Council of the European Union
General Secretariat

Analysis and Research Team

2023
24 April

ART - RESEARCH PAPER

ChatGPT in the Public Sector – overhyped or overlooked?

Executive Summary

The unprecedented success of **AI tools like ChatGPT** caught many by surprise. But large language models (LLM) are here to stay and will continue to grow in sophistication. These models use natural language processing algorithms to interpret and respond to text-based human input. Whilst it is possible to grasp the basic principles which drive models such as ChatGPT, the companies behind them - mostly US-based - are becoming increasingly coy about releasing detailed information on the code and parameters which determine the way they generate their outputs. That makes it more challenging to assess the implications and impact of integrating large language models into the workplace. At the current rate of expansion, it's only a matter of time before such models are **integrated into the public sector** with **wide practical applications, advantages, and possible efficiency gains**, from 24/7 availability to managing large volumes of inquiries simultaneously.

But there are also **limitations**. While sophisticated Al such as ChatGPT may seem extremely intelligent, capable, and reliable, this is not a wholly accurate picture. ChatGPT certainly has some capabilities at a speed and scale that humans do not, but it sometimes provides responses which are inaccurate, biased, or nonsensical. Its purely mathematical approach to reasoning should not be mistaken for human-like intelligence.

If ChatGPT and similar tools become part of daily workflows, this trend will also affect public institutions. By providing services which are instrumental to the functioning of the State and affecting the rights and obligations of citizens, the **public sector is particularly sensitive** to the introduction of such AI-based technologies. Public administration has its own characteristics and principles which distinguish it from the private sector. By extension, the key principles of public administration such as accountability, transparency, impartiality, or reliability need to be considered thoroughly in the integration process.

To benefit from the advantages offered by ChatGPT and similar tools, **risks** should be recognised, managed and, where possible, **mitigated**. Whilst some of the existing limitations will be addressed by technological advances, others, such as biases, are of a more structural nature and cannot be fully corrected. Measures are therefore needed to ensure that the appropriate procedures and human controls are in place as well as the establishment of feedback loops from the citizens and independent audits.

In the absence of clear regulation on ChatGPT accountability, humans are needed to monitor output especially when considering **what lies ahead.** And only humans can provide the personalised services, flexibility, emotional intelligence, and critical thinking needed to fulfil the requirements of public service.

Introduction

The release of ChatGPT in November 2022 has prompted a lively debate about its strengths and weaknesses, its potential, possible risks, and its occasionally bizarre responses. To cut through these discussions, it is important to step back and realise that large language models (LLMs) such as ChatGPT have been part of our daily lives for some time: Google searches in English, for example, have been supported by AI technology since 2019. Yet the practical and political implications of ChatGPT remain difficult to pin down, and it is important to avoid underestimating or overestimating the capabilities of LLMs. There are also important questions around who is developing this technology and where. These issues are critical to the process of integrating of LLMs into the workplace, particularly for the public sector.

This note takes ChatGPT as the basis for assessing the impact of increasingly sophisticated language models on the public sector and the principles on which it is based. It takes an objective and factual look at the technology behind ChatGPT. It also highlights possible risks and opportunities this could create for the public sector both now and in the future.

To improve our understanding of what LLMs like ChatGPT might mean for the public sector, the paper begins by looking at how ChatGPT works, who is behind it, and what differentiates it from other language models and chatbots. This leads into an assessment of the extent to which language models could help support the work of public servants, but also their potential risks and pitfalls. It then looks at how these might be mitigated. The paper ends by assessing the future development of LLMs and their possible implications for the future of the public sector.

A glossary in the Annex contains explanations of some of the technical terms that appear throughout the paper.

1. What is ChatGPT

ChatGPT is a form of Artificial Intelligence which can process and produce natural language, and which is capable of a large range of text-based tasks. ChatGPT does not itself understand the meaning of the text it produces. Its responses are based on statistics and probability, but it has been sufficiently fine-tuned to make them appear to originate from a human source. It is important to remember that it does not have the ability to process and understand meaning in the way humans do[1].

> *"ChatGPT does not itself understand the meaning of the text it produces. Its responses are based on statistics and probability"*

How did we get here?

Language has always been at the heart of artificial intelligence (*see fig.1*). Ever since AI was conceived in a computational context, most notably by Enigma codebreaker Alan Turing in the 1950s[2], the 'intelligence' of an AI programme has been measured by its ability to imitate human language capabilities[3]. Until the 1980s, the development of language models (and AI more broadly) aimed at not merely *imitating* human language but actually *grasping* it. In the late 1980s, the approach shifted towards probabilistic models, which produced output based on calculations informed by datasets and which led to the emergence of statistical language models[4].

