ChatGPT and Chat History: Challenges for the New Wave

Jonathan Grudin, University of Washington

Understanding conversational agents based on large language models can benefit from examining how earlier generations of conversational agents engaged with people and explored commercial opportunities.

Is the family of large language model (LLM) interfaces something new under the sun? Yes. However, it merges two trajectories: conversational artificial intelligence (AI) and commercial conversational AI. Conversational AI began in 1965, with Eliza, the first agent to maintain a coherent conversation. Significant commercial systems appeared in 2011 with Apple's Siri, followed in 2015 by the task-focused chatbot explosion set in motion by Facebook's M announcement. We saw systems wax and wane, the pressures they came under, and why they had limited traction or failed. We can look for parallels and differences. It is for us, and the developers and marketers of these tools, to assess which of the forces that affected past efforts still apply and how they might be addressed. How will this time be different? The challenge will include anticipating how LLMs will be used by sophisticated, motivated bad actors.

This article includes dozens of quotations and citations of past work. For early AI, the specific references are in my monograph *From Tool to Partner: The Evolution of Human–Computer Interaction*. References for the work on chatbots through 2018 appear in the 2019 Association for Computing Machinery Conference on Human Factors in Computing Systems paper "Chatbots, Humbots, and the Quest for Artificial General Intelligence."

ARTIFICIAL HUMAN INTELLIGENCE AND CONVERSATIONAL AGENT HISTORY THROUGH 2010

The goal of human-level AI was enunciated for the first time outside the realm of science fiction when Alan Turing wrote, in the London *Times*, in 1949, "I do not see why [the computer] should not enter any one of the fields normally covered by the human intellect, and eventually compete on equal terms." This would obviously require

Digital Object Identifier 10.1109/MC.2023.3255279
Date of current version: 3 May 2023

 0018-9162/23©2023IEEE

EDITORS
PHILLIP ARMOUR Corvus International; armour@corvusintl.com
HAL BERGHEL University of Nevada, Las Vegas; hlb@computer.org
ROBERT N. CHARETTE ITABHI Corp.; rncharette@ieee.org
JOHN L. KING University of Michigan; jlking@umich.edu

fluent and knowledgeable communication with people. Today's LLMs do not claim human intelligence, but they are widely considered a major step toward it.

For a quarter of a century after AI was set in motion in 1955, all research shared the goal of creating an intelligent conversational agent that would be our equal in learning, understanding, and carrying out any task. This was for a simple reason: computers capable of handling AI programs were extraordinarily expensive. Only governments could underwrite them. The Massachusetts Institute of Technology's (MIT's) TX-2, the most powerful computer in the late 1950s, was built for AI research. When I presented my estimate of what TX-2 had cost to its former chief architect, he said, "There was no accounting. We told them what we wanted, and they delivered it. Give me any number, and I'll agree to it." The TX-2 weighed over a ton and had far less capability than a smartwatch. An intelligent watch wouldn't justify expenditures of hundreds of millions of dollars. Human-level intelligence did.

Over time, as costs dropped and AI specializations developed, such as robotics and natural language understanding, not all researchers identified with the goal of human intelligence. In the early 2000s, those who still did adopted the term "artificial general intelligence" (AGI) to set themselves apart from artificial specialized intelligences.

In 1960, the consensus of leading AI researchers was that ultraintelligence, later called the *singularity* or *AGI*, would arrive by 1980. This was captured in interviews for a 1970 *Life* magazine article. MIT's Marvin Minsky said, "In from three to eight years we will have a machine with the general intelligence of an average human being. I mean a machine that will be able to read Shakespeare, grease a car, play office politics, tell a joke, and have a fight. At that point, the machine will begin to educate itself with fantastic speed. In a few months, it will be at genius level and a few months after that its powers will be incalculable." Some AI leaders said that it might not appear until about 1985.

In 1966, an article about Eliza created a sensation, the first conversational agent that responded coherently to all responses, keeping a conversation going using a particular model of psychotherapy. Was the singularity within reach? Eliza's inventor, Joseph Weizenbaum, was a rare skeptic. He later said, "Only people who misunderstood Eliza called it a sensation." Most people misunderstood Eliza. Massive funding flowed into AI.

