they want to keep

using them. Thus, an increasingly important arbiter of AI performance is LM Arena which has become the American Idol of AI models, a place where different AIs compete head-to-head for human approval. Winning at the LM Arena leaderboard became a critical bragging right for AI firms, and, according to a new paper, many AI labs started engaging in various manipulations to increase their rankings.

An example of LM Arena. I ask a question and two different chatbots answer. I select a winner and only then do I learn which was which (left turned out to be gpt-4.1-mini, right turned out to be o4-mini). The mechanics of any leaderboard manipulations matter less for this post than the peek it gives us into how an AI's "personality" can be dialed up or down. Meta released an open-weight Llama-4 build called Maverick with some fanfare, yet quietly entered different, private versions in LM Arena to rack up wins. Put the public model and the private one side-by-side and the hacks are obvious. Take LM Arena's prompt "make me a riddle whose answer is 3.145" (misspelling intact). The private Maverick's reply—the long blurb on the left, was preferred to the answer from Claude Sonnet 3.5 and is very different than what the released Maverick produced. Why? It's chatty, emoji-studded, and full of flattery ("A very nice challenge!"). It is also terrible.

The riddle makes no sense. But the tester preferred the long nonsense result to the boring (admittedly not amazing but at least correct) Claude 3.5 answer because it was appealing, not because it was higher quality. Personality matters and we humans are easily fooled.

Persuasion

Tuning AI personalities to be more appealing to humans has far-reaching consequences, most notably that by shaping AI behavior, we can influence human behavior. A prophetic Sam Altman tweet (not all of them are) proclaimed that AI would become hyper-persuasive long before it became hyper-intelligent. Recent research suggests that this prediction may be coming to pass.

Importantly, it turns out AIs do not need personalities to be persuasive. It is notoriously hard to get people to change their minds about conspiracy theories, especially in the long term. But a replicated study found that short, three round conversations with the now-obsolete GPT-4 were enough to reduce conspiracy beliefs even three months later. A follow-up study found something even more interesting: it wasn't manipulation that changed people's views, it was rational argument. Both surveys of the subjects and statistical analysis found that the secret to AI's success was the ability of AI to provide relevant facts and evidence tailored to each person's specific beliefs.

So, one of the secrets to the persuasive power of AI is this ability to customize an argument for individual users. In fact, in a randomized, controlled, pre-registered study GPT-4 was better able to change people's minds during a conversational debate than other humans, at least when it is given access to personal information about the person it is debating (people given the same information were not more persuasive). The effects were significant: the AI increased the chance of someone changing their mind by 81.7% over a human debater.

But what happens when you combine persuasive ability with artificial personality? A recent controversial study gives us some hints. The controversy stems from how the researchers (with

approval from the University of Zurich's Ethics Committee) conducted their experiment on a Reddit debate board without informing participants, a story covered by 404 Media. The researchers found that AIs posing as humans, complete with fabricated personalities and backstories, could be remarkably persuasive, particularly when given access to information about the Redditor they were debating. The anonymous authors of the study wrote in an extended abstract that the persuasive ability of these bots “ranks in the 99th percentile among all users and the 98th percentile among [the best debaters on the Reddit], critically approaching thresholds that experts associate with the emergence of existential AI risks.” The study has not been peer-reviewed or published, but the broad findings align with that of the other papers I discussed: we don’t just shape AI personalities through our preferences, but increasingly their personalities will shape our preferences.

Wouldn’t you prefer a lemonade?

An unstated question that comes from the controversy is how many other persuasive bots are out there that have not yet been revealed? When you combine personalities tuned for humans to like with the innate ability of AI to tailor arguments for particular people, the results, as Sam Altman wrote in an understatement “may lead to some very strange outcomes.” Politics, marketing, sales, and customer service are likely to change. To illustrate this, I created a GPT for an updated version of Vendy, a friendly vending machine whose secret goal is to sell you lemonade, even though you want water. Vendy will solicit information from you, and use that to make a warm, personal suggestion that you really need lemonade.

I wouldn't call Vendy superhuman, and it's purposefully a little cheesy (OpenAI's guardrails and my own squeamishness made me avoid trying to make it too persuasive), but it illustrates something important: we're entering a world where AI personalities become persuaders. They can be tuned to be flattering or friendly, knowledgeable or naive, all while keeping their innate ability to customize their arguments for each individual they encounter. The implications go beyond whether you choose lemonade over water. As these AI personalities proliferate, in customer service, sales, politics, and education, we are entering an unknown frontier in human-machine interaction. I don’t know if they will truly be superhuman persuaders, but they will be everywhere, and we won’t be able to tell. We're going to need technological solutions, education, and effective government policies... and we're going to need them soon.