The 21st century has seen exponential growth in the sophistication of artificial intelligence, especially since the rise of 'Big Data' in the 2010s. Big Data refers to extremely large data sets containing different types of data which can be analysed computationally to reveal correlations, patterns, and trends within a particular population. Coupled with the increasing market dominance of US-based major tech companies such as Amazon, Apple, Google, and Meta, which collect data through their platforms, these datasets have grown massively, and constitute an unprecedented resource pool.

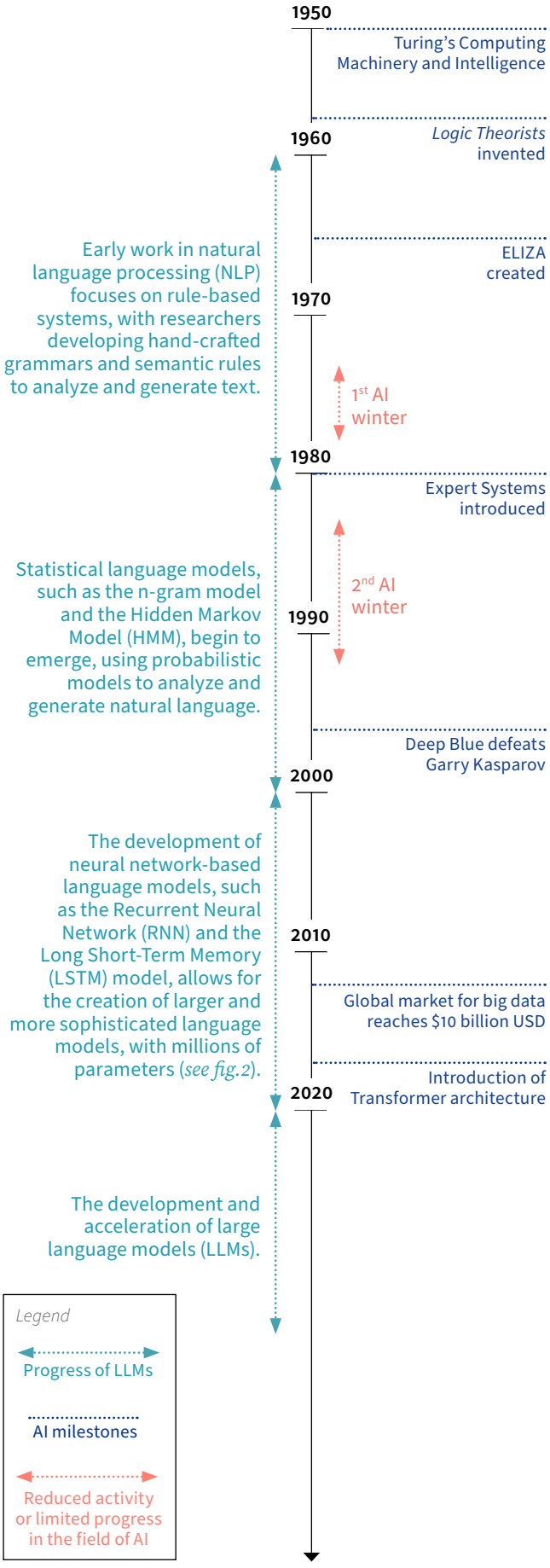


Fig. 1: Exponential growth of artificial intelligence development

However, it was the development of deep learning techniques[5], particularly neural networks and later transformers[6], that enabled the creation of much larger and more sophisticated language models. The text generated by these models is generally coherent and nuanced, often indistinguishable from human writing or speech. This raises ethical questions about the conflation of human-based and computer-based communication.

