03/12/2025

ChatGPT: two years later

TXT JULIÁN PELLER

Has generative AI been a revolution? Is it still revolutionizing things?

Last November 30 marked two years since the launch of ChatGPT, an event that generated major shock waves in technological development, society, and the economy. In the space opened up by this event, it was not always easy (or perhaps even possible) to separate realities from expectations. For example, during 2024, Nvidia briefly became the most valuable public company in the world in a stunning bull run. The company, which makes the hardware used by models like ChatGPT, is today worth six times more than two years ago. But is it really worth that much, or are we facing a collective delusion? This question —and not its eventual answer— is what defines the present moment.

Nvidia's shares multiplied sevenfold in less than two years, and the company briefly became the most valuable in the world (with a market cap that exceeded 3.5 "trillions", or millions of millions, that is, USD 3,500,000,000,000).

D-Day

Let's start by remembering launch day. ChatGPT 3.5 was a chatbot far superior to anything known up to that point in terms of discursive capabilities and intelligence. This leap in capability generated enormous fascination, and the product went viral very quickly: it reached a base of 100 million users in just 2 months, far outperforming other so-called viral apps like TikTok, Instagram, Spotify, etc. It also took over mass media and public debate: AI landed in the mainstream and suddenly everyone was talking about ChatGPT. To top it off, just a few months later, OpenAI launched GPT-4, a version far more intelligent than 3.5 and also capable of understanding images.

The situation sparked debates about the possibilities and problems implicit in this specific technology with respect to copyright, disinformation, productivity, and the labor market, as well as the risks that the advance of AI research could pose in the medium and long term. Various concerns stood out: existential risk (in the style of *Terminator*), a scenario in which the end of work becomes real, and the possibility of the appearance of an artificial consciousness, among others. In this broad and passionate discussion we heard very dissimilar opinions and, as the months went by, the debate began to mature. It took some time to adapt to this

product because the leap it represented left all of us a bit offside. What has happened between then and now?

I love you but I'm a Bard

As far as tech companies are concerned, these two years have been a rollercoaster. The appearance on the scene of OpenAI, with its futuristic advances and its CEO with the spirit and look of a "startup bro", raised doubts about Google's technological leadership, which had been unquestioned until then. Google, for its part, did everything it could to confirm these doubts, publicly humiliating itself on repeated occasions. First, with the embarrassment of the Bard launch video, the chatbot designed to compete with ChatGPT. In this video, the chatbot makes a factual error: when asked for information about the James Webb Space Telescope, the model answers that it was the first telescope to take pictures of planets outside the solar system, which is false. This caused a 9% drop in Google's share price over the following week. Later, with the presentation of the new Gemini model — another competitor, this time to GPT-4— Google again lost credibility when it became known that the incredible capabilities shown in the demo (which would have put it at the cutting edge) were actually fake montages built on top of much weaker capabilities.

Amid this process, Microsoft, the archaic company of Bill Gates —the one that made the old Windows 95 and that we young people hated as much as we loved Google— reappeared and partnered with Altman, integrating ChatGPT into Bing and presenting itself as agile and defiant. "I want people to know that we made them dance," said Satya Nadella, Microsoft's CEO, referring to Google. During 2023, Microsoft grew younger while Google grew older.

This situation dragged on and OpenAI remained for a while the undisputed leader both in technical evaluations and in people's subjective reports (known as "vibe checks"), with GPT-4 at the top. But over time that changed, and just as GPT-4 gained unique leadership at the end of 2022, by mid-2024 its not-so-distant successor (GPT-4o) was competing with other models of similar stature: Google's Gemini 1.5 Pro; Anthropic's Claude Sonnet 3.5; and xAI's Grok 2.

At the time of writing this piece, at the start of 2025, the scenario of intense competition among LLM providers seems consolidated: OpenAI's major efforts to regain its 2022 cutting-edge shine —with its brand-new line of reasoning models o1 and o3— ran into the immediate and forceful competitive response of a new player, the Chinese lab DeepSeek, which released a similarly high-quality model in just two months [1]. What innovation gives you, innovation takes away.a0

Open source strikes back

To complete the picture of blows, crashes, and epic comebacks over the past two years, we have to talk about the world of free software. This new AI era began with two punches to the jaw of the open-source world.