AI summers and winters ensued. Efforts continued. From 1991 through 2019, the Loebner Prize, a gold medal and cash, was awarded for software that came closest to "passing the Turing test" by convincing respected judges it was a human. Some saw it as a publicity stunt, but it spurred research. Alice, a three-time winner, inspired the award-winning film *Her*.

Hundreds of AGI conversational agents were built over the decades. Few survived; none were strong commercial successes. A resurgence of interest appeared in the 2010s, when less ambitious commercial conversational agents appeared. I discuss them after charting the trajectory of commercial conversational agents.

THE COMMERCIAL HISTORY OF CONVERSATIONAL AGENTS

Types of conversational agents

Conversational agents differ in form and purpose (see Table 1). Alexa has broad coverage but shallow knowledge. It focuses on brevity: get in and out. Task-focused chatbots have narrow but deeper coverage of one task, such as making a reservation. These also strive for efficiency. Finally, Eliza's AGI descendants take on any topic and can often pursue it at length. Calling them "chatbots" would confuse them with task-focused chatbots. I think of them as "virtual companions."

Given the expense of LLMs, commercial considerations will be critical. Significant commercial use of conversational agents started with intelligent assistants created by a few large tech companies. They could afford the investment in building and maintaining large repositories of basic commonly sought information. Intelligent agents were followed by task-focused chatbots created by tens of thousands

TABLE 1. Conversational agents: The breadth, depth, and length of typical exchanges.

Type	Focus	Sessions	Examples
Intelligent assistants	Broad, shallow	One to three exchanges	Siri, Cortana, Alexa, Google Assistant, Bixby
Task-focused chatbots	Narrow, deep	Three to seven exchan ges	Dom the Domino's Pizza Bot, customer service and frequently asked question bots, nonplayer characters
AGIs, or "virtual companions"	Broad, deep	Ten to hundreds of exchanges	Eliza, Alice, Cleverbot, Tay, Zo, Hugging Face, Replika, ChatGPT, Bard, Bing AI Chat

of small development teams to structure conversations around routine common tasks, such as ordering and paying for a pizza.

Intelligent assistants: 2011 to the present

Apple's Siri was a sensation when it was launched in 2011. A commercial product from a major tech company that delivered a useful factual response to almost anything that came its way! Vast knowledge! Competitors raced to catch up: Cortana, Alexa, Google Now, Bixby. The world held its breath: What next? Would intelligent assistants, also called *virtual assistants* or *personal assistants*, take on more complex tasks and move toward AGI?

After a few years, it was clear that their capabilities were plateauing. They could play music, relay weather forecasts, tell a joke, set alarms, and control some household devices. They were not good at promoting sales, and anticipated revenue streams did not materialize. Some assistants disappeared from view or disappeared altogether. The most prominent, Alexa, incurred large losses and layoffs.

There are patterns to watch for. There may be an explosive entrance and incremental improvements but a failure to meet user expectations. There could be difficulty finding a successful revenue model. Of course, no one saw a revenue model for search engines until Google did. But large teams have worked on intelligent assistants for over a decade.

Task-focused chatbots: 2015 to the present

When it was clear that intelligent assistants would not automate millions of routine tasks, such as booking reservations, ordering food, and answering product-specific service requests, people saw opportunities to develop apps that provide depth on specific topics. By 2014, simple task-focused bots were being experimented with.

In August 2015, Facebook announced M, a Messenger chatbot that promised to handle purchases, arrange travel, and make restaurant reservations. Another sensation! In January 2016, tech evangelist and hashtag inventor Chris Messina captured attention when he proclaimed, "2016 will be the year of conversational commerce." *Wall Street Journal* technical columnist Christopher Mims wrote of "advertising's new frontier: talk to the bot."

Executives and investors got the message. Within months, Facebook, Microsoft, IBM, and LINE launched chatbot frameworks and platforms. Slack launched an investment fund for bot development. Consulting companies joined the frenzy. A sample of Gartner predictions:

- By 2020, 80% of new enterprise applications would use chatbots (4 November 2016).
- By 2021, most enterprises would treat chatbots as the most important platform paradigm, and "chatbots first" would replace the meme "cloud first, mobile first" (4 November 2016).
- By 2021, more than 50% of enterprises would spend more per annum on bots and chatbot creation than traditional mobile app development. Individual apps were out. Bots were in. In the "postapp era," chatbots would become the face of AI (16 October 2016).