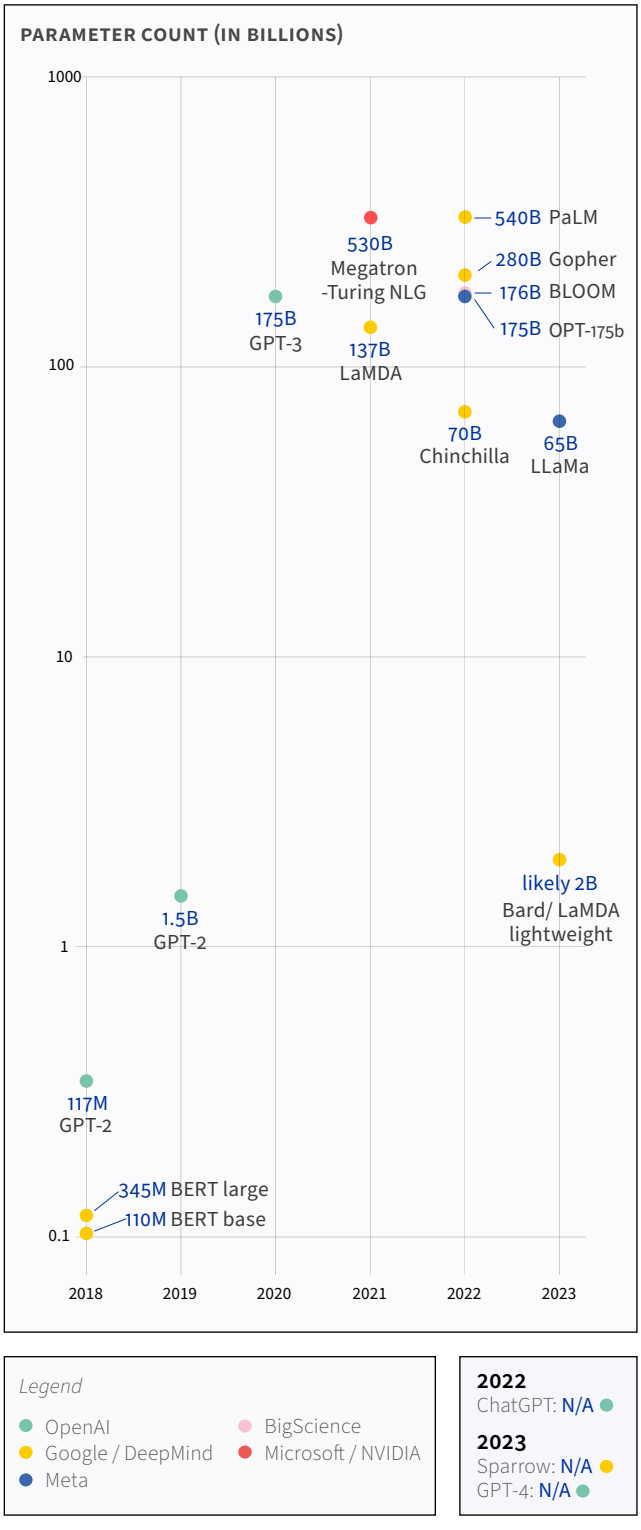


Fig. 2: Parameters of selected models in millions (M) or billions (B)

How does ChatGPT work?

ChatGPT is an AI chatbot[7] overlaying a Large Language Model (LLM), a type of machine-learning model designed to process natural language (*see fig.3)*. These models are a part of what is called general purpose AI systems, which can perform a range of general tasks such as translating, detecting patterns, or answering questions[8]. LLMs use large quantities of text to infer the relationship between different words in these texts and use this information to generate their own human-like texts. At their most basic, LLMs work in a similar way to predictive texting on a mobile phone. They determine which words are statistically likely to appear after each other and use this knowledge to predict the next word.

However, these simpler models are unable to grasp more nuanced connections within the language they are processing. When faced with multiple similar options to generate the next word in a sentence, they are often unable to choose the most logically appropriate one[9]. Models like those used by ChatGPT, so-called Transformers, can solve this issue with a mechanism called self-attention[10]. Self-attention allows these models to subdivide their input data into small segments (tokens) and assign them values based on the importance of each segment for the overall meaning of the input sequence. During its training, the model uses these values to generate billions of normalised weights which provide a basis on which it can calculate the most accurate response to the input it receives. ChatGPT's model performs these calculations several times in a row, which enables it to recognise more subtextual meanings, such as humour. However, this also explains why it can end up generating different answers to very similar questions: varying the position of a word in a sentence triggers a different sequence of calculations, which in turn leads to differently worded answers.

LLMs can be classified both by their parameters and the amount of data used to train them. Parameters are the weights and biases the model has learned during its training. They drive its decision-making processes, like neurons in a human brain. During the training phase these parameters are constantly adjusted and updated by a self-learning algorithm within the model. As a rule of thumb, the more parameters a model has, the broader the range of tasks it can perform[11]. The more training data it has been fed, the more precisely it can adjust its weights, and the more accurately it will respond to a user's query[12].

