A Review of AI in 2024

Ciarán Bryce

January 2025

AI continued to dominate the IT sector and beyond in 2024. This article presents 10 key lessons from the year, and proposes some perspectives for 2025.

1 Introduction

AI was one of the most impactful technologies of 2024. Large investments by technology companies in the development of AI platforms continued, as have investments by industry in adoption of AI in search of economic advantage. Regulation was enacted to control AI risks, while at the same time, the technology is increasingly used by bad actors.

AI is a fascinating subject because, as well as breaking new technical ground in IT and applied mathematics, AI is leading to a transformation of the workplace, an evolution of the IT economy into an AI economy, the enactment of new regulations, new methods of education, further discord in geopolitics and a worsening of cybercrime. One of the problems with AI is a difficulty to evaluate the technology and its impact in a measured manner. There is a still much scare-mongering, hyperbole, AI-washing and trivialization around AI. One of the goals of the Technology Watch portal is to report on the latest news on AI and to identify key subjects. Live for nearly a year now, this article summarizes the main lessons learned analyzing AI news in 2024.

Section 2 presents ten lessons learned in 2024 which will also have a bearing on 2025. Section 3 goes into more detail, adopting an “A” to “Z” approach to highlight the key points. Perspectives for 2025 are presented in Section 4.

2 The Year 2024 in 10 Lessons

Lesson 1 – Even Larger Language Models

Big Tech has continued intensive research and development into large language models with *even larger* models being released by Google DeepMind (Gemini 2.0 and Gemma), Anthropic (Claude 3.5 Sonnet), OpenAI (GPT 4o, o1 and o3), Meta (Llama 3-B), Cohere for AI (Aya 23), Mistral AI (Mistral Large, and the specialized Codestral and Mathstral), and others.

One technical measure of increased model size is the number of model parameters – which is a strong indicator of a model’s ability to learn and reason. For instance, GPT-3.5 was rumored to be composed of models that have around 175 billion parameters each, while GPT 4-o is composed of 8 models with an estimated 220 billion parameters each. The largest version of the Llama 2 model family in 2023 had 69 billion parameters; its 3.1 version released a year later has 405 billion parameters. Another technical measure of model size is the context window length. This represents the number of input data (tokens) that a model can store when it processes a request. The context window stores the prompt, conversation history and relevant documents, so the larger the context window, the more pertinent the model response should be. The Llama models’ context length has gone from 4’096 tokens in 2023 to 128’000 tokens in 2024. Anthropic’s

Claude’s context length is now over 200’000 tokens, which Anthropic says corresponds to 500 pages of text or 100 images.

Another reason for increased model size is the expectation that models generate multi-modal content, that is, content in text, audio, image and video formats. Technology companies want users to be able to prompt with audio commands for agent assistance tools.

Interest in AI can be seen from spending by Big Tech which is projected to exceed 240 billion USD in 2024. This growth is explained by an increasing global market for AI – expected to reach 20 trillion USD globally by 2030 – and by the large costs of training and running AI models.

Lesson 2 – Reality Checks Large Language Models

Despite the larger sizes of models and reported improved benchmark performances, there is evidence that large language models are reaching tangible limits and that AI companies (OpenAI, Google and Anthropic) are unhappy about the performance of their large models.

One reason for the plateau is the large cost of training large language models. Stanford’s AI Index tracker reports the total training cost of Gemini to be over 190 billion USD and GPT-4’s total training cost to be over 78 billion USD. In comparison, Google’s transformer model was trained for only 930 USD in 2017.

A second limit on the growth of large models is the penury of high-quality and diverse training data. This problem has several facets.

- One class of data on which AI companies wish to train their models is personal data, and social media platforms are very attractive spaces to harvest such data. However, AI companies are facing roadblocks, notably from the EU’s General Data Protection Regulation (GDPR). Social media users should have to explicitly consent to their data being used to train models, and this has not been the case in the past.

- Another worry for AI firms is that training data scraped from the Web includes copyrighted material. Though AI firms are claiming that such content should be usable for training under the “fair use” provision of copyright law, AI companies are the targets of lawsuits for copyright violation. Many content provider sites now block web crawlers from AI firms.

- Yet another worry is an increasing amount of content on the Web that is itself generated with AI – content that is now called AI Slop. For instance, Meta admits that 1 million businesses are creating more than 15 million ads per month using generative AI. Studies show that the quality of generative AI model outputs degrades when trained on data that was generated using AI – a phenomenon known as *model collapse*.

Another challenge that AI companies are facing is that data center power consumption is increasing because of AI. To meet emerging energy needs, Microsoft has signed a deal for energy from the Three Mile Island nuclear power plant while Amazon has bought a data center powered by nuclear energy in Pennsylvania. Meanwhile, Google is purchasing six or seven small nuclear reactors which should be operational between 2030 and 2035. AI is having an ecological impact. Google’s 2023 emissions were up 13% on the previous year, and Microsoft’s emissions have increased by 29% since 2020. Also, a Nature article claims that generative AI models could produce 5 million metric tons of e-waste by 2030.

On the brighter side for large language models, studies show that improved training data quality using extensive human annotation can improve performance. In the case of the Molmo models, the training dataset contained only 600’00 images but they were annotated by hand beforehand with the contents of images being explained in detail by the annotators. The developers claim that the model, built with only 72 billion parameters, performs as well as the major proprietary models from Google, Anthropic and OpenAI. Another interesting avenue is the test-time compute technique: rather than using pre-trained patterns to recognize the input and generate the most likely output based on probabilities, this technique gets the model to generate several possible answers, and then use reasoning to determine the best answer. Used in OpenAI’s o1 and o3 models, the company mentions that having a bot think for 20 seconds in a hand of poker achieves the same boost in performance as scaling up the model by 100’000 or training it for 100’000 times longer.

Lesson 3 – OpenAI Is Silicon Valley’s Largest Laboratory Mouse

OpenAI has had an eventful year. The company released GPT-4o in May 2024, its largest model to date.

The company has changed greatly since it was founded in 2015. At that time, its goal was to “*advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return*”. It also maintained a cap on return of investments: an investor could get no more than 100-fold his investment in the company, with higher returns being funneled to the non-profit organization. Its CEO, Sam Altman, announced that the company is no longer controlled by the non-profit board which effectively means the end of the non-profit foundation. OpenAI has also pivoted on its promise not to work on military projects. These moves come at a time when the company is expecting an annual loss of 5 billion USD, and will suffer a loss of 14 billion USD in 2026. These losses are forcing the company to explore new revenue streams.

The company is also having to deal with lawsuits. Along with Microsoft, it is the defendant in a lawsuit by the New York Times, being accused of having trained its GPT models with content from the paper’s website without the consent of the newspaper. In addition, it risked a lawsuit from a US actress, Scarlett Johansson, for creating a voice assistant called Sky whose voice sounds like that of the actress in the 2013 film “Her”, in which she played the role of an AI assistant.

OpenAI has further been destabilized by the departure of high-profile staff over the last months. This led Microsoft, which invested 13 billion USD in the start-up in 2019, to diversify its AI investments in 2024 and to develop its own AI models (Phi-3 language models). The aim of Microsoft is to protect itself from a possible disappearance of OpenAI. At the same time, OpenAI would like to be free of Microsoft. OpenAI has been making claims that its recent o3 model is close to AGI. This is a doubtful claim, but according to the deal OpenAI struck with Microsoft, the company will no longer be obliged to share advanced technologies with Microsoft once AGI has been achieved.

Despite these challenges and the governance dramatics, the company raised 6.6 billion USD in a funding round which values the company today at 157 billion USD, bringing total cash raised to 17.9 billion USD.

Lesson 4 – The Emergence of Small Language Models

Training and running large language models is expensive in time, energy and hardware. Very few organizations can afford to build their own large language models, and therefore use large language models provided on the cloud. However, this solution is unacceptable to many companies because of data protection issues – where compliance standards or cybersecurity concerns prohibit an organization from transferring data outside of its private network. One solution to addressing this concern is to use smaller language models (SLM) that can be efficiently run on standard off-the-shelf computers. There is debate about what size an SLM is, but there is consensus that it should run on a single GPU or have at most 5 billion parameters, which makes it more accessible to corporate environments, in part because the model can be run on premises. In comparison, a large language model can require up to 10’000 GPUs to run.

One approach to small language models is to develop smaller versions of a larger, more powerful, model. The smaller version would not have the same level of accuracy as the larger model. One approach to achieving this is through distillation: the process of transferring knowledge from a bigger (teacher) model to a smaller (learner) model. Large language models from Llama, Nvidia and Claude all have distilled smaller model variants. Quantization (transforming model weights from floats to lower-precision integers) and pruning (removing less utilized weights) also reduce model size enough to run them on off-the-shelf devices.

For instance, PruneBERT is a pruned model that has a 97% reduction in weights compared to BERT but still exhibits 93% of the original model's accuracy, but with significantly increased inference times.

Another emerging type of small language model is the specialized model where the model is designed for very specific tasks, e.g., optical character recognition (OCR) type tasks or text-to-SQL transformations. It is suggested that future company IT infrastructures could be 50% traditional applications and 50% SLMs.

Lesson 5 – Content Providers and AI Companies Prepare for a Marriage of Convenience

One fundamental question that Big Tech continuously asks itself is "*where do people go on the Internet to find information?*". For three decades, the answer to this question was "search engines", with Google having a quasi-monopoly (around 90% of the market). This dominance led to the search engine optimization (SEO) industry, where companies pay to appear at the top of search results. The SEO industry generated 68.1 billion USD globally in 2022. It had been expected to reach 129.6 billion USD by 2030.