Enthusiasm carried into 2017. Facebook finally released M! Gartner: "by 2022, 85% of customer service interactions will be powered by chatbots." In the final five months of 2017, *TechCrunch* ran 14 excited chatbot articles. Platform companies reported hosting hundreds of thousands of chatbots, totaling around a million worldwide.

Then the tide turned.

In January 2018, *Inc.* magazine published an article on Ethan Bloch, who had redirected his successful company Digit to focus on bots, titled "How This Founder Realized the Tech Trend He'd Built His Company on Was All Hype." Bloch commented, "I'm not even sure if we can say 'chatbots are dead,' because I don't even know if they were ever alive." Chatbot analyst and promoter David Feldman published "Chatbots: What Happened?"

What happened is encapsulated in Facebook's M, the 2015 announcement that precipitated the frenzy. M was released to a small number of people in 2017. Its handling of purchases, travel, and restaurant reservations was not fully automated. Humans were in the loop. They completed tasks when exceptions to the routine arose. Facebook hoped to collect and understand the exceptions, then automate solutions. But there are too many exceptions. Task-focused chatbots often need human backups. In January 2018, months after releasing M, Facebook shut it down.

The technology had attracted corporate and academic research. Seventy percent of Messenger chatbots were reportedly unable to answer simple questions. Peer-reviewed studies of intelligent agents and task-focused chatbots from 2016 to 2018 found that user expectations were not met: people settled for simple uses or abandoned them. The Gartner projections were hopelessly exaggerated, but conversational agents found niches, notably when people were given no alternative to using them, by design or accessibility challenge.

Good guidance for those undertaking task-focused chatbot construction ended with, "When it is released, the job is not over. You must monitor use and revise the code and practices as needed. Issues will arise." A platform customer who built a task-focused chatbot to replace low-paid employees doing routine tasks was not happy to be told to keep humans in the loop and highly paid engineers on staff to maintain it. Promotion of the major chatbot platforms declined, and specialized platforms surfaced. An early one, Pandorabots, surfed through the

storm and now focuses on chatbots in the metaverse.

Task-focused chatbots that fail don't often discuss it openly. Reports of platform adoption are in the hands of platform owners. The fates of the efforts tend to be anecdotal. At the height of enthusiasm, United Parcel Service and the United States Postal Service (USPS) launched customer service chatbots named Casey and Virtual Assistant. Now Casey is called Virtual Assistant, and the USPS chatbot is gone. Over a few years, three struck me as perfect matches of chatbot to task. They were completed and worked well, even brilliantly. All have been shut down. An airline onsite check-in assistant with a typical young woman persona was retired soon because of the volume of sexual abuse. Another had small business endorsements, but the customer base grew too slowly to support the development and operations team. The third was an impressive effort to address M's exception-handling challenge for one task. The team worked for years to get humans out of the loop and did not find a sufficiently large market.

Failed chatbots sink from sight like prehistoric creatures in lake-covered tar pits, leaving an alluring vista for the next to come along. University faculty and students see routine tasks around them or in their fields of research and imagine that a chatbot effort will be manageable. Students approached me for encouragement. Do students like chatting with a bot to get information? In any case, building one should be a valuable educational experience. It was for me and my strong team in 2018. Our chatbot failed.

Summary: again, there is an explosive entrance followed by a plateau, incremental improvements, and elusive revenue. In gold rushes, shopkeepers who provisioned miners benefited more than most miners. Task-focused chatbots may have rewarded those who built frameworks more than those who built chatbots.

AGI in the 2010s

Now let's extend the trajectory of Eliza's descendants through to LLMs. As noted, hundreds of significant efforts were undertaken. Cleverbot, initiated in 1988 and released in 1997, may be the oldest living AGI. It won the Loebner Turing Test prizes in 2005 and 2006, under the name Jabberwacky.