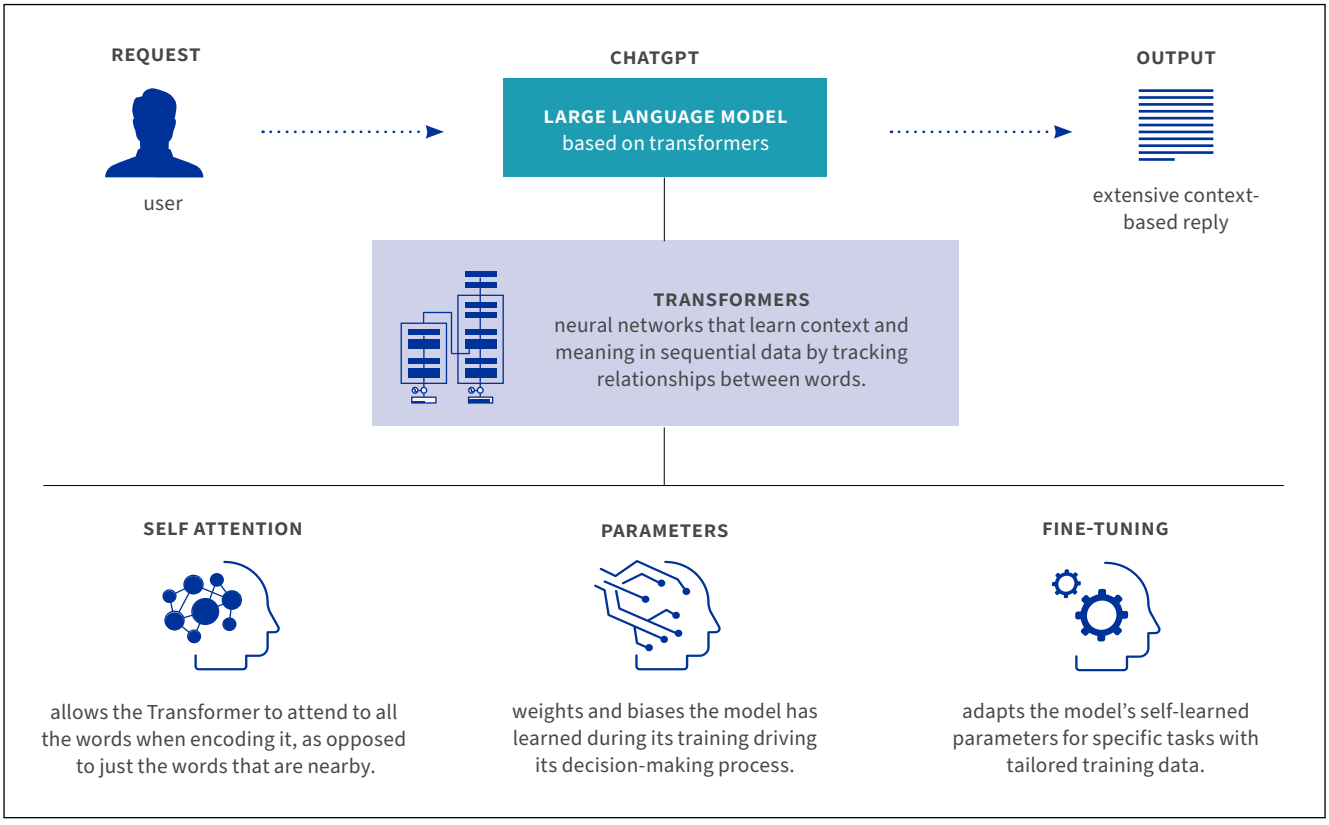


Fig.3: Functioning of ChatGPT

After some initial training, LLMs are often fine-tuned to generate responses which are more closely aligned to the user's intent. Fine-tuned models can be based on larger models with more general capabilities. For example, the model behind the initial version of ChatGPT is based on an existing LLM that has been optimised for creating more natural-sounding conversations[13]. What differentiates ChatGPT from other language models is the way human feedback was used during the fine-tuning process to ensure the output is more closely aligned with the intentions of the user[14].

In addition to text-generating models such as ChatGPT, there are now also LLMs specialized in generating images or even videos from written input. Techniques from large language models are also used to complete and generate spoken sentences[15].

Which companies are behind it?

ChatGPT was created by a company called OpenAI, an AI research and deployment company (*see fig.4*). Founded in 2015 by Sam Altman, Elon Musk[16] and several former researchers from other AI companies (*see. fig.5*), OpenAI aims to "ensure that artificial general intelligence (AGI) - by which we mean highly autonomous systems that outperform humans at most economically valuable work - benefits all of humanity"[17]. Initially a non-profit research lab, OpenAI has since created a business subsidiary[18], which enabled it to enter an exclusive commercial partnership with Microsoft in 2019, a deal that included investment of $1 billion. In January 2023, both companies announced the extension of their partnership, with Microsoft investing another $10 billion[19].