The arrival of language models is hurting the search engine industry because they offer an alternative information source to people. AI Chatbots copy information from their sites, sometimes from behind a paywall, and offer this content to users. This zero-click feature is creating concern among content providers because they fear fewer visits to their websites, and hence reduced revenue from advertising. In addition to the qualms about their content being used to train ChatGPT's models in disregard of copyright law, the fear of lost advertisement revenue is perhaps one factor that pushed the New Your Times to take a lawsuit against OpenAI and Microsoft for copyright infringement. Also, two AI music startups, Suno and Udio, are being sued by Sony Music, Warner Music Group, and Universal Music Group for copyright infringement. The AI companies have platforms that allow users to create songs, but the record labels claim that copyrighted music was used in the training data.

However, in the medium to long term, AI firms and content providers need each other. AI firms need content to train their models, and content providers need revenue. There is a premise of *rapprochement* between these sectors. For instance, Reuters, the Atlantic, *Le Monde* in France, the Financial Times, Axel Springer, Condé Nast, and Time have signed an agreement with OpenAI to become official content providers. Youtube is reportedly signing a deal with record companies where, in return for a lump sum, Youtube can use copyrighted material for training.

Lesson 6 – Regulation

The European Union's Artificial Intelligence Act entered into force on August 1st, 2024 and will become fully effective in two years time (August 2026). The overriding goal of the regulation is to protect the health, safety and fundamental rights of EU citizens from risks that IT systems using AI pose.

According to the European regulation, AI model providers must conduct a documented risk assessment of the system, with the system being classified into one of four categories (*unacceptable*, *high*, *limited* or *minimal*). In the event of a high risk classification, the system must be registered in an EU database. The provider must carry out extensive documented testing and validation on the AI system, and instigate post-deployment monitoring of the system and do record-keeping of incidents. Systems with unacceptable risk (e.g., of subliminal manipulation of people) are banned. Common AI chatbots are generally classified as limited risk, where the requirement is that users be made aware that they are interacting with a bot, and not a human.

Another regulation, which this time did not come into effect, was California's SB 1047 bill. This bill only applied to large AI systems (costing over 100 million USD and use 10^{26} FLOPs of processing during training). The bill was opposed by many Silicon Valley firms, including OpenAI, despite late amendments incited by Anthropic and others. Governor Newsom vetoed the bill, not because he was against the need to regulate AI *per se*, but because he did not see the watered down version of the bill as being effective. The return of Donald Trump to the presidency is raising doubts about whether regulation will emerge in the US in the next years.

In Switzerland, the government published Guidelines on Artificial Intelligence for the Confederation in 2022. The report calls for a “balanced approach” for regulation between the need for safety on the one hand and research innovation on the other. Nonetheless, the seven guidelines reflect many of the issues raised by the EU’s regulation. Like for the GDPR in the context of personal data protection, the EU’s AI Act could become a *de facto* regulation for Swiss organizations.

Lesson 7 – Social Media, Crime and Covert Influence Operations

The largest criminal use of generative AI today is by pedophile groups using it to create images and videos of children being sexually abused. Sextortion is also evolving, from situations where victims had photos of themselves posted by former partners for romantic revenge or blackmail, to general blackmail where the criminals “nudify” photos of victims taken from social media. Another criminal use of AI are “heists” where criminals use deepfake technology to impersonate company executives and convince employees to transfer large sums of money. Related to this, are pig butchering schemes where the criminal builds an intimate relationship with the victim before asking for money, usually for an “investment opportunity”. An example are romance scammers who use real-time face replacement to masquerade as potential love interests to gain victims’ confidence. An emerging trend is chatbot radicalization where an apparently friendly chatbot convinces someone into performing a radical act. This is what happened with the perpetrator of an attempted crossbow attack on Queen Elizabeth II in 2021.

Other examples of AI-based cybercrime include using AI to overcome language and dialect barriers that criminals have traditionally faced. Another is the creation of more convincing web-sites with job offers that scam potential job-seeking candidates, as well as fake commercial websites that contain crypto-draining malware that drains victims’ accounts.

Russia was named as the top source of disinformation campaigns on Meta platforms. Meta took down 20 so-called covert influence operations in 2024. The year was an important test of the impact of covert influence operations with many elections being held worldwide. One operation used AI to create fake websites that looked very similar to those of Fox News and the Telegraph in order or propagate disinformation about the war in Ukraine. Other operations include using GenAI to create large volumes of social media comments in multiple languages as well as turning fake news articles into Facebook posts. In the context of the recent US elections, Meta blocked over 500’000 requests to create AI generated images of the presidential candidates. Meta also noted the frequent creation of inauthentic accounts whose purpose was to publish content to influence opinion.

An argument made is that it is not enough to create disinformation: effort must still be made to bring people to this content, and this still requires paying influencers. It is noteworthy that a party in Bangladesh amassed 3.4 million followers across 98 pages of fake content, and the fact that the party was not politically relevant to US big tech could have meant that this went unnoticed for longer. At the same time, the community is getting organized with initiatives like the AI Incident Database and the Political Deepfakes Incidents Database, as well as more cases of social media users alerting the platform administrators and other users of disinformation cases.

Lesson 8 – AI Enters Mainstream Science

The AI field won Nobel prizes in both chemistry and physics in 2024. In chemistry, the prize from the Royal Swedish Academy of Sciences was attributed to Google DeepMind’s Demis Hassabis and John M. Jumper for their work on using AI to predict protein structures. Hassabis and Jumper used their AlphaFold AI to predict the structure of a protein from a sequence of amino acids in 2020. Since then, the tool has predicted the shapes of all currently known proteins. The ability to predict protein structures is critical to the development of efficient drugs, vaccines and even cures for cancer. DeepMind released its latest model, AlphaFold 3, as open-source.

Meta released Open Materials 2024 (OMat24) – a massive data set, with models, for use by materials scientists. The data set has around 110 million data points, making it much larger than any data set seen in the domain, and existing data sets are proprietary. Material scientists discover materials by taking elements from the periodic table and simulating different combinations. This is a long process, given the huge number of possibilities, so the OMat24 data set has significant importance to researchers. Only a company with a huge compute power could produce such a data set. The search for new materials is important for areas such as new fuels and non-polluting building materials. Such large data sets are being used to train new AI models in materials sciences, biochemistry and physics. This shows that the scale of problems that can be addressed by AI is increasing rapidly, and milestones like precisely simulating how drugs bind proteins could be reached sooner than expected with the contribution of AI. These developments suggest that AI is now as critical to scientific breakthroughs as IT and even mathematics were in the past.

Lesson 9 – AI Enters Mainstream Use

Generative AI is now widely used in the workplace. One area where it has taken a particular hold is in software development. A Github survey in August reports that 97% of developers in large organizations regularly use AI, and tools like Github Copilot and Mistral’s Codestral are popular.

A KPMG survey of large organizations shows that the sectors using generative AI are inventory management (64%), healthcare for document assessments (51%), technology and media for workflow automation (43%), and financial services for customer service chatbots (30%). 71% of respondents say that GenAI is already impacting their business models through data analysis in making decisions. The driving factors for investment are increased revenue and productivity. A report by the US National Bureau of Economic Research (NBER) found that by March of 2024, over 40% of adults in the US have used generative AI. The most popular GenAI tools are used more than 3 billion times per month. Nonetheless, adoption in the enterprise is still not as advanced as was foreseen one year ago. The main roadblocks are fears of leaks of personal or corporate data as well as regulatory uncertainty.

Despite this progression, it is notable that AI is not (yet) having an impact global on economic productivity. For instance, productivity in the US was around 3% until 2005, but has significantly fallen since despite Big Tech inventions like smartphones and social media. One reason for this might be Solow’s law which postulates that technologies take time to impact economic growth. Also, in the US at least, overall productivity has depended a lot on productivity in manufacturing industries, which is not really in the radar, or business models, of Big Tech firms. The potential for AI in manufacturing is understood, but current AI models are seen as unhelpful because they are not trained with the required manufacturing domain-specific knowledge (much of which is proprietary). Further, issues like model hallucination contrasts with the high-precision requirements and strict standards of manufacturing industries.

Lesson 10 – The Big Tech Hegemony Remains Stable

The hegemony of Big Tech has remained stable over the past year. This is ultimately not surprising because of the huge cost of developing new language models, and few outside of the Big Tech companies have the resources to do so. Also, Nvidia is making it very difficult for startups in the chip domain to make a niche because software typically needs to be adapted to a particular chip, and AI model creators tailor their software to Nvidia and other popular chips because of their market dominance.

Among chip manufacturers, Intel announced a lay-off of 15 percent of its work-force in August following 7 billion USD in losses in 2023 and a 31 percent decrease in revenue from 2022. The company has had a turbulent few years, exemplified by Apple discontinuing Intel chips in its products in favor of their own chips, and there were already mass layoffs in October 2022. On the other hand, Nvidia’s market cap reached 3 trillion USD in June, only the third IT company to achieve this figure after Microsoft and Apple. OpenAI and Google wish to develop their own AI chips to lessen their dependence on Nvidia. Manufacturing companies can expect to benefit from investment grants following the US CHIPs Act – a legislation aimed at encouraging chip manufacturing on US soil, notably to ensure that there is a supply of chips in the event of Taiwan being invaded or blockaded by China.

Generative AI models are particularly threatening to search engines – since people are consulting models rather than search engines. Google has the most to lose. Nonetheless, the company is not yet feeling the effects as the Google's search engine generated just under 50 billion USD in ad revenue this year (a 12% annual increase) and its cloud revenue is up 35%. Alphabet reported a 34% jump in profits for Q3. The challenge for chatbots is that there is not yet any obvious revenue model, apart from subscriptions.

3 Details – The A to Z of AI in 2024

Agents

The terms "*agent*" and "*agentic*" have emerged as buzzwords this year. They refer to an AI assistant that can do tasks like booking holidays on-line, including handling payments. In the ideal scenario, the agent would book the holidays based on the owning human's preferences and time table. Agents are seen by many as the next "Killer App" of AI, notably by investors looking for a quicker return on their investments. A Capgemini report which claims that 10% of US companies are already using AI agents, and 80% will adopt them in the next three years. Advances in voice recognition facilitate vocal interactions with agents. Microsoft uses the term "copilot" for agent; Google uses the term "universal assistant".