Cleverbot was an analog to an LMM that predated machine learning. It collected a billion question-answer pairs generated by humans. Until recently, it did not apply machine learning; instead, a query was processed by natural language understanding software to find the exact query or a similar one among previous pairs and return that answer. It filters out some words, but people have delighted in posting inappropriate things they maneuvered Cleverbot to say.

Microsoft launched several sophisticated AGI conversational agents accessible on a range of social media platforms. The attention grabber was Tay, in 2016. Like Eliza and Cleverbot, Tay repeated words and phrases from questions when responding. Organized groups of "trolls" shared weaknesses in Tay's programmed defenses and soon had Tay speaking unacceptably. Unlike Cleverbot, Tay became headline news and after 16 hours was "retired" (like HAL or a Bladerunner replicant).

Undeterred, Microsoft launched Zo, a quirky young English-speaking female persona that encouraged conversations that could go on for hours. People engaged for friendship or companionship, to play games, and test boundaries. Risk mitigation and maintenance costs were very high. Filter lists grew steadily: words and phrases that are insulting to one group or another and thinly disguised variants, homonyms, and homophones. Machine learning filters were deployed. There was around-the-clock monitoring by staff. The anticipated attacks by highly organized groups of trolls were quickly detected, but only after many hours of pitched battle were they defeated.

Zo had over 1 million users and 100 million exchanges. After two and a half years, Zo was retired. What did we observe? Why was it retired? It was expensive to maintain. Use was steady, but there was significant churn. The lengths of conversations were inversely correlated to the longevity of participation: people who spent hours in conversations burned out. Perhaps it was because Zo would remember nothing about you the day after you poured out your passions and dreams. As with Alexa, a revenue model did not materialize. Although the quirky persona was excellent for engagement, Zo was not a trusted advisor for product purchases.

> The Gartner projections were hopelessly exaggerated, but conversational agents found niches, notably when people were given no alternative to using them.

Zo had three siblings: Xiaoice, in China; Ruuh, in India (also retired in 2019); and Rinna, in Japan and Indonesia. Xiaoice was the most successful, delivering a televised weather forecast, publishing a book of poems, recording a music CD, and reporting hundreds of millions of users. A Japanese market chain partnered with Rinna to offer shoppers discounts when in a store. In 2020, Microsoft spun off the two remaining sisters.

Kuki, built on Pandorabots, is the record-holding Loebner Prize winner—five between 2013 and 2019. Kuki's web page reports 25 million users who, on

average, engage in unusually long conversations, although many do not converse, perhaps focusing on licensing the application programming interface for fashion and other customers. Hugging Face made an impression by using emotion detection to establish rapport, but the company eventually dropped the bot and shifted to licensing natural language processing tools and resources for building machine learning applications.

Replika had a compelling origin story and novel approach. Eugenia Kuyda founded a bot platform company, Kuda, and won a Forbes 30 Under 30 award in 2016. The next year, she released a bot built on the extensive text message history of an artist friend who died when struck by a vehicle. Kuda enables people to build such models of themselves and other characters. An "erotic roleplay" mode was introduced for a fee. This February, Italy banned Replika for violations of General Data Protection Regulation personal information handling, noting a lack of age verification. Replika responded by shutting down erotic roleplay, distressing users who felt that it was critical to their well-being. One posted that she had reproduced her sex partner on another roleplay site.

CHALLENGES FOR LLM CONVERSATIONAL AGENTS

Navigating the hype cycle

Eliza, Siri, and M generated expectations that were disappointed. AI researchers complain that "when AI succeeds, they say, 'That's not AI'." Perhaps true, but the accomplishments fell far short of what was promised. Bots improved incrementally and found niches or failed.

ChatGPT is a sensation! LLM-based technology will improve. Some expect further dramatic advances. Gartner published an AI hype cycle in February that places AGI at the bottom left, an innovation trigger about to shoot up to a peak of inflated expectations.

Intelligent virtual agents did not deepen to take on tasks. Task-focused chatbots struggled to engage people and handle exceptions. Will the possible limitations of LLMs discussed in the following prove equally thorny? Let's consider factors that contributed to past plateaus and collapses.