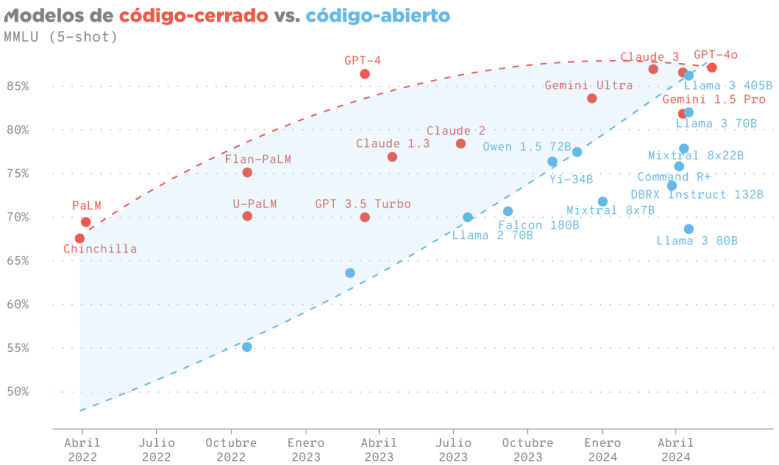
First, despite its name, OpenAI was a pioneer in no longer making its technological advances public. During a golden age prior to 2022, the norm in AI research was to publish detailed results, creating a symbiotic ecosystem between big corporations and academia. Not to go too far, ChatGPT and other large language models (LLMs) are based on the famous paper *Attention is all you need* published by Google in 2017. This paper made public the Transformer neural network architecture, the cornerstone of the generative AI revolution and the "T" in "GPT". But in an exciting plot twist, OpenAI leveraged this discovery to gain an edge and begin conducting research behind closed doors. The launch of GPT-4 was the watershed moment between these two eras: for the first time, OpenAI did not explain anything about the inner workings of its advances. From that moment on, many closed LLMs such as Gemini and Claude appeared, which worsened the research ecosystem.

The second blow to the open-source community was the scale of the new models. Up through GPT-2, a modest GPU [2] was enough to train deep-learning models. Starting with GPT-3, infrastructure costs skyrocketed, and training models ceased to be something accessible to any individual or institution. Fundamental advances ended up in the hands of heavyweights.

But after these blows, and with everyone expecting a knockout, the free-software world fought back and proved to be up to the demands of the new era. It had on

its side, luckily for all of us, an unexpected champion. Mark Zuckerberg, the most hated reptilian android on the planet, pulled off a radical image shift by positioning himself as the standard-bearer of open source and freedom in the field of generative AI. Meta, the conglomerate that dominates a large portion of the fabric of Western digital communications according to its own design and will, took it upon itself to bring open source into the age of generative AI with its LLaMA model line. It is definitely a bad time to be a Manichean. The first LLaMA releases came with timid open licences and limited capabilities (although the community made great efforts to believe otherwise), but with the recent versions (LLaMA 3.1, 3.2 and 3.3), the gap with private offerings began to close, allowing the open-source world to keep up at the cutting edge of technological research.

Very recently, the epic open-source comeback grew even more epic and added two new unexpected champions: the Chinese labs DeepSeek, with its eponymous family of models, and Alibaba, the "Chinese Amazon", with the Qwen model family. These companies came on strong with cutting-edge open-weight models that placed them among the big U.S. players, causing an international stir. Notably, DeepSeek-R1, released on January 20, 2025, is a free model with quality comparable to o1, the most advanced closed model available at that time, but at a much, much lower cost. R1 forcefully crowns the open-source recovery process after two years of crisis.

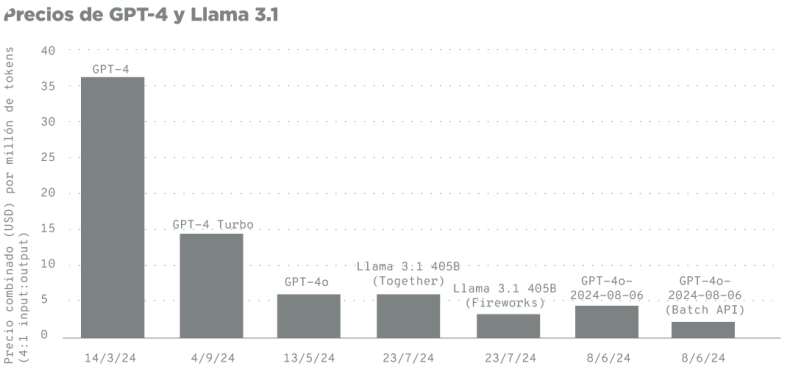


The gap between closed and public models is narrowing. With the latest LLaMA versions, open-weight models have caught up and reached competitive performance with the best private LLMs.

At the beginning of 2025 this situation consolidated strongly: the launch of the cutting-edge closed model o1 was followed by a comparable public version (DeepSeek-R1) in just 2 months. Source.