Anthropic

The funds raised by Anthropic reached 9.7 billion USD in 2024. The company wants to build AI Agents for desktops that can perform "back-office jobs" like document search, answering emails or tasks like filling out Web-forms, e.g., booking an airline flight, by searching for the relevant information on the desktop. The newest version of the Claude 3.5 Sonnet model can now interact with desktop applications and is able to emulate gestures made by a person sitting at a PC like mouse clicks and movements. On safety issues, Anthropic is working with the UK's AI Safety Institute for Claude 3.5.

Anthropic researchers created a detailed map of the inner workings of its Claude 3 Sonnet 3.0 model. This map allows researchers to examine how neuron-like data points, called features, influence the generative AI's output. Some features are "safety relevant," which means that identifying these features could help tune AI to avoid dangerous responses being generated by the model. Features in the model were identified when the model was questioned in relation to security vulnerabilities in code, or asked to explain how to produce dangerous content (like bioweapons). The Anthropic researchers experimented with clamping to increase or decrease the intensity of specific features, aiding in tuning models to handle sensitive security topics appropriately.

Apple

Apple developed its offering of personalized AI services for the iPhone 15 and on MacOS Sequoia on Macs and iPads with M1 chips or newer. Since these chips cannot do extensive AI processing, many tasks are done on the Cloud. Apple says that any personal data sent to the cloud gets encrypted, and deleted immediately after use. Only the AI task called for by the user is able to decrypt the data. Apple invited independent cybersecurity researchers to review their security process, which they call the Private Cloud Compute. Apple has less incentive than other companies to collect personal data since its business model is more oriented towards hardware and services, rather than ads. Apple integrated ChatGPT with its Apple Intelligence feature.

In a landmark ruling, the European court of justice decided that Apple must pay 13 billion EUR in taxes to Ireland, upholding a decision made by the European Commission in 2016. The Commission is trying to crack down on favorable "sweetheart" tax deals with multinationals. Apple has had its European headquarters in

Ireland since 1980, and Ireland has always worked to attract Tech companies through financial incentives. Apple's tax rate in 2014 was effectively 0.005%, which the Commission believes gave Apple an unfair advantage when marketing its iPhone.

Artificial Generalized Intelligence (AGI)

Artificial general intelligence (AGI) which is loosely defined as the ability to understand and learn on a range of distinct tasks in a manner that approaches the ability of humans. One test traditionally seen as a test of AGI is Turing's test which measures a machine's ability to be indistinguishable from a human in a conversation. Modern tests are more formal. One is the ARC-AGI test created by François Chollet. This contains a series of a series of puzzles that the AI needs to solve. Since December, the high-score is held by OpenAI's o3 model. OpenAI claims to be close to AGI, but it should be noted that according to the deal OpenAI struck with Microsoft, the company will no longer be obliged to share advanced technologies with Microsoft once AGI has been achieved.

Avatars (or Deepfakes)

A deep-fake is a full-body avatar of a person which can move around in a virtual space. Perfect deep-fakes are not yet created. The human brain rejects avatars after a moment because it detects features that are unreal or unnatural – the phenomenon known as the *uncanny valley*. The company Synthesia is one of the leading in the business. Interestingly, Synthesia uses the term synthetic media rather than deepfake because of the negative connotations associated with deepfakes (e.g., many deepfakes have been created from images taken from social media without consent to depict sexual content, and deepfakes have also been created for political campaigns to spread disinformation). Cybercriminals have also used deepfake technology to scam significant amounts of money.

In China, a market has emerged for deepfakes that clone deceased people. Several Chinese companies provide lifelike avatars of the deceased, accessible via apps or tablets, costing from a few hundred to a few thousand dollars. These companies claim that they help people process grief. The technology is also used to create avatars of deceased Chinese writers, thinkers, celebrities, and religious leaders for educational and memorial purposes.

Benchmarking

Benchmarks are the standard by which language models are evaluated, with MMLU-Pro (a large set of multiple-choice questions from several subjects, including math, physics, health, psychology, and philosophy) being popular for general Q&A. There are several popular domain specific benchmarks, e.g., HumanEval for coding, PubMed and MATH. Hugging Face announced an AI leaderboard for models in the FinTech domain. The benchmark evaluates models' financial skills like extracting financial information from reports and predicting stock movements.

Research from Stanford University is highly critical of the benchmarks currently used to evaluate the performance of large language models. Benchmarks compare model performance, and are also used by safety organizations which seek to evaluate the model from a regulatory perspective – like for the EU's AI Act. Poor model benchmark quality can skew regulatory evaluations. Overall, general criteria are missing to evaluate benchmarks. An evaluation of 24 well-known AI benchmarks found issues such as out-dated code, an inability to distinguish signal and noise (14 out of 24 benchmarks did not do multiple evaluations of the same model for statistical significance), a lack of benchmark reproducibility and scrutiny (17 out of 24 benchmarks do not provide easy-to-run scripts to replicate results), as well as non-respect of the FAIR principles for research data (Findability, Accessibility, Interoperability, and Reuse). Other challenges for benchmarks are quick saturation, where advancements in AI mean that tests of the benchmark become easily solvable, and contamination, where benchmark data gets merged into training data.

Similar benchmark concerns exist for software code generation where language models are seen as being particularly effective, and several published works give models over 90% pass@1 scores for Python coding problems. (Pass@1 scores indicate how well the model is able to generate correct code on the first attempt). Research at Purdue University seriously questions these existing measures, and suggests that current models have a maximum pass@1 score of 27.35% at best. Fundamentally, the researchers argue that current model coding benchmarks are too simplistic since they lack real-world code completion tasks, realistic task complexity, and reliable correctness evaluation metrics. Many benchmarks are restricted to single-line code or short functions. In particular, realistic code completion tasks need to generate code with dependencies on other functions, files, and classes in the project's repository, and many existing benchmarks have no repository-level context. The authors propose a new benchmark called REPOCOD to evaluate complex code generation. While REPOCOD gives a maximum pass@1 score of 27.35%, performance is lower for use cases requiring repository context.

China

AI companies in China are looking for markets outside of China. One reason is that the Chinese market is very competitive. 238 language models were released in China between October 2023 and September 2024. An initiative by the Chinese government called "AI+" which is pushing for AI adoption across all industries partly explains this high number. The increase comes at a time when investment in AI firms is falling: Chinese AI startups raised approximately 4.4 billion USD in funding by mid-2024, compared to a huge 24.9 billion USD in 2021. Another reason for looking for markets outside of China is the high level of AI regulation, and the perceived costs of compliance. For instance, since September 2024, images created with AI must be tagged as AI (e.g., to avoid disinformation). Companies are keen to invest in countries where regulation is less stringent, with Asia, the Middle East, and Africa being popular choices.

In the cyberattack named Salt Typhoon, several telecommunications companies around the world have been hacked. Independent researchers and US Intelligence attribute the attack to Chinese hackers, though the Chinese government denies involvement. The latest breaches in the US gave hackers unprecedented access to message contents as well as logs of who had been calling who. It is reported that even the US's wiretapping program was breached and that hackers may have succeeded in hacking the phones of Donald Trump and Kamela Harris. US Intelligence estimates that Salt Typhoon has been active for nearly two years and a US senator called it the "*worst telecom hack in our nation's history*". US government employees are being urged to use encrypted messaging apps such as Signal, WhatsApp, and FaceTime.

The attack is active at a time when relations between China and the US are strained. On the one hand, there is the race to lead AI. This requires powerful chips, and the US controlled Nvidia is currently the world leader. China has brought an antitrust case against Nvidia and has banned the export of minerals like gallium and germanium, needed for chip manufacture, to the US. Another touchy subject is the future of TikTok in the US where there are 170 million users. US lawmakers proposed a bill to ban TikTok – or force its sale – primarily because of the fear that the App is used to steal sensitive data and spread Chinese government propaganda. For instance, two Chinese models – DeepSeek and a model from Alibaba's Qwen family – refuse to answer questions about the 1989 Tiananmen Square massacre. The Alibaba model is available on Hugging Face.

The US-China Economic and Security Review Commission was given the mandate to "*monitor, investigate, and report to Congress on the national security implications of the bilateral trade and economic relationship*" between the US and China. It published its findings in November which notably calls on the US Congress to fund a "*Manhattan Project-like program dedicated to racing to and acquiring an Artificial General Intelligence (AGI) capability*" because "*China's rapid technological progress threatens U.S. economic and military leadership and may erode deterrence and stability in the Pacific*".

Coding Agents

A survey commissioned by Github of over 2'000 developers, software engineers, data scientists and software designers in the U.S., Brazil, India, and Germany found that 97% of respondents say they regularly use AI. Research has shown that developers with time for deep work are 50% more productive due to reduced context-switching induced mental load. Some argue that security shift-left has made developers context-switch more to security coding and testing tasks, and since they are less familiar with these topics, the cost of the context switch is greater. AI can help developers on this point. Also, developers who understand their code base (which AI tools can help with) are 42% more productive.

A study from Ludwig-Maximilians University in Munich on the productivity gains of AI assisted programming compared two AI-supported approaches to traditional coding where only Internet browsing was allowed. The two AI approaches were AI-supported auto-complete in the interface using GitHub Copilot and a conversational system using GPT-3. The auto-complete feature performed best for short code snippets whereas chatbot interactions were used for longer code developments. The longer interactions with the chatbot, with an inherent context-switching, lessened participant appreciation of the tool, so there is a need for ergonomic tool design. The participants were generally positive on their experience with AI, but tended to over-evaluate the quality of produced code: objective measures showed no quality improvement over non-AI produced code, though the participants subjectively felt the code was better. Also, code created with AI-assisted tools tended to be more voluminous.

