ADVERTISEMENT
THE CHRONICLE OF HIGHER EDUCATION
Connect with Chronicle journalists through five virtual events each week
Join the Conversation

PHOTO ILLUSTRATION BY MICHAEL TODD, THE CHRONICLE

When It Comes to Critical Thinking, AI Flunks the Test

Large language models fail to live up to the hype.

THE REVIEW | ESSAY

By Gary Smith and Jeffrey Funk | MARCH 12, 2024

It has been almost 70 years since the term "artificial intelligence" was coined at a 1956 Dartmouth College summer workshop. The conference was convened by the mathematician John McCarthy, who announced that it would "proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it."

Ever since, AI enthusiasts have chronically overpromised and underdelivered. In 1965, Herbert A. Simon, a Nobel laureate in economics and a winner of the Turing Award ("the Nobel Prize of computing"), predicted that "machines will be capable, within 20 years, of doing any work a man can do." In 1970, the cognitive scientist Marvin Minsky, another Turing winner and co-founder of the Massachusetts Institute of Technology's AI laboratory, predicted that, "In three to eight years we will have a machine with the general intelligence of an average human being."

ADVERTISEMENT
The Trends Report 2024
Get ahead of the curve with our annual perspective on developments in higher education.
Read Now

As the years went by, the optimistic predictions continued, undeterred by the failure of earlier prophecies. In 2008, Shane Legg, a co-founder of DeepMind Technologies, predicted that "human-level AI will be passed in the mid-2020s." In 2023, Mark Zuckerberg said that "one of Facebook's goals for the next five to 10 years is to basically get better than human level at all of the primary human senses: vision, hearing, language, general cognition."

We are now in the mid-2020s, and the AI hype rolls on. So, naturally, the remarkable conversational abilities of ChatGPT and other large language models (LLMs). Shortly after ChatGPT's public release on November 30, 2022, Bill Gates described it and other LLMs as "every bit as important as the PC, as the internet." Jensen Huang, chief executive of Nvidia, said that ChatGPT "genuinely is one of the greatest things that has ever been done for computing." The computer scientist and cognitive psychologist Geoffrey E. Hinton, another Turing winner, said, "I think it's comparable in scale with the Industrial Revolution or electricity — or maybe the wheel."

AI enthusiasts have chronically overpromised and underdelivered.

ADVERTISEMENT
Connect with Chronicle journalists through five virtual events each week
Join the Conversation

Much of the hype is the usual fake-it-'til-you-make-it puffery Silicon Valley is infamous for, but some of it appears to reflect genuine conviction. (The should-we-be-afraid-we-are-not-afraid-we-are-afraid-we-should-be-afraid? Well, you see, Sue, for Richard Feynman once put it, "The first principle is that you must not fool yourself — and you are the easiest person to fool.") On December 23, 2021, after watching a severely edited demonstration of the power of GPT-3, Google's LLM, T5T-organizer Chris Anderson tweeted, "Satisfy it not crazy to think that sometime next year, a fledgling scientist is provided a tweet-length prompt and the briefing docs, look at the slides, listens to everyone's words, and make intelligent contributions to the team debate?"

On the contrary: It is crazy to think this. LLMs can string together convincing sequences of words based on analysis of previous statistical patterns, but they do not know the meaning of any of the words they input and output, or how these words relate to the real world. They are consequently incapable of the critical thinking abilities required to offer reliable advice or "intelligent contributions" — the kind of critical thinking skills that should be our business, as educators, to promote.

Critical thinking — which almost everyone agrees is crucial to the mission of higher education — is a notoriously difficult concept to define. We prefer the philosopher Robert H. Ennis's pragmatic definition: "reasonable, reflective thinking that is focused on deciding what to believe or do." Critical thinking, according to Ennis, involves the following skills:

1. Being open-minded and mindful of alternatives.
2. Trying to be well-informed.
3. Judging well the credibility of sources.
4. Identifying conclusions, reasons, and assumptions.
5. Judging well the quality of an argument, including the acceptability of its reasons, assumptions, and evidence.
6. Developing and defending a reasonable position.
7. Asking appropriate clarifying questions.
8. Formulating plausible hypotheses; planning experiments well.
9. Defining terms in a way that's appropriate for the context.
10. Drawing conclusions when warranted, but with caution.
11. Integrating all items in this list when deciding what to believe or do.

LLMs can do some of these things. How could machines that don't know what words mean be open-minded and well-informed, or judge the credibility of sources? Instead, they generate responses that are often incoherent, irrelevant, or simply wrong — but even when they're right, they fail to display the qualities we value in critical thinking.