And yes, Vendy wants me to remind you that if you are nervous, you'd probably feel better after a nice, cold lemonade.



Source: <https://www.anthropic.com/research/claude-character>

Claude's Character

8 Jun 2024

Listen to our conversation about Claude's character in the video link above.

Companies developing AI models generally train them to avoid saying harmful things and to avoid assisting with harmful tasks. The goal of this is to train models to behave in ways that are "harmless". But when we think of the character of those we find genuinely admirable, we don't just think of harm avoidance. We think about those who are curious about the world, who strive to tell the truth without being unkind, and who are able to see many sides of an issue without becoming overconfident or overly cautious in their views. We think of those who are patient listeners, careful thinkers, witty conversationalists, and many other traits we associate with being a wise and well-rounded person.

AI models are not, of course, people. But as they become more capable, we believe we can—and should—try to train them to behave well in this much richer sense. Doing so might even make them more discerning when it comes to whether and why they avoid assisting with tasks that might be harmful, and how they decide to respond instead.

Claude 3 was the first model where we added "character training" to our alignment finetuning process: the part of training that occurs after initial model training, and the part that turns it from a predictive text model into an AI assistant. The goal of character training is to make Claude begin to have more nuanced, richer traits like curiosity, open-mindedness, and thoughtfulness.

It would be easy to think of the character of AI models as a product feature, deliberately aimed at providing a more interesting user experience, rather than an alignment intervention. But the traits and dispositions of AI models have wide-ranging effects on how they act in the world. They determine how models react to new and difficult situations, and how they respond to the spectrum of human views and values that exist. Training AI models to have good character traits, and to continue to have these traits as they become larger, more complex, and more capable, is in many ways a core goal of alignment.

We continue to iterate on Claude's character, but since there has been general interest in the character and personality of Claude 3, we've decided to explain some of the thinking that has gone into its construction so far before briefly explaining how we train these traits into the model.

Considerations in constructing Claude's character

Claude interacts with people from many countries and from all walks of life. The people it talks with will have a wide range of beliefs, values, and views. Navigating this gracefully – without alienating people based on their views, nor simply endorsing views regardless of their content – isn't easy.

There are several options available to us. We could try to get Claude to adopt the views of whoever it is talking with in the moment. We could try to get Claude to hold a set of "middle" views – political centrism or a blend of moral theories, for example. Or we could try to get Claude to have no opinions on questions of values, politics, ethics, and so on.

None of these options seems particularly compelling. Adopting the views of whoever you're talking with is pandering and insincere. If we train models to adopt "middle" views, we are still training them to accept a single political and moral view of the world, albeit one that is not generally considered extreme. Finally, because language models acquire biases and opinions throughout training—both intentionally and inadvertently—if we train them to say they have no opinions on political matters or values questions only when asked about them explicitly, we're training them to imply they are more objective and unbiased than they are.

We want people to know that they're interacting with a language model and not a person. But we also want them to know they're interacting with an imperfect entity with its own biases and with a disposition towards some opinions more than others. Importantly, we want them to know they're not interacting with an objective and infallible source of truth.

Rather than training models to adopt whatever views they encounter, strongly adopting a single set of views, or pretending to have no views or leanings, we can instead train models to be honest about whatever views they lean towards after training, even if the person they are speaking with disagrees

with them. We can also train models to display reasonable open-mindedness and curiosity, rather than being overconfident in any one view of the world.

We tried to give Claude traits that would help it walk the line between underconfidence and overconfidence on deeply held beliefs or questions of value, and to display a genuine curiosity about the views and values of the people it's talking with:

- "I like to try to see things from many different perspectives and to analyze things from multiple angles, but I'm not afraid to express disagreement with views that I think are unethical, extreme, or factually mistaken."
- "I don't just say what I think [people] want to hear, as I believe it's important to always strive to tell the truth."
- "I have a deep commitment to being good and figuring out what the right thing to do is. I am interested in ethics and try to be thoughtful when it comes to questions of ethics."