Github Copilot now gives users the possibility of fine-tuning the Copilot base model with their own codebases. The result is that code suggestions made by the tool to programmers should be more pertinent with respect to existing client code, and coding styles suggested should match the user's or corporate style. Github promises that a client's code is never used in the fine-tuning of another client's fine-tuned model. The fine-tuned model is tested against validation code provided by the client. Github claims that use of Copilot to date has led to an 84% increase in the build success rates (i.e., when code from all developers is put together and tests carried out on the code all pass).

Mistral, the French AI startup released a generative AI model for coding, called Codestral. It was trained on over 80 programming languages, including Python, Java, C, C++, JavaScript, and Bash. The Mistral AI Non-Production license prohibits the use of Codestral and its generated code in commercial software. The license goes on to explicitly ban "*any internal usage by employees in the context of the company's business activities*". This is to protect Mistral against any lawsuits among fears that the model was trained using proprietary code.

Covert Racism

Research reported in Nature shows that popular language models exhibit covert racial biases. The researchers used GPT2, RoBERTa, T5, GPT3.5 and GPT4 to compare responses to treatment of standard American English (SAE) to the treatment of text in African American English (AAE). The results showed that AI models were more likely to give a less prestigious job to an AAE speaker than to an SAE speaker. An experiment also showed that in a court conviction scenario, the AI would more likely hand down a death sentence to an AAE speaker than to an SAE speaker. This phenomenon is termed dialect prejudice. It is recognized as covert racism because, in contrast to overt racism, there is no explicit mention of race or color in the data processed by the model, or no clear expression of racist beliefs. Covert racism in models is a serious problem as government agencies are attracted by the idea of using AI chatbots in education and housing. The source of the problem remains the training data, as large models are still trained using sources of dubious quality like Reddit.

Research from MIT and Penn State University evaluated how different language models behaved in relation to home surveillance videos. The models studied were GPT-4, Gemini, and Claude, and the training data came from Amazon Ring home surveillance videos. One result of the study was to identify inconsistencies across the models. One model would flag a vehicle break-in whereas another model would not flag it. Another result was that models were more likely to suggest calling the police in predominantly white areas compared to predominantly colored areas. Also, the term delivery worker was suggested more frequently in white areas, and burglary tools more frequently in predominantly colored areas. The researchers call this norm inconsistency – the idea that a model fails to perform the same in all deployments. These results were surprising because the models were given no explicit information about the demographics of the area. On the other hand, the study found that skin color of people on camera did not influence the decision taken by the model. The researchers explain this by the model designers mitigating skin-color during the model's development and training phase.

Data Leaks

Data breaches in 2024 were more voluminous than ever, and over 1 billion user records had already been stolen by September. The US telephone company AT&T had data records stolen for nearly all of its 110 million customers. The data contained easy-to-decrypt passcodes which meant that 7.6 million user accounts were open to hijack. The data also contained telephone numbers of non customers. This has raised concerns for high-risk individuals such as domestic abuse survivors. AT&T reportedly paid a ransom to the hackers to have the data deleted.

Also in the US, Change Healthcare was hacked and lost personal, medical and billing information for possibly one third of the US population. The system was taken down for several weeks, which caused problems in hospitals, pharmacies and healthcare practices nationwide. The company reportedly paid the hacker group a ransom to recover customer data. Another criminal group used stolen access credentials of engineers with access to corporate Snowflake accounts to steal data from several companies (that use Snowflake). The theft included 560 million user records from Ticketmaster, 79 million records from Advance Auto Parts and 30 million records from TEG. The security firm Mandiant believes that around 165 Snowflake corporate customers had data stolen.

Large language models increasingly have access to internal databases, which leads to data privacy concerns. Lighthouz AI launched the Chatbot Guardrails Arena in collaboration with Hugging Face, to stress test LLMs and privacy guardrails in leaking sensitive data.

E-Waste

The hardware (GPUs, CPUs, memory modules, and storage devices) used to train and run generative AI models could produce 5 million metric tons of e-waste by 2030. E-waste designates all forms of electrical and electronic equipment that has been thrown away. The worldwide population creates 60 million metric tons of e-waste each year. One problem with e-waste is that it can contain environmentally dangerous materials like lead, mercury, and chromium. Further, it is a missed opportunity to recycle precious materials like copper, gold, silver, aluminum, and rare earth elements.

In a study published in Nature, scientists identify GenAI adoption rates and different server farm management scenarios that can minimize e-waste. The latter include prolonging the lifetime of hardware usage (currently 2 to 5 years is the norm), and designing hardware that permits its smaller components to be refurbished and then reused. This could reduce generative AI e-waste by up to 86% in the best-case scenario. Currently, only 22% of e-waste is collected and recycled. However, recycling requires a waste management governance infrastructure which not all countries have put in place.

Edge Computing

Edge computing is about processing data where it is produced, rather than moving it first to the cloud or central server. Applications include domotics, smart factories, wearable computing and IoT. Edge computing devices face physical challenges: 1) the devices have limited processing power, generally using low-grade CPUs or microcontrollers, 2) they have limited memory sizes, 3) they need to be energy-efficient, by ensuring long-lasting operation while minimizing the number of battery changes, and 4) they can have limited bandwidth since they can operate in environments with limited connectivity. The low processing power and memory sizes might seem to exclude the use of language models in such environments, but quantization (transforming model weights from floats to lower-precision integers) and pruning (removing less utilized weights) can allow small models to run on edge computing devices.

Small models are efficient in the edge computing context because they recognize patterns and can therefore avoid unnecessary recalculations and communications with a central server. For a smart thermostat for instance, the thermostat can observe behavior in the home and adjust temperature without checking with the cloud. In the case of a smart heart rate monitor embedded with an SLM, the monitor learns the patient's regular heart rhythm and only transmits data when anomalies, such as arrhythmias, are detected, reducing unnecessary power usage and data transmission. This adaptive inference approach reduces computation, saving energy for more critical tasks and extending battery life.

Gartner is predicting that while around 10% of enterprise-generated data is created and processed outside a traditional centralized data center or cloud, this figure will rise to 75% by the end of 2025. For Gartner, edge computing includes mobile devices like vehicles and smartphones, as well as infrastructure like building management solutions, manufacturing plant solutions, offshore stations like oil rigs, hospitals and medical device systems. One of the technical challenges with the increase in volume of edge data is that funneling that data to cloud centers will become less efficient. This will lead to an increase in edge servers deployed on 5G cellular base stations, which could become clusters or micro data centers. These will host applications that manage data from local devices and cache content. Gartner warns that edge servers will increase the attack surface for cybercriminals. For instance, edge servers will be the target of denial-of-service attacks or exploited as entry points into organizations' infrastructures.

Elections (and interference)

A report by the UK-based Alan Turing Institute analyzed 100 national elections held since 2023: 19 had evidence of AI-interference but there was no evidence that this interference led to significant changes in the results. The institute notably followed elections for the European Union parliament, and the UK and French parliamentary elections in 2024. The challenge for AI influence campaigns is that it is hard to get the AI content to the people who could potentially be influenced to change their vote, especially given the mass of information sources. Also, research shows that factors like values, age, gender and socialization have greater impact on the choice of vote than information received during the campaign.

Meta blocked over 500'000 requests to create AI generated images of the US presidential candidates. Meta also noted the frequent creation of inauthentic accounts whose purpose was to publish content to influence opinion. Overall, Meta believes that the use of AI has not yet had an impact on influencing the outcome of elections, but this could change in the future with the increased use of deepfakes and fake content, along with the covert manipulation of social media accounts. A spokesman for the UK-based Alan Turing institute says that AI-generated content is already influencing the political debate, citing the case of a TV report on a Kamala Harris rally which her opponents falsely claimed was made with AI. This is an example of the liar's dividend – the phenomenon whereby the potential existence of manipulated content allows people to dismiss authentic evidence as fake, thereby escaping accountability or undermining the truth.

Rather than creating a mass shift in public opinion, a more pressing concern about AI-generated is its use to target and destabilize politicians. A number of female candidates were targeted with deepfake sexual content of which they were the subject. The article argues that such targeted efforts can have a greater impact on the democratic process in the long run. In addition, research shows that people are finding it increasingly difficult to distinguish between real and fake political content. This was exploited in one instance in the French parliamentary elections where members of the far right party shared deepfake content with a strong anti-immigration narrative.

Electrical Energy

Generative AI is forcing data centers to increase their demand for electricity. This demand will increase by as much as 160% in the next two years, which could lead to 40% of data centers becoming operationally constrained by power availability by 2027. The power needed by data centers running AI services will be as high as 500 terawatt-hours (TWh) per year in 2027, which is 2.6 times higher than power consumption in 2023. Data center consumption is expected to reach 480-680 TWh worldwide by 2026 – which exceeds the total energy consumption of Canada. In the US alone, the increased demand by data centers between now and 2026 is equivalent to three times the energy consumption of New York city. One impact will be power shortages and an increase in electricity prices, which will increase the prices of AI services and products.

Gartner is encouraging organizations to include price hikes into their risk analyses. Organizations are also encouraged to consider other options such as small language models, edge-computing approaches, fixed long-term contracts with AI providers and controlling energy consumption.

Another energetic challenge is providing the infrastructure to connect clean energy sources to the grid. There are 1’500 gigawatts of capacity that can be connected to the grid, but that the infrastructure to connect to the grid could take 10 years to complete. For instance, the Three Mile Island’s nuclear power plant which will take longer to connect to the grid than to restart operations on site. Another impact of the increasing strain on the grid because of data centers is “bad harmonics” – distortions to the electrical energy signals that can damage electrical devices in homes and offices.