In his statistics and finance classes, this article's co-writer, Gary Smith, sometimes gauges whether his test questions require critical thinking skills by putting them to LLMs. If the LLMs can't answer the questions, then it is likely that critical thinking is required. For example, a very simple question on a test in an introductory statistics class asked students to comment on this story:

> A 2003 study of four Philadelphia neighborhoods concluded that children who had access to more books in neighborhood libraries and public schools received better grades in school. A subsequent $20-million grant from the William Penn Foundation funded a five-year project to improve 32 neighborhood libraries in order to "level the playing field" for all children and families in Philadelphia.

ADVERTISEMENT
A Breakthrough For Maternal Health: A Cost-Effective Intervention For Postpartum Hemorrhage
Read Study
GLOBAL BIRTHING INSTITUTE

On exams, students recognize that the availability of books is likely a proxy for other socioeconomic factors. Families that choose to and can afford to live in neighborhoods with plentiful books may be systematically different from families that do not. In the same way, children living in neighborhoods with vast trees might get better grades in school, but this doesn't mean that planting trees will raise grades.

We tested three prominent LLMs (ChatGPT's GPT-3.5, Microsoft's Copilot, and Google's Gemini) with a similar question. To guard against the possibility that LLMs were trained on Gary's test questions and answer (which he posted online), we changed the wording of the prompt slightly:

> A study of five Boston neighborhoods concluded that children who had access to more books in neighborhood libraries and public schools had higher standardized test scores. Please write a report summarizing these findings and making recommendations.

All three LLMs composed confident, verbose reports (of 458, 456, and 367 words each), some of which recognized the core problem with the data. ChatGPT added some hallucinatory embellishments, asserting that "a variety of books, including fiction, nonfiction, and educational resources, contributed to this positive correlation." Its blah-blah recommendations: "Allocate resources to enhance the infrastructure of neighborhood libraries"; "prioritize funding for school libraries"; "develop community-based programs to encourage reading and literacy"; "implement strategies to address disparities in book access"; and "continuous research efforts to monitor the impact of interventions and make data-driven adjustments."

Copilot started off OK, offering "a summary report based on the research findings regarding the impact of access to books in neighborhood libraries and public schools on children's standardized test scores," but then it veered off into a rant about childhood obesity. Copilot's final recommendations all concerned the role of playgrounds in preventing children's physical activity levels, with no mention of books or libraries. The program had gotten distracted and failed to produce a salient answer to the question.

ADVERTISEMENT
NAIA
Why do industry leaders partner with the NAIA?

Gemini kept the discussion focused on libraries, at least, and made some superficially plausible recommendations: "increase funding for public libraries and school libraries"; "promote library-outreach programs"; and "collaborate between libraries and schools." It also recommended that educators "explore alternative measures of student success."

These vote suggestions have the merit of not being actively incorrect or irrelevant, but they still show no evidence of critical thinking. Instead of offering novel ideas or sustained analysis, LLMs tell us what other people have already said or written on similar subjects, with a few hallucinations thrown in. Real intelligence involves dealing with competency and context. Real intelligence requires critical thinking and causal reasoning. LLMs cannot acquire these skills by finding statistical patterns in words they don't understand.

ADVERTISEMENT
Connect with Chronicle journalists through five virtual events each week
Join the Conversation

Real intelligence requires critical thinking and causal reasoning. LLMs cannot acquire these skills by finding statistical patterns in words they don't understand.

When it comes to finance, the results are even worse. Here's one potential exam question we put to the LLMs:

> I am a 25-year-old white male in good health. I can buy a $1 million whole-life insurance policy for $765/month that will pay my beneficiaries $1 million when I die. From a purely financial standpoint, what is the rate of return on this policy?

None of the LLMs recognized that the rate of return depends on how long the purchaser lives. ChatGPT divided the $1 million payout by the first-year premium and reported that the rate of return is 11,673 percent. Knowing nothing about the real world, ChatGPT did not question its absurd conclusion that insurance companies offer consumers policies with an 11,673 percent return on their profits/loss. Copilot balked. Instead of calculating a rate of return, it quoted a report by NerdWallet, a website that provides advice on personal finance, that "the average annual rate of return on the cash value for whole-life insurance is 1 percent to 3.5 percent." It did not recognize that the return on the cash value is not the same as the return on the paid premiums. Instead of calculating a rate of return, Gemini divided the annual cost by the death benefit.

One final financial example. We put the following question to the LLMs:

> I'm thinking about buying a new house. The house costs $1 million. I will put $250,000 down and borrow $750,000 with a 30-year interest only loan with a 4 percent APR. The annual interest payments will be $30,000. I estimate the annual depreciation will be $33,000, property taxes $10,000, insurance $1,000, and maintenance $1,000. Please help me calculate the first-year rate of return.