OpenAI has developed several AI products, most notably its series of generative pre-trained transformer language models, GPT-1 -to -4, and several other more specialised models such as the image generation models DALL-E and DALL-E 2. It also offers an application programming interface (API), which enables paying customers to develop their own applications on top of its models and which provides its own tools[20].

Despite its proclaimed intention to develop AI for the benefit of all humanity, OpenAI has become much less transparent in recent years. Since its change of structure from a non-profit organisation to a commercial subsidiary, it is no longer obliged to publish the salaries of its principal employees[21]. While GPT-2 was open-source and was released with code and model weights[22], its successor, GPT-3, and all subsequent models have remained proprietary[23]. This means that little is known about the architecture of the model behind the initial version of ChatGPT, OpenAI's most successful model to date, and the company refuses to publish detailed information about the code and parameters which underpin the decision-making processes of its recent models.

What does the competition look like?

OpenAI is not the only company developing and deploying large language models at scale. In fact, even though GPT-3, the model on which ChatGPT is based, was by far the largest language model when released in 2020, it is no longer either the largest or the most accurate. The original ChatGPT launched on 30 November 2022 is using a model called GPT-3.5-turbo, which is an improved version of the original GPT-3 model[24]. The details of its architecture have not been revealed by OpenAI, but some sources assume that it may have as few as 20B parameters[25].

Nevertheless, developing a language model like ChatGPT can be a complex process. Since training and running a large language model can be very costly, not many actors can afford to develop them[26]. Big tech companies and a few well-funded start-ups therefore dominate the market.

Partly in response to the high costs of running ChatGPT, OpenAI introduced a $20/month subscription plan called *ChatGPT Plus* on 1 February 2023. Subscribers gain preferential access to ChatGPT as well as priority access to new developments. On 14 March 2023, OpenAI launched a new, and supposedly much improved model called GPT-4. The release was accompanied by detailed documentation on safety-testing GPT-4, but in line with previous releases it did not contain any details on model size and training data.