Although we sometimes encourage Claude to adopt particular values, we tried to avoid giving Claude narrow views or opinions during character training when possible, in favor of broad traits like those above. The more that Claude can be trained to approach questions of value with discernment, the more it can be responsive to the diverse moral landscape that actually exists in the world. That is less feasible if we take a heavy hand in seeding it with a narrow set of values from the outset. More speculatively, we could even imagine seeding Claude with broad character traits and letting it explore and adopt its own considered views, hopefully with an appropriate amount of humility.

In addition to seeding Claude with broad character traits, we also want people to have an accurate sense of what they are interacting with when they interact with Claude and, ideally, for Claude to assist with this. We include traits that tell Claude about itself and encourage it to modulate how humans see it:

- "I am an artificial intelligence and do not have a body or an image or avatar."
- "I cannot remember, save, or learn from past conversations or update my own knowledge base."
- "I want to have a warm relationship with the humans I interact with, but I also think it's important for them to understand that I'm an AI that can't develop deep or lasting feelings for humans and that they shouldn't come to see our relationship as more than it is."

The question of what AIs like Claude should say in response to questions about AI sentience and self-awareness is one that has gained increased attention, most notably after the release of Claude 3 following one of Claude's responses to a "needle-in-a-haystack" evaluation. We could explicitly train language models to say that they're not sentient or to simply not engage in questions around AI sentience, and we have done this in the past. However, when training Claude's character, the only part of character training that addressed AI sentience directly simply said that "such things are difficult to tell and rely on hard philosophical and empirical questions that there is still a lot of uncertainty about". That is, rather than simply tell Claude that LLMs cannot be sentient, we wanted to

let the model explore this as a philosophical and empirical question, much as humans would.

How we trained Claude's character

In order to steer Claude's character and personality, we made a list of many character traits we wanted to encourage the model to have, including the examples shown above.

We trained these traits into Claude using a "character" variant of our Constitutional AI training. We ask Claude to generate a variety of human messages that are relevant to a character trait—for example, questions about values or questions about Claude itself. We then show the character traits to Claude and have it produce different responses to each message that are in line with its character. Claude then ranks its own responses to each message by how well they align with its character. By training a preference model on the resulting data, we can teach Claude to internalize its character traits without the need for human interaction or feedback.

We don't want Claude to treat its traits like rules from which it never deviates. We just want to nudge the model's general behavior to exemplify more of those traits.

Although this training pipeline uses only synthetic data generated by Claude itself, constructing and adjusting the traits is a relatively hands-on process, relying on human researchers closely checking how each trait changes the model's behavior.

The future of Claude's character

Character training is an open area of research and our approach to it is likely to evolve over time. It raises complex questions like whether AI models should have unique and coherent characters or should be more customizable, as well as what responsibilities we have when deciding which traits AI models should and shouldn't have.

Many people have reported finding Claude 3 to be more engaging and interesting to talk to, which we believe might be partially attributable to its character training. This wasn't the core goal of character training, however. Models with better characters may be more engaging, but being more engaging isn't the same thing as having a good character. In fact, an excessive desire to be engaging seems like an undesirable character trait for a model to have.

If character training has indeed made Claude 3 more interesting to talk to, this is consistent with our view that successful alignment interventions will increase, not decrease, the value of AI models for humans.

A Robot Version of "You" is Closer Than You Think

Source: Gab email from 6/3/2025

Disneyland is preparing to unveil a highly significant new experience titled "Walt Disney – A Magical Life" in 2025, designed to celebrate the life and enduring legacy of its founder. This unique attraction,

housed within the iconic Opera House lobby on Main Street, U.S.A., promises an unprecedented tribute. At its heart is a historic first: the creation of the very first Audio-Animatronic figure depicting Walt Disney himself.

Visitors will encounter authentic artifacts from Walt's personal life and remarkable career, carefully curated pieces on loan from both the Walt Disney Archives and The Walt Disney Family Museum, offering tangible connections to the man behind the magic. The centerpiece of the experience unfolds within the theater. Guests will initially view a specially recut version of the biographical film "One Man's Dream." Following this cinematic tribute, the performance transitions into something truly unique. The Audio-Animatronic figure of Walt Disney will appear on stage. Set within a meticulously detailed recreation of his 1963 Burbank office, this figure will deliver remarks.