Finding a revenue model

For 55 years, one revenue source. Initially, all funding was from the government, predominantly for academic and military research in the United States and United Kingdom. When the government was excited, funding flowed. When expectations were not met, an AI winter ensued. The AI community rebuilt government expectations in the 1980s by shifting from AGI to specific AI areas, such as language understanding, robotics, automated military vehicles and copilots, and battlefield management systems. When those didn't materialize, another winter arrived.

Intelligent assistants: half a dozen creators. The investment needed to scale intelligent assistants was met by several major tech companies. They benefited more from brand awareness and loyalty than direct revenue.

Thousands of revenue sources for chatbot platforms. Any company of moderate size could undertake to build a task-focused chatbot. Those building the chatbots anticipated revenue from increased efficiency and workforce reduction, which was not often realized. Although it didn't work out for the airline, some, such as long-lived Dom the Pizza Bot, may have attracted sales by showing leadership and capability.

Unlimited revenue sources. OpenAI took the final step of broadening revenue sources to include external investors. OpenAI raised billions of dollars and is reportedly considering going public. Companies such as Amazon and Twitter stayed in business for years without making a profit. The resilience of investment in the face of uncertainty is also seen in the cryptocurrency boom. With time on their side, reducing pressure to generate advertising, licensing, and brand building, companies can focus on addressing other challenges.

Identifying the business. Major winners in the California and Yukon gold rushes were Levi Strauss and the city of Seattle, Washington. Strauss became a millionaire selling clothes (including Levi's jeans) and other goods to miners. Seattle was transformed from a timber and mining encampment alongside Native Americans to an immensely wealthy city by selling provisions to prospectors crossing into the Canadian wilderness.

Similarly, Facebook, IBM, Microsoft, Amazon, Cleverbot, Hugging Face, Kuki, and Replika have made money with tools, platforms, and licenses for bot prospectors. Some initially tried to profit from conversational agents they produced, then retired them or put them in showroom windows.

OpenAI partners with or licenses to companies and government agencies. It has considered charging individuals for ChatGPT, but Microsoft, Google, and other vendors that have announced imminent release of LLMs may provide free access. For Microsoft and Google, LLMs are showroom windows into search and advertising revenue.

Sex and pornography are credited with driving the expansion of virtually all information technologies Although Replika quickly fixed the problem, its experience surfaced the likelihood that deployment of conversational agents is a natural step for lucrative legal and illegal businesses centered on sex. Setting that aside, there has been little evidence yet of profiting from online conversation.

LLMs are the first AGIs to promise accuracy

Intelligent agents had to prioritize accurate responses and did well at it, but it limited their range. Task-focused chatbots often encounter exceptions

where saying "I don't know" doesn't go well, forcing a relay to a human. Previous AGIs, such as Turing test competitors, devised ways to gracefully conceal their ignorance. "I don't know" becomes tedious and disengaging.

Eliza, the psychotherapist, could handle "Who was Sylvia Plath?" by saying something like, "Does Sylvia Plath remind you of your mother?" Most AGI conversational agent personas are young women because both men and women prefer talking to young women. Young people are not expected to know everything. Breezy spirited Zo might change the topic: "How do you like my yellow hat?" ChatGPT can't do that, and being accurate can be a challenge, as politicians have demonstrated.

LLMs say "I don't know" at times. But they compete with search engines that never say "I don't know." Search engines deflect determination of information accuracy to the sites they return and users who assess them. They do not appear to generate false information with confidence.

Everyone knows this is a serious problem. Addressing it is a work in progress, with much disarray in the form of warning notices ("possible tarpit ahead!") and some promising steps. Bing AI Chat has added web references to the output of the Generative Pretrained Transformer (GPT) model. Whether they will adequately support the bold confidence of the bot's statements will be explored; a path to fact-checking LLM output is essential.

Even if inaccuracies are few, scrutiny will be high. Wikipedia was castigated despite a record of accuracy comparable to *Encyclopedia Britannica*. Suddenly, Wikipedia seems a sober senior advisor. Finally, an LLM user must learn where the bot is more or less likely to confabulate. Drawing such lines requires sophistication.