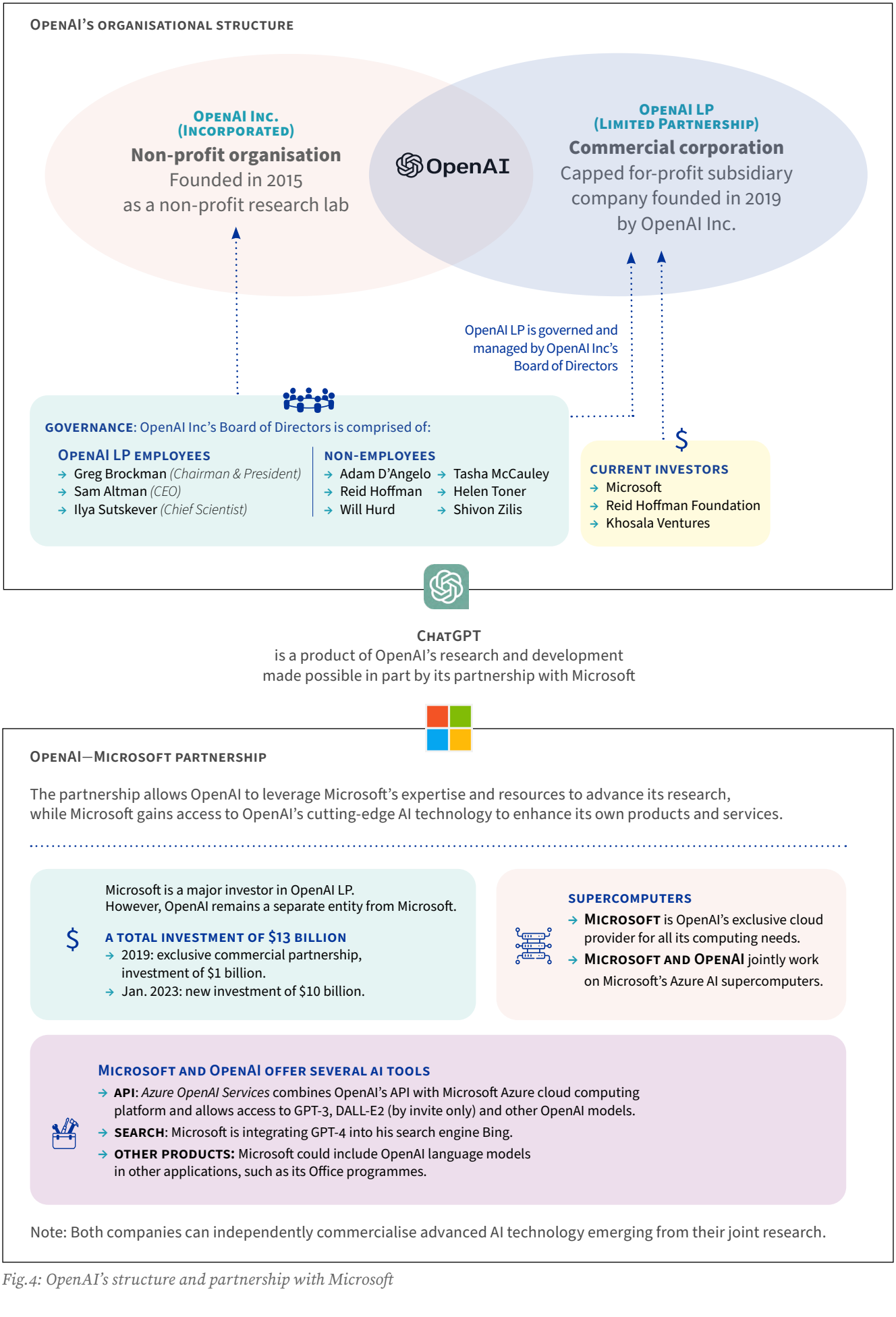


Fig.4: OpenAI's structure and partnership with Microsoft

What makes ChatGPT stand out are the underlying model's accessibility and training process. The initial ChatGPT uses one of the largest LLMs accessible to the public, and the first to be trained by using human feedback - all that is needed to use it is a free ChatGPT account[27]. The way the model has been fine-tuned to enable conversation-like interaction and even follow-up questions make it unique so far compared to other LLMs[28]. This does not mean that ChatGPT is always accurate: several pre-print articles document that even though ChatGPT performs very well when carrying out various reasoning and dialogue tasks, it struggles with more complex reasoning and mathematics[29]. Even with its elaborate fine-tuning, it still makes up information, and its built-in safety features on undesired content can be circumvented[30]. Despite these shortcomings, ChatGPT has gained rapidly in popularity. Since its launch in November 2022, the number of daily users has been rising steadily, with over 40 million daily visits by late February[31].

Until the release of ChatGPT, no other commercial company had granted access on this scale to such a versatile and accurate application. Most state-of-the-art large language models are proprietary with paid access only. This necessarily limits their availability. Large established companies have tended to be reluctant to give free public access to their AI-applications, mainly out of concern for reputational damage should their technology produce offensive or inaccurate responses. However, OpenAI's unusual strategy of releasing ChatGPT to the public has prompted a race between big tech companies on the release of their own AI to a wider audience[32]. In February 2023 alone, Microsoft introduced an AI-powered Bing search engine,[33] Google announced the introduction of its own conversational AI called Bard,[34] and Meta released their new LLM called LLaMa under a non-commercial license[35].

> *"[The LLM ecosystem] is largely dominated by a handful of mostly US-based large tech companies and a group of well-funded start-ups"*

Overall, while there are a few other players in the large language model ecosystem, it is largely dominated by a handful of mostly US-based large tech companies and a group of well-funded start-ups, whose employees often move on to create their own start-ups. ChatGPT has attracted the attention of a wider public thanks to its easily accessible user interface and the way it has been fine-tuned to respond to queries.