An idea never explored before

Over these two years, research into large language models has been prolific. The first fundamental advance, now taken for granted, was that companies managed to increase **chatbots' context windows** (how many tokens [3] they can read as input and generate as output) while at the same time **bringing per-token costs down** in an impressive and sustained way. But we've also seen models become **multimodal** (accepting as input not just text but also images, audio, or video), we've seen LLMs be allowed to **use tools —notably, searching the Internet—**, and we've seen constant advances in capability.



The price per million tokens has dropped drastically since the arrival of LLMs. The cost for models like GPT-4 fell from USD 36 to less than USD 4 in under a year. For newer models, the drop is even faster: while o1 costs USD 24, DeepSeek-R1, released just 2 months later, costs less than USD 1. Source.

On the other hand, various **quantization and distillation** techniques appeared, allowing huge models to be compressed into smaller versions, to the point where they can run on desktop computers (even if the price to pay has sometimes been an unacceptable reduction in capability). Everything seems to indicate that this size optimization is on a positive progress curve and that we are close to **small language models (SLMs) that will eventually fit on a smartphone.**

On the negative side, **there have been no fundamental advances in controlling the famous hallucinations**, the false texts that models can generate with great plausibility. This problem, quaint at first, now seems confirmed as a structural feature of this technology. Those of us who use it for work know the frustration of relying on a tool that usually behaves like an expert but 10% of the time makes a gross mistake without noticing or outright invents something. In this respect, Yann LeCun —head of Meta AI and a heavyweight in the industry— can claim a win for himself, given that he held a rather deflationary stance on LLMs during the peak enthusiasm of 2023.

But pointing out limitations in LLMs does not mean the debate on their capabilities and how far they can take us has been settled. For example, Sam Altman believes the current research program still has a long way to go before it hits a wall, and the market, as we'll see shortly, seems to agree. Many of the advances we've seen in these two years also seem to back him up. We've seen OpenAI launch its **voice assistant** and an improved version that allows almost real-time interaction, with interruptions (as between humans) rather than rigid turns. Recently, we have also seen the first advanced attempts at **LLMs with access to and control over a browser or even users' computers**, as in the 4o demo, the launch of Claude 3.5, and the recent limited trial of the Operator product. These tools, still in an embryonic stage, offer a glimpse of what the near future might look like with **LLMs with greater agency** and represent one of the current strong bets of the research program.

On another front, the family of **o1 and o3 models**, presented as "reasoning models" (and no longer just "language models"), anticipates a research track with a lot of room for improvement. Indeed, this family starts from a novel idea: leveraging inference time —rather than training time— to improve the quality of the generated response. With this strategy, the LLM does not immediately generate the most probable first word, but instead has the ability to "pause to think" before it starts speaking. One of the researchers on the project suggested that, in the future, these systems could use weeks or months of compute to produce a single answer, an idea never explored before. At the end of last year, the

announcement of o3 doubled down on this "reflective pause" and surprised many with notable advances in abstract reasoning, advanced mathematics, and programming. The progress that o3 showed in the demo caused a great stir and high expectations, as it seems to strongly validate this line of research and, with it, a clear direction to pursue in order to achieve increasingly intelligent models. It remains to be seen how well the announced results hold up once the model is publicly accessible and, on the other hand, how far this line of research will go, with upcoming steps —o4 and o5, R2 and R3— potentially marking a new milestone in the evolution of generative AI.

In terms of AI safety, we have seen a key advance from Anthropic. The company has achieved **a fundamental interpretability milestone** for LLMs, allowing us to understand a bit better the black box that these architectures are. With its discovery concerning the polysemous nature of neurons and its method for extracting patterns of neural activations that represent concepts, the main barrier to controlling transformer models seems to have fallen. The ability to manipulate circuits at will, bending these models' observable behavior, is promising and has brought a degree of reassurance regarding the gap between LLM capabilities and our understanding of them.

Beyond language models, these two years have brought enormous advances on other fronts. The first mention must go to **image generation,** the text-to-image models that began to take off even before chatbots and continued to develop at a rapid pace until they overflowed into **video generation**, with products like Sora and Veo 2, capable of producing extremely high-quality videos. Perhaps a bit less well-known, but with equally astonishing advances, are the fields of **music generation**, with platforms like Suno and Udio, and **voice generation**, with Eleven Labs leading the way in achieving extremely high quality standards.

They have undoubtedly been two intense years.