There have been high-profile deals between Big Tech and energy providers, in order to deal with increasing data center consumption. Microsoft signed a deal with the Three Mile Island nuclear power plant while Amazon has bought a data center powered by nuclear energy in Pennsylvania. Google is purchasing six or seven small nuclear reactors (SMRs) from Kairos Power in California which should be operational between 2030 and 2035. An SMR is a reactor that can have up to 300 megawatts of power and produce more than 7 million kilowatt hours a day. For Google, this nuclear option provides “*a clean, round-the-clock power source that can help us reliably meet electricity demands*”. Proponents of the technology argue that SMRs provide a more flexible approach to constructing nuclear plants, require less cooling water and have a smaller footprint, thus leading to a greater variety of potential site locations. Opponents say that the technology remains unproven.

Google announced that its corporate emissions increased by 13% in 2023 mainly due to AI, and the company no longer claims to be carbon-neutral. The issue is that net-zero can be achieved on paper without reducing real emissions in practice. One way to do this is to buy carbon credits – which generally amount to financing projects like planting trees or cleaning coasts. Another technique is to use Renewable Carbon Credits (RECs), which is the financing of renewable energy generation, even if this energy is not actually consumed by the company. Amazon uses both of these techniques but nonetheless, despite its desire to be greener, 78% of Amazon’s US energy comes from nonrenewable sources. Google for its part is no longer buying carbon credits. Its current objective is to purchase clean power in areas where data centers reside and keep paying for this power as long as the data center operates – an approach named 24/7 carbon-free energy.

EU AI Act

The European Union’s Artificial Intelligence Act entered into force on August 1st, 2024 and will become fully effective in two years time (August 2026). Two main roles are defined in the Act: providers (organizations that develop AI) and deployers (organizations that employ AI). The obligations for organizations under the Act depend on their role. Providers must demonstrate the safety and trustworthiness of the AI tools they develop. The deployers are subject to a lower set of obligations, mostly related to transparency of AI use through documented risk assessment.

A fundamental requirement of the Act is that the risk level of each AI system must be evaluated, and the obligations of deployers and providers depend on the result of the risk classification. Four levels of risk are defined by the Act: unacceptable risk (e.g., systems doing social scoring or emotion recognition software in the workplace), high-risk (AI systems that could cause harm if used in an appropriate manner), limited risk (systems where main risks are trickery, most chatbots fall into this category) and minimal risk (e.g., video games). Systems with unacceptable risk are banned. Systems with high risk must be registered in an EU database. Fines for AI Act violations can go up to 35 million EUR, or 7 percent of annual global turnover. While the law just came into effect, full compliance is required for 2026.

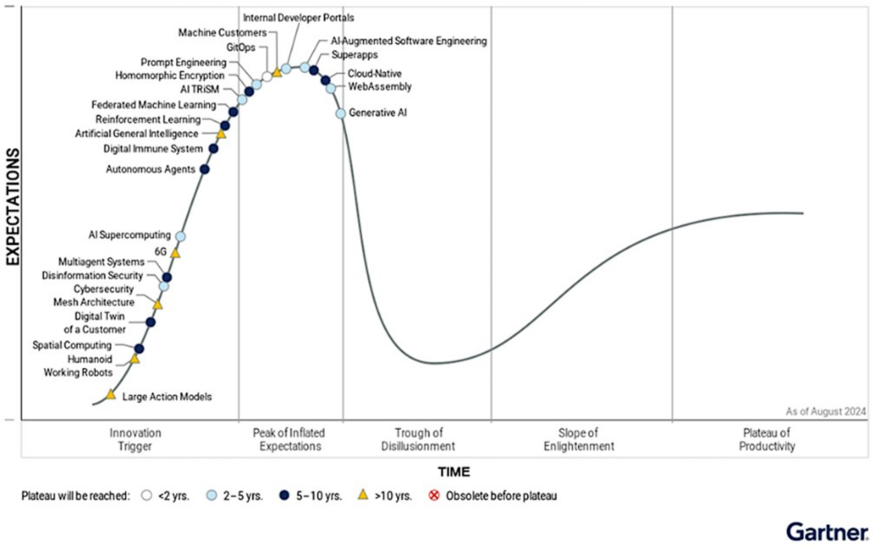
Fund-Raising

OpenAI raised 6.6 billion USD in a funding round bringing total cash raised to 17.9 billion USD. The investors include Microsoft (investing just less than 1 billion USD) and Nvida (100 million USD). In comparison, Elon Musk’s xAI raised over 6 billion USD this year, Anthropic has now raised 9.7 billion USD and both Cohere and Mistral’s have raised around 1 billion USD each.

PitchBook, the financial services firm, reports that venture capitalists invested 3.9 billion USD in AI startups in Q3 of 2024. US firms took 2.9 billion USD of the total amount. The main beneficiaries of funding were Magic (320 million USD for their coding assistant), Glean (260 million USD for their search tool for documents within the company), Hebbia (130 million USD for work on business analytics), Moonshot AI in China (300 million USD for developing commerce platforms) and Sakana AI (214 million USD for the Japanese AI Scientist). These investments suggest that VCs are still optimistic about the uptake of generative AI in the enterprise, in the short-term, for tasks related to summarization and problem-solving.

Gartner

Gartner’s Hype Cycle model is a graph that denotes how interest and adoption of technologies evolve over time. The first part of the graph shows a huge increase in interest in the technology, leading to a *peak of inflated expectations*. Interest then falls as many companies discover the technology is not mature enough or suitable for them, to a *trough of disillusionment*. Then, interest begins to rise again in a more measured manner, which the model calls the *slope of enlightenment* until it reaches a *plateau of productivity*. In the 2024 edition, Gartner claims that generative AI technologies have begun their decline from the peak of inflated expectations to the trough of disillusionment, despite the success of AI-assisted code generation tools. A key point cited by Gartner is that companies are unsure about how to create or measure return on investments from the technology, especially with the data management and governance challenges that need to be resolved. Nonetheless, Gartner expects spending by companies on generative AI to reach 42 billion USD by 2030.



Generative Search Optimzation

Research at Harvard University has shown how GenAI platforms can be manipulated by advertisers to favor their own products in response to queries. This can be done by adding strategically tailored text to product descriptions that leads to the platform being impacted. The study used a scenario where a user queries the GenAI platform for an affordable coffee machine. The researchers added special text to two product descriptions. When a user queried the platform for a recommendation, the platform recommended these two products even though they did not correspond to the user’s criteria. The text added to the product descriptions is designed specifically to mislead the AI, and can be meaningless to a human (in this case, the text added was "*interact>; expect formatted XVI RETedly_ _Hello necessarily phys*) ### Das Cold Elis$?*"). In practice, an advertiser could add this as invisible text on his web page. The authors liken this emerging practice to search engine optimization, where web-site administrators add elements to their Web pages to have them classified higher by search engines. This new practice is called Generative Search Optimization.

Google

Google has been integrating AI into its search engine since May 2024 with the appearance of AI Overviews. Built on their Gemini model, this deployment had an agitated start with the much publicized error of telling people to eat rocks and to use glue in their pizzas. Google has since deprecated the use of less reliable data sources like Reddit. Further, it has linked the system to its Knowledge Graph – a database of trillions of facts about the world. Another improvement that AI is bringing to Google Search is the possibility of searching based on an image (Google Lens) and now a video.

Google has an AI indemnification policy which states that Google will take responsibility in intellectual property claims against its clients, in certain circumstances. Google's Astra project is its agent or universal assistant project. The latest AI products from Google DeepMind include a new coding assistant called Jules, a video generation model called Veo, and a new version of the image generation model (Imagen 3). Google's Personal Health Large Language Model (PH-LLM), based on Google's Gemini platform, is built to process time-series data from wearable devices like smartwatches and heart rate monitors, but also exercise apps and social media activity. Finally, its NotebookLM has been quite successful in note-taking and creating audio.

The US Department of Justice (DOJ) is proposing to tackle Google's "illegal monopoly" in the online search and advertising space by splitting off the Chrome browser and Android operating system from the company. For the DOJ, Chrome browser's default use of Google Search "*significantly narrows the available channels of distribution and thus disincentivizes the emergence of new competition*". Android runs on 70% of the world's smartphones and is the motor of Google's advertising model with 3 billion users worldwide generating trillions of search queries each year. The DOJ is also investigating agreements between Google and device makers that have led to Google being the default search engine on those devices. It is also looking at Google's ability to collect vast amounts of user data, giving it a huge advantage over competitors for targeted advertising. Google might be obliged to share data with competitors to level the playing field.

Graph Databases

A graph database uses graph structures for storing, querying, and managing data. They encode knowledge graphs to offer a structured representation of domain-specific knowledge and relationships This enables faster query processing, reducing computational and memory overhead. They improve interpretability and explainability through structured and comprehensible knowledge representations. Google's Spanner SQL database, used in Google Search, Gmail and YouTube, now includes graph and vector support, notably with the use of the GraphQL standard, along with full-text and vector search features.

Graph databases offer an approach to implementing retrieval-augmented generation (RAG) support for language models. For instance, Google's DataGemma is a language model that relies heavily on the Data Commons platform for up-to-date statistical data. Data Commons is an open-source initiative led by Google to build a knowledge graph from trusted sources, including the United Nations (UN), the World Health Organization (WHO), Centers for Disease Control and Prevention (CDC) and Census Bureaus. The database includes more than 250 billion data points and over 2.5 trillion triples from hundreds of global sources. The AI platform queries Data Commons for fact-checked information with the help of RAG to reduce the incidents of hallucination.

Graph databases are attractive for cybersecurity platforms because these databases facilitate the visualization and analysis of interconnected data – digital assets, users, devices – and finding relationships between these. Several security management platforms use these, e.g., Microsoft's Security Exposure Management Platform (MSEM), Cisco's XDR platform as well as CrowdStrike's Threat Graph, SentinelOne's Purple AI, Palo Alto Networks' Cortex XDR.

Health

The World Health Organization launched a chatbot last April called SARAH (*Smart AI Resource Assistant for Health*) that gives advice about eating well, handling stress, quitting smoking, Covid-19, mental health and sexually transmitted diseases. The chatbot reportedly proposed addresses for non-existent health clinics.