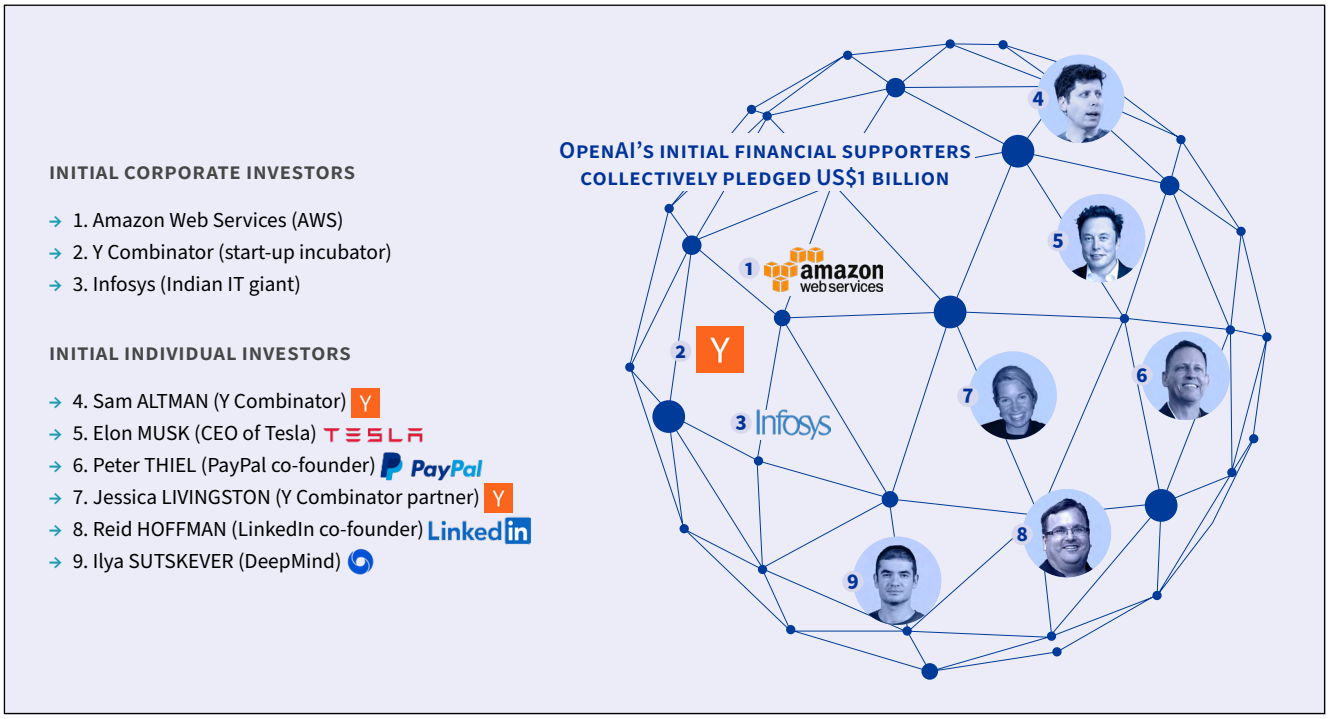


Fig.5: OpenAI's initial corporate and individual investors

What about the EU industry?

When it comes to large language models, European companies have struggled to keep up. To date, all cutting-edge large language models have been developed outside the EU[36] (*see fig. 6*). Much of European AI development is supported by academia, which can be a disadvantage in a field where most cutting-edge developments come from commercial entities (*see fig. 7*). According to a study by Large European AI Models (LEAM), an initiative of national AI associations across Europe, the US industry is up to two years ahead of European and open-source competitors when it comes to LLM capabilities[37]. LLMs released by EU member states tend to be smaller, less intensively trained models that improve or imitate existing models. BLOOM, an open-source LLM developed by a large scientific collective and trained on a French supercomputer, and Luminous, the LLM developed by the German start-up Aleph Alpha, are two notable exceptions[38]. The main reason for this is the difficulty in accessing venture capital, a lack of computational capacity, and an insufficient pool of available talent[39]. Raising capital is one of the biggest hurdles for new market entrants, which puts European competitors at a significant disadvantage. Another is the lack of commercially available supercomputers, where European countries rely much more on public initiatives than the US and China[40].

There are several ongoing initiatives aimed at improving Europe's standing on LLM. Aleph Alpha is launching Europe's fastest commercial

> *"The US industry is up to two years ahead of European and open-source competitors when it comes to LLM capabilities"*

AI data centre[41]. LEAM has produced a feasibility study supported by the German government, which estimates that building a supercomputing infrastructure suitable for AI development in Germany will cost €350-400 million[42]. The European High Performance Computing Joint Undertaking (EuroHPC JU) is a joint initiative between the Commission, European countries, and private partners to develop an ecosystem of European supercomputers[43]. It is developing eight supercomputers, one of which is currently the third fastest globally[44]. The consortium for High-Performance Language Technologies (HPLT) aims at developing multilingual training materials and train language models that support European languages[45]. In addition, the Commission has issued a €20 million tender for Natural Language Understanding and Interaction in Advanced Language Technologies through the Horizon programme, to foster a "human-centred and ethical development" of language models[46]. Finally, OpenGPT-X, a collaborative project between science, business and technology funded by the German government, builds and trains LLMs for the EU economy and intends to offer open-source versions of its models[47].