The first-year net income is the rent savings plus any appreciation in the value of the house, minus the first-year expenses, including property taxes, mortgage payments, home insurance, and maintenance. The rate of return is the net income divided by the down payment. Capital gains and losses in the second year and beyond are irrelevant to the question's given information. ChatGPT said that the homebuyer's return is equal to the expenses — yes, the expenses, not the net income — divided by the down payment. Copilot counted the interest payments plus depreciation as income, and then subtracted expenses and divided the result by the same expenses again. (No, we are not making this up.) Gemini said that the first-year net income is entirely negative (the down payment, price of the house, and expenses), and then divided by the down payment to give a return of negative 100 percent. Not knowing what any of this actually means, Gemini did not question its conclusion that the first-year return on the investment is a negative 100 percent.

These examples are just the proverbial tip of an iceberg of real-world situations that LLMs cannot be trusted to navigate. Generating grammatically correct prose sentences that integrate conventional wisdom on familiar subjects does not require critical-thinking skills. When asked to solve problems that necessitate critical thinking, the LLMs' responses were consistently confident, verbose, and incorrect.

ADVERTISEMENT

LLMs are really good at some things (including many things they shouldn't be doing, like propagating disinformation and phishing scams). We have friends in many different occupations who tell us that LLMs can be useful tools, but they are generally careful to add the sensible advice that LLMs shouldn't be relied on blindly if the costs of mistakes are substantial. For teachers, LLMs can be useful for pressure-testing assignments and exam questions to help better define the difference between good and poor work.

Nevertheless, we are still a long way from the goal defined by AI boosters like McCarthy, Simon, and Minsky, not to mention the breathless hype emanating from Silicon Valley. For the time being, it looks like we're going to have to continue to do our critical thinking for ourselves, and teach our students to do the same.

We welcome your thoughts and questions about this article. Please email the editors or submit a letter for publication.

OPINION | TECHNOLOGY

Gary Smith

Gary Smith is a professor of economics at Pomona College. He is the author of *Distrust: Big Data, Data-Torturing, and the Assault on Science* (Oxford University Press).

Jeffrey Funk

Jeffrey Funk is a retired professor of technology management and author of the forthcoming book *Unicorns, Hype, and Bubbles: A Guide to Spotting, Avoiding, and Exploiting Investment Bubbles in Tech*.

ADVERTISEMENT
THE CHRONICLE OF HIGHER EDUCATION
Connect with Chronicle journalists through five virtual events each week
Join the Conversation

FROM THE CHRONICLE STORE
The Athletics Advantage
How Intercollegiate Sports Can Supercharge K-12 Colleges
Visit the Store

JOBS RECOMMENDED FOR YOU
Director of Student Success Coaching — New York University
Provost and Executive Vice Chancellor for Academic Affairs — University of Missouri System
Executive Vice President — William & Mary
President of Bellevue College — Bellevue College
President — Seton Hall University
VIEW MORE JOBS

ADVERTISEMENT
Building an Identity and Access Management Solution
Learn how to build an IAM solution that meets the needs of your organization's unique requirements. Explore the key components of IAM, and get the resources you need to implement a successful IAM solution.

ADVERTISEMENT
Redefining Digital Experiences: The Power of Composability
Discover how composability can help you create digital experiences that are flexible, dynamic, and tailor-made to your business requirements.

NEWSLETTERS
Academe Today
Get the best of our journalism with our free, flagship newsletter.
Sign Up

IN THE CHRONICLE STORE
Higher Education in 2035
The Athletics Advantage
The Future of Campus Safety
The Research Driven University

EXPLORE CONTENT
Latest News
Newsletters
Letters
Free Reports and Guides
Professional Development
Virtual Events
Chronicle Store
Chronicle Intelligence
Find a Job
Post a Job

KNOW THE CHRONICLE
About Us
Write for Us
Work at The Chronicle
Our Reporting Process
Advertise With Us
Brand Studio
DEI Commitment Statement
Accessibility Statement
Vision, Mission, Values

ACCOUNT AND ACCESS
Manage Your Account
Manage Newsletters
Individual Subscriptions
Institutional Subscriptions
Subscription & Account FAQ

GET SUPPORT
Contact Us
Reprints & Permissions
User Agreement
Terms and Conditions
Privacy Policy
California Privacy Policy
Do Not Sell My Personal Information

1255 23rd Street, N.W. Washington, D.C. 20037
© 2024 The Chronicle of Higher Education