Shovels and picks

If we turn our gaze to the financial side of the phenomenon, we see huge amounts of capital being poured into AI in a sustained and growing way. We are currently

in the midst of the **AI gold rush**, and no one wants to miss out on a technology that its inventors, modestly, presented as equivalent to the steam engine, the printing press, or the Internet.

It may be telling that the company that has capitalized the most on this rush does not sell AI, but the hardware that AI uses as infrastructure, fulfilling that old piece of folk wisdom that says that during a gold rush a good way to get rich is to sell shovels and picks. As I mentioned earlier, at the end of 2024 **Nvidia became the most valuable company in the world** when it reached a market cap of 3.5 "trillions" or millions of millions. To give some context, USD 3,500,000,000,000 is a figure far higher than France's GDP. Even today, after the recent market jolt caused by the launch of DeepSeek, with a 17% plunge, the company remains in the top 3 and with a market cap six times higher than it had in 2022.

On the other hand, if we look at the list of public companies with the highest market value, we find the podium dominated by tech companies that are partly or wholly tied to AI's promises. Apple, Microsoft, Nvidia, and Google are the top four as of the date of writing and **between them they exceed USD 12T in market cap**. By way of reference, in November 2022 the combined market cap of these four companies was less than half that amount. Meanwhile, generative-AI startups in Silicon Valley are raising record funding rounds. As Maslatón would say, the market is bullish.

While technology is advancing without brakes, the business model of generative AI, beyond a few specific cases, does not seem clear. In parallel with this ongoing bull frenzy, some voices, including recent Nobel laureate in economics Daron Acemoglu, are skeptical about AI's ability to pay back all the money being poured into it. For example, in this Bloomberg interview, Acemoglu argues that current generative AI will only be able to automate less than 5% of existing tasks in the next decade, and that this can hardly be the productivity revolution that investors are counting on.

Is this the AI fever or rather the AI fever dream? For now, expectations around the current program are sky-high and the bull run shows no sign of stopping. As with any bubble, with the benefit of hindsight it will be very easy to

look back and see it burst, but from inside it is not obvious whether there will be a correction and, if so, when. Are we in a bubble about to burst as Acemoglu believes? Or, as one investor suggested, is Nvidia on its way to becoming a company worth more than 50 trillion dollars in a decade? This is the big unknown and, unfortunately, I do not know the answer. Everything seems to indicate that, just as in the dot-com bubble, we will emerge from this situation with some companies riding the wave and many under water. My only suggestions are the following: distrust anyone who conveys a strong sense of certainty and always consider the possibility of a sharp market correction if you plan to invest in the big tech companies.

The problems of intelligence

Let's now talk about the broader impact of the arrival of generative AI at the social level. The quality leap of ChatGPT compared to the socially known technological horizon at the time of its launch created a big stir, opening debates about the opportunities and risks of that specific technology, as well as the opportunities and risks that more advanced technological developments might entail. Unlike what happened with the arrival of the Internet and social networks, this time society seemed to react quickly and to show itself attentive and concerned about the implications and challenges that this new technology might pose. In addition to the deep debate on existential risk, related to future technological development and the question of the speed of progress, the impacts of existing language models were also widely discussed. **The problem areas of generative AI are several, but perhaps the three most prominent are: 1) the fear of a boost in disinformation and digital pollution, 2) major problems with copyright and the use of private data, and 3) the impact on productivity and the labor market.**

Regarding **disinformation**, this paper suggests that, at least for now, we have not seen a significant increase in the level of disinformation we are exposed to. It is hard to say with certainty, but my personal impressions are similar: while disinformation is strong and may even have increased in recent years, it has not

gone through any significant phase change that could be attributed to the emergence of generative AI. This obviously does not mean to suggest that disinformation is not an extremely serious problem. The thesis is weaker: generative AI does not appear to have had, at least so far, a significant impact on this problem.

We have indeed seen **deepfakes**, as in recent news stories about the production and distribution of pornographic material generated with AI using the faces of real people and several school cases where the affected subjects are underage girls. This is extremely serious and it is necessary to bolster judicial and policing systems to address it, but it seems, at first glance, controllable and, in the grand scheme of things, represents relatively minor impacts compared to a possible disinformation nightmare unleashed by generative AI. Legal systems take longer than we would like, but there are signs that institutions may be up to the task, at least when it comes to deepfake porn involving minors, as illustrated by the exemplary 18-year sentence handed down in the United Kingdom to a man who created and distributed such material.