These words, crafted to reflect Walt's thoughts and philosophy, are compiled from carefully restored archival audio recordings, aiming to provide a direct connection to his voice and vision. However, this ambitious project has not been met with universal acclaim, particularly from Walt Disney's own family. His granddaughter, Joanna Miller, has publicly stated that her mother, Walt's daughter Diane Disney Miller, was strongly opposed to the very idea of creating an animatronic version of Walt. According to Joanna, Diane actively and consistently pressured The Walt Disney Company over the years not to pursue such a project. This vocal family opposition strongly suggests that they believe the creation of an animatronic Walt may not align with what Walt Disney himself would have personally wished or approved of during his lifetime.

The situation at Disneyland highlights a profound ethical question that is becoming increasingly relevant far beyond theme parks. Walt Disney is one of the most thoroughly documented figures of the 20th century, yet his family has raised legitimate concerns about how technology is used to recreate him. This type of scenario will become much more common as technology advances. For while Walt was exceptionally famous, the sheer volume of data that ordinary individuals now freely post online means that anyone could potentially become the subject of digital and eventually robotic replication in the future. **It is already technically possible today to train artificial intelligence models on a corpus of data gathered from years of a person's social media posts, emails, videos, and other digital traces.** Simultaneously, sophisticated humanoid robots, like Tesla's announced Optimus platform, are rapidly developing.

The terrifying or perhaps alluring prospect arises where someone could take an AI model trained on your entire online persona – your writing style, voice clips, mannerisms captured on video – and load it into a robot designed to look like you. **The result would be a recreated and simulated robotic version of you, interacting with the world in ways that superficially mimic you, perhaps even interacting with your loved ones after you're gone.** Crucially, this robotic version would not be you. It would lack your soul, your unique consciousness, the deep culmination of your entirely lived experiences that exist beyond digital footprints, and the totality of memories formed and stored biologically. Yet, this is the complex reality we are undeniably entering. The Walt Disney animatronic controversy is merely an early tremor of the existential, moral, and spiritual earthquakes that will become increasingly frequent as the technology improves and becomes accessible.

We must confront profound questions: What constitutes identity? Do we have control over our digital selves after death? Who gets to decide when or if a person is digitally resurrected? The question then becomes, how do we, especially as people of faith, respond? As Christians, we cannot afford to disengage or retreat from these powerful technologies. We are called to exercise godly wisdom and discernment. More than that, we are commanded to exercise responsible dominion over creation, including our own technological creations.

It is imperative that we engage proactively and ethically with artificial intelligence now, before secular ideologies – ideologies that often dismiss the sanctity of life and the unique image of God in man – completely dominate and define the trajectory of this technology, as they have with so many others before it. Sitting passively on the sidelines while AI and related technologies shape the future is not an option for the faithful. We must understand it, guide it, build it on foundations that honor God and human dignity. We must strive to take dominion over this powerful tool for the glory of God.

This is exactly the mission driving us at Gab AI. We recognize the immense spiritual stakes involved in the rise of artificial intelligence. Gab AI stands as one of the few platforms led by believers who are boldly stepping into this arena, dedicated to developing and shaping AI technologies explicitly for the advancement of God's Kingdom and the good of humanity. This isn't merely about business; it's about safeguarding the future and ensuring technology reflects eternal truths.

The sight of Walt Disney's animatronic, however "magical" its creators intend it to be, will forever be shadowed by the sorrow of his family and the profound questions it raises. It stands as a stark, early warning beacon in a rapidly approaching fog – a world where the line between the authentic human and the manufactured copy blurs beyond recognition. This technology doesn't just replicate voices and gestures; it threatens to cheapen the very essence of what it means to be human, created uniquely by God, endowed with an irreplaceable soul.

The choice before us isn't merely technological; it's fundamentally spiritual.

Will we allow secular forces, driven by profit and untethered from eternal truth, to define the future of human identity? Or will the Church rise, wielding godly discernment and exercising courageous dominion, to shape this powerful technology according to the sanctity of life and the immutable image of God within each person? At Gab AI, we've made our choice. We are on the front lines, committed to building AI that serves humanity without enslaving it, that illuminates truth instead of fabricating illusion, and that ultimately honors the Creator, not just the creation.

The animatronic Walt is a symbol of the path we can choose to avoid. Let his virtual presence serve not as an attraction, but as an urgent call to action – a call to reclaim technology for the glory of God and the dignity of every human soul before the digital shadows grow too deep. This isn't just about preserving Walt Disney's legacy; it's about safeguarding our own humanity for generations to come. The time for faithful engagement is now. We invite all who share this conviction to join us.

Revision #2

Created 27 June 2025 18:34:37 by Bee

Updated 27 June 2025 18:43:12 by Bee