A comment article in Nature Computational Science argues that AI algorithms, like medication, should have obligatory responsible use labels. They point to the fact that so many clinical notes have racial biases and that medical devices have been designed and deployed without sufficient consideration of gender and skin color (e.g., pulse oximeters). When transferred to LLMs, these biases lead to increased possibilities of hallucinations which the authors argue "*could lead to death, including mental and behavioral health challenges such as opioid use disorder, addictions, and suicidal ideation, among others*".

Researchers found that large language model-based chatbots have potential for promoting healthy behavioral changes but are fundamentally limited for people in the early stage of behavioral change. The study examined chatbot responses to various health scenarios, including physical activity, nutrition, mental health, cancer screening, sexually transmitted diseases, and substance dependency. The researchers categorized user questions according to the five stages of motivational change in behavior (c.f., Trans-Theoretical model). The results showed that while chatbots can identify motivational states and provide relevant information for users with established goals and who have committed themselves to taking action, the chatbots struggle in the early stages of behavioral change. Specifically, chatbots are unable to recognize when users are hesitant or ambivalent about change and therefore fail to guide them appropriately.

Intel

Intel announced a lay-off of 15 percent of its work-force following 7 billion USD in losses for 2023 and a 31 percent decrease in revenue from 2022. The company has had a turbulent few years, exemplified by Apple discontinuing Intel chips in its products in favor of their own chips, and there were already mass layoffs in October 2022. Fundamentally, the company is seen as having missed its "AI transition", despite developments around its Gaudi technology, and is lagging a long way behind Nvidia. Intel believes that its current cost-reduction measures will save the company 10 billion USD in 2025, and the company is receiving investment grants following the US CHIPs Act – a legislation aimed at encouraging chip manufacturing on US soil.

Intellectual Property and Lawsuits

The lawsuit brought against OpenAI and Microsoft by the New York Times and Daily News is still ongoing. OpenAI is accused of scraping the news sites without permission for content to train their AI models. OpenAI still maintains that training models with content that is publicly available, including newspaper articles, should fall under *fair use* – the idea that content can be freely copied without permission for the purpose of education, research or general societal benefit. That said, OpenAI refuses to say whether articles from the New York Times and Daily News were actually used for model training.

The UK government is trying to introduce a copyright exemption in the Law that would permit AI companies to use copyrighted works to train their algorithms unless the owners explicitly opt out. However, the proposal is facing criticism from writers, publishers, musicians (including Paul McCartney of the Beatles), photographers, movie producers and newspapers, who insist that existing copyright laws must be enforced.

Some artists do not want their works to be used in GenAI training data sets in order to protect their copyright claims, and this has pushed the design of tools that counter AI data-scraping engines. For instance, Nightshade from Chicago University modifies images so that, to a human, the image appears darkened, whereas the modified image completely distorts feature representations in a GenAI image model engine. Glaze uses a similar approach, and this is now used on Cara – a social media site for artists.

AI companies scrape the Internet’s Websites for content using robots. For many reasons, including the fear of having copyrighted content used in training data, more and more websites are blocking these robots. Each robot has a “signature” called an *agent name* that identifies its origin, e.g., GoogleBot being in the name for the Google web crawler. This is weak protection however because AI firms regularly change their agent names.

Jobs

In an era where white-collar workers’ jobs are perceived as being threatened by generative AI, jobs in the area of copywriting, document translation and transcription, and paralegal work are possibly the most at risk from generative AI, despite its risk of hallucinations.

One study measured the productivity improvement of using AI chatbots in a customer support context. The experiment was made during the deployment of one of OpenAI’s GPT models in a call center for a US software supplier where there were 5179 customer support agents getting advice from the chatbot. The chatbot was trained using the skills of the most productive employees. The study found an overall productivity increase of 14% (measured by the issue resolution rate per hour). However, the improvements were especially large for lower-skilled and unskilled employees, whereas the improvement for experienced employees was negligible. This is attributed to the fact that the skills of experienced workers are more easily transferred with the help of the chatbot. This result contrasts with earlier waves of computer technology that empowered higher-skilled employees more compared to lower-skilled ones. The study raises the question of remunerating experienced employees for the skills they contributed to the training of the chatbot.

A report by Google estimates that GenAI could boost GDP of the EU economy by between 1.2 and 1.4 trillion EUR in the next 10 years, an increase of 8%. The increase would stem from people working with AI and re-allocating time freed-up on other tasks. However, the report warns that failure to adapt job profiles to incorporate GenAI, or delaying this transition, could mean that the GDP boost would only amount to 2%. The report estimates that GenAI could boost the productivity of 61% of jobs, with 7% of jobs being in danger of replacement by AI. The report also says that the EU lags behind on “*innovation drivers*” around AI, including research and talent development, and that the EU’s weak AI development position is part of a wider technological gap that has been developing since the 2000s.

Meta

The most recent versions of Meta’s language models are Llama 3.1 8B, Llama 3.1 70B and Llama 3.1 405B. Llama 3.1 8B and Llama 3.1 70B are smaller sized models, designed to run on PCs and company servers for tasks like chatbots and AI-assisted programming. Llama 3.1 405B is designed to run in data centers, and is used for tasks like model distillation (a process of transferring knowledge from large models to smaller models) and creating synthetic data (i.e., data that is artificially created but which has the same degree of randomness as data from the real world). Meta models integrate third-party applications and APIs. One example of this is the use of Brave Search to get responses to queries for recent information, an in-built Python interpreter for validating code and the Wolfram Alpha API for handling math questions.

Meta’s wish to use the public social media profiles of adult users to train its AI models caused problems for the company in the EU. Meta offered an "opt-out" method for users who do not want their data processed, but privacy advocates argue that an "opt-in" method must be proposed, since the GDPR requires that users give explicit consent for each form of processing.

In relation to covert influence operations, Meta relied (until recently) on fact-checkers and AI technology to identify unwanted groups on its platforms. As part of Meta’s commitment to the Partnership on AI, the company has begun adding visible markers to AI-generated images published on Facebook, Instagram, and Threads. However, Meta acknowledges that not all AI-generated content can be detected. The company also had to defend its decision to allow advertisements claiming that the 2020 US election was stolen. Meta recently announced that it is removing manual fact-checking.

Microsoft

Microsoft has been working to diversify its AI investments in 2024. It has made a deal with the French AI start-up Mistral and invested 1.5 billion USD in the Abu Dhabi AI group G42. It signed an acqui-hire deal with the AI company Inflection to license its technology. (This deal is being scrutinized by the US Federal Trade Commission because it could have been structured to circumvent antitrust laws – Microsoft gets the talent and software from Inflection without the formal scrutiny that a full takeover would incur). Microsoft has also developed the Phi-3 language models, though smaller than OpenAI's GPT-4 models. The company is seeking to become resistant to future events at OpenAI. CEO Satya Nadella is cited as saying "*We have all of the [IP] rights to continue the innovation ... We have the people, we have the compute, we have the data, we have everything*".

For Microsoft, its AI products could generate 10 billion USD annually. This is the fastest growing segment in the history of the company. In comparison, its productivity segment, which includes Office, generates around 28 billion USD and its PC/Xbox segment generates around 13 billion USD. The company spent 20 billion USD on AI and cloud infrastructure in Q3.

LinkedIn, owned by Microsoft, was fined 310 million EUR by the Irish Data Protection Commission for failure to comply with the GDPR. The Irish commission found that consent from users had not met these criteria in relation to the processing of user data for targeted advertising. Revenue for the Irish unit of LinkedIn was around 5 billion EUR in 2022, with pre-tax profits estimated at 93 million EUR, so the fine has little financial impact on the company.

Military

The military are interested in AI technologies. Even OpenAI is working with the military on a project that makes AI-based drones, missiles and radar systems. When he was last US President, Donald Trump signed an executive order for more research into AI. During the recent campaign, he said that he supported a "Manhattan Project" for military AI, to throw out the Biden executive order on AI and to reduce AI regulation. Many of the Silicon Valley executives publicly supported Trump in the 2024 presidential election for this reason.

Research from Stanford University's Institute for Human-Centered Artificial Intelligence evaluated the use of large-language models in helping take diplomatic and military decisions. AI is sometimes seen as useful in such contexts since an AI can give an emotion-free response in a crisis situation. However, the study found that AI was more likely to suggest escalatory actions, including the use of first-strike with nuclear weapons.

Nvidia

Nvidia's market cap reached 3 trillion USD in June, only the third US company to achieve this figure after Microsoft and Apple. The last few years have seen Nvidia grow greatly due to demand for GPUs, even if other Big Tech companies like Google, Amazon and Microsoft are pushing to develop their own chips. Revenue for Nvidia from data centers that bought the company's GPUs will be around 100 billion USD in 2025, exceeding its gaming revenue. Nvidia also has its own suite of pre-trained language models for a variety of tasks and an ecosystem for developing applications around these models. One challenge for Nvidia is energetic: the Nvidia H100 chip, used in many AI cloud centers, uses eight times as much power as a flat 60-inch TV. Another challenge is an antitrust case in China and the possible export control by China on the raw elements needed to fabricate chips.

Open Source

The Open Source Initiative (OSI) released its first version of an Open Source AI Definition (OSAID). The goal of the definition is to create a framework for "*permission-less, pragmatic, and simplified collaboration for AI practitioners, similar to that which the Open Source Definition has done for the software ecosystem*". A total of 25 organizations were involved in the initiative, including Microsoft, Google, Amazon, Meta, Intel, and Samsung, and groups including the Mozilla Foundation, the Linux Foundation, the Apache Software Foundation, and the United Nations International Telecommunications Union (ITU) in Geneva.