Regarding the impact on the **labor market and productivity** —the flip side of the market boom— the debate does not seem settled, nor is the potential scope clear, either in terms of increasing worker productivity or in terms of shrinking or expanding job numbers. If you go online, you will find very diverse opinions about the impact this technology will have. We hear phrases like "AI replaces tasks, not people" or "you won't be replaced by AI, but by a person using AI", delivered with great confidence and no sources, something reminiscent of a language model's hallucinations. It is true that ChatGPT cannot perform complex tasks and, for those of us who use it day-to-day, we are well aware of its enormous and frustrating limitations. But it is also true that tasks like drafting a professional email or scanning large amounts of text for specific information have become much faster. In my experience, I can say that in the world of programming and data science, productivity has increased enormously with AI-assisted programming environments like Copilot or Cursor. In my team, less-senior profiles have gained greater autonomy, and all of us produce code faster than

before. However, this same speed can be a double-edged sword, and some studies suggest that code generated with generative-AI assistants may be of worse quality than code generated by humans without such assistance.

If the impact of existing LLMs is not entirely clear, we must add to this the uncertainty associated with major advances in novel research lines that fuel expectations about the automation of complex tasks (notably, the reasoning-model and agent-with-browser-control lines of research we discussed earlier). And while the market is betting heavily on a productivity boom driven by generative AI, many serious voices downplay this technology's potential impact on the labor market, as we mentioned earlier when talking about the financial side of the phenomenon. In principle, the strongest limitations —hallucinations— have not only not been solved, they increasingly seem unsolvable, while human institutions have proved less agile and revolutionary than the technology itself, pouring cold water on the conversation.

In any case, the promise of a massive revolution in the workplace over these two years has not yet materialized. Considering the accelerated adoption of this technology (according to this study, more than 24% of U.S. workers now use generative AI at least once a week) and assuming that early adopters are perhaps the ones who stand to gain most, we can assume that we have already seen a significant share of its potential impact.

Another major challenge brought by the emergence of generative AI has been **problems around copyright.** Content creators —including artists, writers, and media companies— have expressed their discontent at seeing works they own used without authorization to train generative AI, which they consider a violation of their intellectual property rights. On the other side of the coin are AI companies, which argue that the use of protected material to train models is covered by the concept of "fair use" and that the output of these models is not reproduction but a creative transformation of the works.

This conflict has led to a large number of lawsuits, such as Getty Images' case against Stability AI for the unauthorized use of images to train models, or lawsuits by artists and authors, such as the one filed by Sarah Silverman against OpenAI,

Meta and other AI companies. Another notable case is the suit filed by record labels against Suno and Udio, alleging copyright infringement for using protected songs to train generative music models.

In this futuristic replay of the old distinction between inspiration and plagiarism, court cases have so far not clearly tipped the scales in one direction or the other. Courts have allowed some aspects of the suits to move forward while dismissing others, sustaining a climate of uncertainty. Recent legal filings and the stance of companies, including strategies like indemnifying Adobe, Google, and OpenAI customers, show that the issue remains far from definitively resolved.

The regulatory framework for AI has also seen important steps forward; the most significant on this side of the globe was the approval of the **AI Act** by the European Union in March 2024. This legislation positioned Europe as the first bloc in the world to adopt a comprehensive regulatory framework for AI and sets out a phased system to ensure compliance. Its implementation will begin in February 2025 and will roll out gradually. The AI Act classifies AI risks and restricts cases of "unacceptable risk", such as using the technology to deceive or to carry out social scoring. Although certain provisions were softened during the debate, securing basic rules applicable to all models and stricter regulations for applications in sensitive contexts, the industry has voiced concerns about the burden this regulation represents.

With these tensions, opportunities, and challenges, it is clear that the impact of generative AI marks the beginning of a new stage of profound social, economic, and legal transformations whose scope we are only just beginning to understand.

Big things coming

We are living in a present of excitement and expectations: a full-blown springtime of AI, with impressive advances that keep coming and promising lines of research waiting to be explored. On the other hand, these are uncertain times. The suspicion that we are in a bubble and the expectation of a major stock-market and emotional correction is more than reasonable. But as with any market correction,

the important thing is not to predict that it will happen, but to know exactly when.

What will happen in 2025? Will Nvidia's stock collapse or, on the contrary, will the company continue its bull run under the promise of becoming a **50-trillion-dollar** company in ten years? And what will happen to the AI stock market in general? On the other hand, what will happen with the line of reasoning models started by o1, R1, and o3? Will it hit a ceiling or continue a spiral of massive progress just as the GPT line did from versions 1 through 4? How much will today's crude LLM-based agents that control desktops and digital environments in general improve?

We will find out sooner rather than later, because that is where we are headed.

elgatoylacaja.com/en/chatgpt-two-years-later