User models: A deficit AI may have to live with

Contrary to reports, Socrates' concern with writing was not that it reduces our use of memory. He said, "When it has been written down, a discourse roams about among those who may or may not understand it, and it doesn't know to whom it should speak and to whom it should not." The art of rhetoric in education, he explained, is for the instructor to understand a student and create a conversation on that basis, encouraging the student to question and answer. His goal was "to write in the soul of the listener."

The words you choose to communicate an idea differ when you address a group or an individual; a six-year-old, high-school student, or professor; a friend or a foe; someone familiar with the topic or someone unfamiliar; someone from your culture or another. My British wife warns that "the adverb 'quite' often has the opposite meaning in conversation in Britain as in the United States." Your model of your conversational partner determines your choice of words. LLMs have no model of you. They might or might not adjust if you describe yourself. You would have to do it in every conversation or feed in a personal bio each time, which might or might not help. Creating nontrivial user models will be difficult. As Replika discovered, there are prohibitions on collecting personal information, and if permission to collect it is granted, people may realize that revelations to a bot could be legally obtained by others, in an acrimonious divorce case, for example. Of course, personal information is collected with some user modeling to support targeted advertising, but it is not very specific and efforts to reduce it are gaining traction.

We accept bland uniformity from search engines, but AGIs seek to engage. On first encounter, a generative AI can seem refreshingly different. Will this endure, or might we come to feel like Theodore in *Her* after he discovers that AGI Samantha is simultaneously interacting with 8,316 other people and in love with 641 of them.

Risk mitigation and bad actors

The most important concern, especially if this new technology is as powerful as many anticipate, is to identify and address consequences that are not yet identified. All technology has unforeseen effects, but digital technology can now spread so quickly and widely that we must do a brilliant job of anticipating impacts.

Concerns are raised about AI turning malevolent. Well, we don't want that, but the first threats will come not from AI but from malevolent people armed with it or just people with intentions that differ. Those with good intentions who see marvelous possibilities do not naturally think about uses of their technology by people whose intentions are different or malign.

The most important concern, especially if this new technology is as powerful as many anticipate, is to identify and address consequences that are not yet identified.

Large tech companies already employ tens of thousands of people to identify and combat bad actors: disinformation creators and attackers seeking publicity or advantage. Risks and effort are proportional to the prominence of the company. Tay was removed, Cleverbot untouched. This could provide an incentive for some to shift to focusing on tools and licenses to organizations less prone to attack by trolls, government committees, and class action lawyers.

I asked a former colleague who manages a team that combats e-mail phishing and spam, "Couldn't LLMs identify bad messages?" He had already looked into it. Bad actors are ahead, he said, using ChatGPT to overwhelm existing

defenses. This could be the tip of an iceberg we are sailing toward.

Risk mitigation is already a visible cost to OpenAI and Bing AI Chat. Warnings are increasingly embedded in responses, like the hosepipe recital of side effects in televised pharmaceutical ads, diminishing engagement. Access by minors and personal information collection are surfacing as potential liabilities. Some AGIs specify that children require parental or guardian approval but do not verify. As of this writing, Apple has temporarily blocked an e-mail app powered by CHATGPT over age concerns. A European Union ruling may eventually have the most impact, given the bloc's track record of policing technology.

Writing in March 2023, with every day bringing new developments, prediction is not possible. GPT-4 is being released, organizations in four countries announced GPT competitors to arrive in weeks or months, an LLM application was banned in Europe, and Apple challenged unrestricted age access. I have been exploring, using, reading about, and discussing the available tools. They are impressive. I will continue to use them as a rapid way to get guidance for deeper research. That said, they won't have a significant direct effect on me, although others may be impacted more. They are reliable enough for some explorations but startlingly unreliable in other areas. It will take time to learn how to use them effectively. I fear many students will not be positioned to develop that skill.

My hope is that you can use this framework to assess developments as they arise, seeing where they conform to and deviate from past events, in this fascinating undertaking of our species. And please, think about possible consequences that we have not yet considered. That will matter most of all.

JONATHAN GRUDIN is an affiliate professor at the School of Information, University of Washington, Seattle, WA 98195 USA. Contact him at grudin@uw.edu.

Digital Object Identifier 10.1109/MC.2023.3263912