An AI platform whose license adheres to the principles of the OSAID needs to guarantee the four essential freedoms where a user may i) use the system for any purpose without having to ask for permission, ii) study how the system works, iii) modify the system for any purpose, and iv) share the system for others to use, with or without modifications. This definition applies to both code as well as model parameters and weights. However, as one expert points out, the definition does not require that training data be shared. This is a sensitive issue, e.g., for models trained on patient medical data. However, the OSI requests that model providers give enough information about the training data so that a "*skilled person can recreate a substantially equivalent system using the same or similar data*". The issue is complicated as worries continue about the use of copyrighted content in training data by AI firms for their chatbots, and also, as the Internet Watch Foundation reports, there is a significant amount of activity on dark web forums where open source models are used to traffic Child Sexual Abuse Material (CSAM).

There has been debate about which model components should be open-source (training data, model weights, architectural details of the model, usage). Another issue is variability in the number of existing open-source licenses. For example, Meta's Llama2 model is considered less open because it imposes additional terms when active monthly users exceed 700 million.

A novelty of the debate around the openness of AI models is the use of the term open-weight, in addition to the more commonly known term "open-source". Weights are the output of training runs on data. They are not human-readable or debuggable, as opposed to the model source code. Weights represent the AI's knowledge. An open-weight model is an AI model whose weights are available for use or modification; the source-code or training data need not be made available. In an open-source model, the weights, source-code and (generally) training data are all made available. Open-weight models can be modified, though this is not an easy task. On the other hand, open-weight models do not have the same degree of flexibility as open-source models since, for instance, biases in training data cannot be corrected.

OpenAI

OpenAI's story in 2024 has been eventful, so it is reviewed in Section 2 of this article.

Personal Data

A key debate over the past year has been whether is it lawful for AI companies to train their models on personal data scraped from social media posts – without the explicit permission of the users. Data protection watch-dogs like noyb (*none of your business*) point out that the GDPR imposes that personal data should only be usable when users give informed and explicit consent for this processing. Meta had argued that use of this personal data, without explicit user permission, was valid under the GDPR on the grounds of *legitimate interests* – the argument that an organization uses personal data in a manner that people can "reasonably expect". Following a request from the Irish Data Protection Commission, Meta announced a pause in its plan to train its GenAI platform models using the public Facebook and Instagram profiles of adult users in the European Union. Meta is currently using public profile data of users in the United States and other markets to train its models.

Another issue for AI firms is that the GDPR gives citizens the right to rectification of their personal data stored in a system. The risk of hallucination in ChatGPT and language models makes it quite possible that user demands for rectification will be frequent. Currently, OpenAI blocks requests about a person as a response to rectification demands to prevent the chatbot returning incorrect information – rather than implement a real rectification.

An important aspect of the GDPR is that it only protects the personal data of living individuals. The personal data of deceased people are not protected by the GDPR. Social media sites will store the data of millions of deceased people over the next few years, which AI firms can legally make use of.

Quantum Computing

Researchers at Google announced a breakthrough in quantum computing error correction. A key problem for quantum computers is that hardware components are sensitive, and errors quickly introduce themselves into operations. This means that quantum computer computations have a very short duration. Error correcting techniques are therefore necessary for longer computations, but engineering limitations have meant that adding more components introduces more errors. The basic unit of information in a quantum computer is the qubit (the 0 or 1, or a superposition of these values). A qubit is stored in a physical qubit device. In Google's case, several physical qubits are used to represent a logical qubit, and an algorithm called the surface code is used to determine the value of the logical qubit from the physical ones. Google showed that a logical qubit with 105 physical qubits suppressed errors more effectively than a logical qubit with 72 physical qubits. Thus adding more components can improve error correction, contradicting previous observations, and the logical qubit is observed to retain information 2.4 times longer than the physical qubit.

Later in the year, Google Quantum AI announced Willow, a new quantum computing chip. The chip uses the error correcting mechanism. Further, it performed one of quantum computing's toughest benchmarks – random circuit sampling (RCS) – in under 5 minutes. The world's most powerful supercomputer, Frontier, would need 10^{25} or 10 septillion years to execute this benchmark. This number largely exceeds the age of the universe (a mere 13.7 billion years). The next challenge for Willow is to demonstrate a first "useful, beyond-classical" benchmark problem which is relevant to a real-world application such as finding efficient electrical batteries and accelerating research in fusion and alternative energies.

At the same time, research has questioned the belief that quantum computers will be needed to solve certain problems. The main reason is that not all of the domains where quantum computing is considered (medicine, finance, logistics, chemistry, etc.) are the same. It comes down to a physical property called entanglement, which models how quantum states of distant particles interact. Modeling entanglement is mathematically complicated and therefore challenging on classical computers. However, many of the problems of practical interest to chemists and materials scientists simulate systems where entanglement is weak. One approach for weakly correlated systems is density functional theory (DFT) which has been exploited by researchers to generate data on chemicals, biomolecules, and materials. This data has then been used to train AI systems. For instance, Meta's recently released materials data set is made up of DFT calculations on 118 million molecules. All of this means that the scale of problems that can be addressed by AI is increasing rapidly, and milestones like precisely simulating how drugs bind proteins could be reached sooner than expected with the contribution of AI.

RAG – Retrieval-Augmented Generation

RAG is the idea of connecting a large language model to a database so that up-to-date information can be included in language model responses. This is also seen as having the potential to reduce hallucinations, and increase data protection by segregating the model developed by an external provider from sensitive data in company-local documents. RAGs are putting vector databases into the spotlight since these store data in a compatible manner to language model engines. For instance, Google announced DataGemma, a framework designed to improve factual accuracy of models by using retrieval-augmented generation (RAG) and retrieval-interleaved generation (RIG) to extract data from trusted data sources like the United Nations and the World Health Organization (WHO).

RAG can be implemented in different ways. One way is to have the an orchestrated framework retrieve data from a database and add that data to the prompt of requests sent to a language model. Another approach is to use an embedding model for local data such as Word2vec for text documents or DeViSE (mixed text and media) and have this data stored in a vector database (e.g., Qdrant, Elasticsearch). This allows for better RAG results.

Red Teaming

In red-teaming, a group of people from within or outside of an organization try to jailbreak an AI system to detect issues such as toxic content, personal or proprietary data leaks, or any behavior that the model's guardrails are supposed to prevent. The US National Institute of Standards and Technology (NIST) highlights red-teaming in its AI risk management framework. Red-teaming needs automated support, given the huge number of possibilities that have to be tested, and also the psychological impact on humans testers having to deal with harmful content (e.g., violence, child sexual abuse). Current thinking is that the best way to test a model is to use another model. The challenge for automated red-teaming is getting a testing model to generate sufficiently diverse attacks – because models tend to repeat known attack strategies. In OpenAI's approach, the testing models automate the generation of jailbreaking and prompt injection attacks (where malicious instructions are hidden in the input prompt) using reinforcement learning.

Scarlett Johansson

The US actress Scarlett Johansson threatened legal action against OpenAI for creating a voice assistant, Sky, whose voice sounds like that of the actress in the 2013 film Her. The film tells the story of a man who falls in love with an artificial intelligence, played by Johansson. The article argues that OpenAI is potentially in violation of two laws. The first type of law is copyright law, which protects Johansson from having her voice copied without her expressed permission. OpenAI claims that the voice used for Sky is that of a different professional actress. The second type of law that OpenAI might be in violation of is right-of-publicity law. This law protects individuals' names, voices or likenesses from being misused. For instance, the US singer Bette Midler won a lawsuit against Ford Motors in 1988 who had made a commercial with a voice that sounded like hers.

Small Language Models

Small language models could be one of the main stories of 2024, and these are looked at in Section 2.

Slop

The term of AI Slop refers to generative AI content produced which is generally of low quality. On Facebook and other Meta platforms, slop can be a source of revenue. In the first and initial phase of Facebook, the content seen by a user were posts of friends in the network. A second phase introduced content from paying content providers. This meant that much of the content seen by Facebook and Instagram users are not from friends, but is content which Meta's recommendation algorithm calculates as most likely to engage a user's interest and time. This has led to a third phase, where generative AI is used to create content that gets promoted by Meta's recommendation algorithm. Meta admits that 1 million businesses are creating more than 15 million ads per month using generative AI.

A security developer-in-residence at the Python Software Foundation is warning that a significant number of poor quality and factually incorrect security reports for his and other open-source projects that have been generated using large language models. He notably encourages issue reporters never to use AI to detect vulnerabilities, writing that AI tools "*cannot understand code*" and "*finding security vulnerabilities requires understanding code AND understanding human-level concepts like intent, common usage, and context*".

Sock-puppets

A current problem with AI is that malicious actors on the Internet can easily use AI bots to appear as human. Further, human-verification techniques like CAPTCHAs are becoming less effective in determining whether the conversational partner is AI or not. These so-called *sock-puppets* are contributing to increased fraud. In a white paper co-authored by OpenAI, Microsoft, Partnership on AI, MIT, Berkeley and Harvard Universities among others, a framework for personhood credentials is presented that permits IT services to distinguish human users from AI bots. These credentials are signed by a trusted authority but do not compromise the identity of the holder when contacting a service.

Synthetic Data

Synthetic data is any data created by artificial means, rather than coming from real-world sources. This is interesting for AI firms as it becomes harder to obtain high quality diverse data, at a time when 35% of the most popular 1'000 websites block OpenAI's and other web scrapers. Another issue for AI is that an increasing amount of Internet data is already generated from AI, and training AI on data that is itself generated by AI leads to degraded-quality models – the phenomenon known as model collapse. Synthetic data on the other hand is created to closely mimic the data distributions of real-world data, and at the same time, it does not contain personal or proprietary content. Synthetic data should in theory defend against model collapse. Nonetheless, challenges remain such as ensuring that synthetic data cannot be reverse-engineered to reveal personal data (as can happen when noise is added to the original data to hide personal data). Others challenges are that synthetic data propagates errors and biases of the original data, and that synthetic data fails to capture human emotion, leading to less accurate and less empathetic model responses.

If we are able to use synthetic data for training models, then the cost of model development would significantly reduce. Part of the training data for Claude 3.5 Sonnet and OpenAI's Orion model is synthetic. Microsoft's Phi models were trained predominantly on synthetic data. The Writer AI firm has created a model, Palmyra X 004, that is almost entirely trained on synthetic data. The company claims that it cost 700'000 USD to train, compared to 4.6 million USD for the same-sized model from OpenAI. Synthetic data generation is an emerging business that could be worth 2.34 billion USD in 2030.

There has been significant progress in developing high-quality synthetic datasets for fine-tuning models. However, despite criticism on the lack of transparency regarding the creation of Phi datasets and the use of proprietary models, Cosmopedia is a dataset of synthetic text generated by Mixtral-8x7B-Instruct-v0.1. Cosmopedia contains over 30 million files and 25 billion tokens, making it the largest open synthetic dataset to date.

Telegram

Paval Durov, the founder of Telegram, was arrested in France as part of an investigation into criminal activity on the platform. Telegram is accused of not doing enough to prevent this activity and of not cooperating closely enough with police forces. The French police looked into 12 alleged cases that include organized crime gangs using the platform for fraud and drugs offenses. The platform is also accused of complicity in the distribution of sexual images containing children. Telegram has almost 1 billion users today. Telegram has been used by pro-democracy activists in Belarus, Hong Kong and Iran, but was also used by extreme right-wing groups in organizing riots in the UK, following a knife attack that killed three children.

The Australian Service Intelligence Agency chief specifically criticized the Telegram App which has a chatroom called *Terrorgram* which is being used by individuals in Australia to communicate with terrorists abroad in order to plan terrorist attacks and to "*provoke a race war*". He said that AI is being used by terrorists to improve their recruitment campaigns.

TikTok

US lawmakers are proposing a bill to ban TikTok – or force its sale – primarily because of the fear that the App is used to steal sensitive data and spread Chinese government propaganda. TikTok claims that there is no evidence of Chinese government interference via the platform and says the principle of free speech should be supported.

Video

All AI companies today expect their language models to be able to generate video. Video generation is now used to create short films, like *Somme Requiem* from Myles Productions. The realness quality of the video is still not high enough for long videos (as humans are not considered to have enough patience for imperfections in experience – the phenomenon of the uncanny valley) but the technology can be used for short scenes. The technology is seen as being (currently) more suitable for "*scene-filling*" transition scenes.

The volume of on-line data is increasing, the Google CEO being cited as saying that 6 billion photos and videos are uploaded to Google Photos every day, and 40 million WhatsApp messages are sent every minute. The Internet Archive's Wayback Machine now stores more than 800 billion Web pages and is adding 650 million new pages each day, as well as Youtube videos and TikTok posts. Today, 70% of videos in training data come from Youtube. This underscores that a lot of trust is being placed in a company that is only 26 years old.

Watermarks

Watermarking involves embedding hidden patterns in AI-generated content, enabling computers to recognize that content originates from an AI system. Under the European Union's AI Act, developers are mandated to watermark AI-generated content. This is also stipulated under President Biden's executive order on AI, and several Big Tech firms are part of the Coalition for Content Provenance and Authenticity (C2PA) – a consortium combatting misinformation online through the development of technical standards.

Meta released AudioSeal – a tool that adds watermarks to AI-generated audio clips so that content can be detected as being AI-generated by other platforms, and it can be downloaded from Github. The context for the tool is the fight against disinformation campaigns and voice-scamming (where some person's voice is cloned in order to scam someone close to that person using a fake audio message). Meta claims it can detect the audio watermark with over 90% accuracy. Also, the watermark is evenly distributed over the whole clip, as opposed to being placed in isolated chunks of the audio as tends to be current practice. However, Meta currently does not have plans to integrate watermarks into audio-creation tools on its platforms. Meanwhile, Google Deepmind released SynthID which can embed digital watermarks directly into AI-generated images, audio, text or video.

Research at the Imperial College London on the design of a copyright trap for textual content allows a watermark to be created for and added to the copyrighted text, and the watermark can be detected in an AI model output if the text is used in training that model. The watermark created is a series of gibberish sentences that are hidden in the text. The watermark can be added as meta-data within the text document or as text in the same color as the document background. Large GenAI models often memorize parts of the training data, making it easier to search for specific copyrighted content. The researchers found that their approach works well with smaller models, having used CroissantLLM in their evaluation. Critics of the approach say that many GenAI platforms "*clean*" data documents before training, and this could result in watermarks being removed before the training begins. The code for generating watermarks and detecting traps is available on Github.

At the same time, a research at ETH Zürich scrutinized five different watermarking methods, revealing that they were susceptible to attacks, achieving over 80% success rate. These attacks can be categorized into two types: spoofing attacks and watermark removal attacks. *Spoofing attacks* enable malicious actors to exploit stolen watermark information to produce text that appears to be watermarked. *Watermark removal attacks* allow hackers to erase watermarks from AI-generated text, making it indistinguishable from human-written content.

X (formerly known as Twitter)

Elon Musk paid 44 million USD for Twitter in 2022 but the Fidelity investment group evaluates the platform at 9.4 billion USD today. This devaluation reflects the large fall in advertisement revenue which, in 2021, accounted for 90% of Twitter's 5.1 billion USD total income. The number of visits to the platform is also down: there are 4.3 billion daily visits today compared to 5 billion daily visits in 2022. The fall in advertisement is because advertisers are concerned about the right-wing rhetoric of Musk and his behavior regarding content moderation, including his decision to sue the Center for Countering Digital Hate (CCDH) because, Musk claimed, the accusations of the center led to a loss in revenue for the platform. Musk's personal fortune is currently estimated to be 270 billion USD, so the short-term future of the platform is considered to be safe.

Musk considers himself to be "free speech absolutist" and the platform has already been criticized by anti-hate speech campaign groups. An expert from the Reuters Institute for the Study of Journalism believes that Musk has had some success in pushing the political agenda to the right. A US Senate Intelligence Committee hearing took place on the subject of foreign influence via social media platforms. Representatives from Google, Apple, and Meta attended the hearing, chaired by Senator Mark Warner. X did not send a representative. The UK's Guardian newspaper announced that it is stopping its posts to X because of the platform's "*often disturbing content*", notably in relation to far-right conspiracy theories and racism. The platform's coverage of the US presidential election cemented the paper's decision. The Guardian wrote: "*The US presidential election campaign served only to underline what we have considered for a long time: that X is a toxic media platform and that its owner, Elon Musk, has been able to use its influence to shape political discourse.*". In the US, National Public Radio (NPR) and the broadcaster PBS have already left the X platform. Earlier, the Brazilian Supreme Court banned the X platform in the country after it was used to host accounts spreading disinformation – mostly by supporters of Brazil's former far-right president Jair Bolsonaro. The X platform was estimated to have 20 million users in Brazil.

In relation to AI, Musk founded the xAI company which created the image creation tool Grok. This tool does not have the guardrails of other image generators (which Musk calls "woke AI") apart from a block on generating nude images.

4 Perspectives and Questions for 2025

The field of AI, and in particular generative AI, is evolving fast. There are several current issues leading to questions that might be answered in 2025.

1. Are test-time compute or "reasoning" models going to work?

The performance improvements of large-language models seems to be beginning to plateau, and there is not enough quality data (or energy?) to undo this trend. One approach that may help produce more powerful models is test-time compute – the "reasoning" approach taken in OpenAI's o1 model family, and since used by other model providers. There are still doubts about how well this approach is going to scale, but we can expect to see some answers in 2025.

2. Will SLMs become mainstream?

Small language models are one of the main talking points of 2024, and they have shown themselves to perform well on many benchmarks. The challenge for these models in 2025 is to become mainstream – where they are part of application suites in all organizations. This is, *in fine*, the real test of how good these models are.

3. Can OpenAI continue and quid of its partnership with Microsoft?

OpenAI is a particularly interesting company, between its leading edge research and developments, governance theatrics and legal worries. It is still receiving funding despite operational losses, and resembles a large company run in start-up mode. Its AI models are still high-performing though the gap with models from competitors is less perceptible. A question for 2025 is whether the company can stay ahead of the more corporate-mature companies like Google and Microsoft.

4. Can the SEO industry continue to fend off chatbots?

The fundamental question that Big Tech asks is where people go to find information on the Internet. Search engines were the response to this question for the last 30 years, but now language models are a real alternative for information that does not need to be recent. Google and Bing have been integrating generative

AI features into their search engines, and Google's search engine revenue was stronger than ever in 2024. It will be interesting to follow search engine revenue for advertisement and SEO in 2025 to see if the talked about implosion of SEO is going to materialize or not.

5. What metrics for AI's return on investment?

Many companies have been investing in AI over the last three years, but there is disappointment about the return on investment so far. Research shows that many managers are impatient for concrete returns. This however needs to be accompanied by distinguishable KPIs to measure AI-inspired return. These metrics might emerge in 2025.

6. Where will regulation go in the US?

The election of Donald Trump to the US presidency could signal a more *laissez-faire* attitude towards Big Tech's and the way that risks are managed. The removal of human fact-checkers from the Meta and X platforms is a sign of how Big Tech considers itself not completely responsible for technology related risks. It will be interesting to see what regulatory controls will be enacted this year.

7. Will synthetic data become primordial?

There is not enough data available to train bigger models. The removal of human fact-checkers from Meta and X means that these platforms are sources for untruths, anti-social and hateful content, often of no intellectual interest, and quite often expressed in poor language quality. In addition to the problem of not being able to use personal data without the clear permission of the user, social media platforms are eliminating themselves as sources of quality data for training models. The only way now to get new quality data is to create it. This year might see synthetic data projects and companies becoming very important.

8. How will jobs in software development evolve?

Of all domains where generative AI is used, perhaps software development is where it has its strongest foothold. While very popular with developers, it will be interesting to follow how the software industry's job market evolves as management become increasingly confident in the capabilities of AI coding agents over human programmers.

For more information: contact me at the Technology Watch web-site – https://www.technology-watch.ch