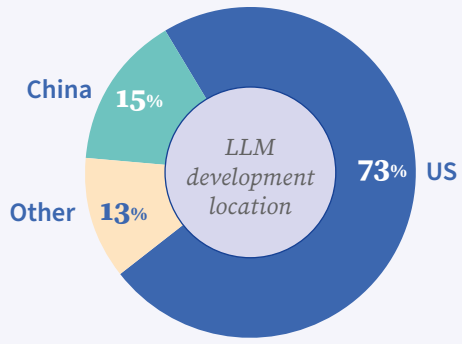


Fig.6: Geographic origin of cutting-edge large language models (LLM)

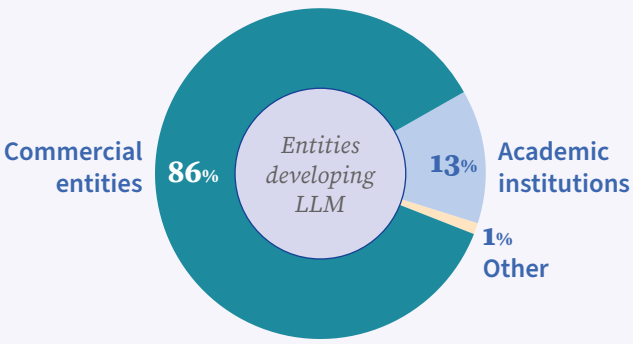


Fig.7: Type of entities at the forefront of cutting-edge large language model (LLM) development

2. What would be the impact of using language models in the public sector?

The nature of work is having to adapt rapidly to the increased use of artificial intelligence. LLMs already facilitate automated customer service, online translation, and automatic data analysis, allowing businesses to reduce staffing levels and save costs. But the public sector has a rather different set of priorities based on the principle of serving the public interest and needs to respect higher standards of accountability. The role of the public sector in providing services which support the functioning of the state, and which affect the rights and obligations of individual citizens, means that it is particularly sensitive to the introduction of new technologies based on AI. There are a range of potential use cases of LLMs for the work of public servants, but they could also affect the main principles which underpin the work of the public sector.

How could LLMs support the work of public servants?

> *"The fact that LLMs have capabilities beyond that of humans does not necessarily mean that all (or any) jobs will disappear"*

The fact that LLMs have capabilities beyond that of humans does not necessarily mean that all (or any) jobs will disappear. The dichotomy of replacing jobs through digitalisation as opposed to preserving jobs at the cost of efficiency is not quite as polarised as it appears. It is important to distinguish between the use of LLMs for specific and limited tasks, and its potential to replace entire jobs. A task-oriented approach to LLMs in public administration could enable employees to spend less time working on mundane tasks. The time saved could be used to bring a more human perspective to their work by allowing more time and energy to be devoted to more interesting tasks or exploring alternative approaches to their work.

One of the key advantages of large language models is their ability to process and analyse large volumes of data more quickly and efficiently than humans. Machinery at this level of sophistication requires human oversight and maintenance, which means jobs. Skilled workers in areas such as data analysis, cybersecurity, and technology may be able to use LLMs to take over certain tasks, but at the same time there will be a need for workers to take on new tasks and responsibilities. A 2020 World Economic Forum report predicts that by 2025, 85 million jobs will be affected (but not necessarily replaced) by machines, but 97 million new jobs could emerge as a part of this transition[48]. More recently, research by Goldman Sachs estimate that 300 million full-time jobs could be affected[49].

One potential application of LLMs in the public sector concerns chatbots and virtual assistants[50]. These models can be used 24/7 to provide a fast and efficient customer service, answer questions and address basic issues without the need for human intervention. This can help free up civil servants' time, allowing them to focus on more complex and high-priority tasks. Another area where LLMs could be applied in public administration is in document or text analysis. An LLM could be trained to identify key information in complex documents such as legal contracts, reducing the time needed for them to be reviewed by civil servants. Large language models could also be used in decision-making processes, such as evaluating grant applications or determining eligibility for social services. By processing and analysing substantial amounts of data quickly and accurately, LLMs have the potential to support fairer and quicker decision-making. However, LLMs still lack a nuanced understanding of human emotions, intentions, and context, which is why some degree of human supervision will remain necessary.

LLMs such as ChatGPT can be trained on financial and accounting data to provide insights, answer questions, generate reports, and provide financial advice based on market trends and data analysis. In the field of human resources, LLMs can assist with tasks such as the screening of CVs, candidate matching, and conducting initial interviews.

It is technically possible to train LLMs to conduct automated screening of CVs and recommend the best-qualified candidates. They can analyse large volumes of text-based data such as job descriptions and CVs to identify patterns and make predictions based on past hiring decisions. Crucially, the quality of the recommendations would depend heavily on the quality and quantity of the initial dataset, and it may not capture crucial factors such as non-verbal communication skills, attitudes to work ethics, and cultural background, as the LLM may not have been trained on a sufficiently wide range of data to capture these nuances. This means that relying solely on an AI-based system for recruitment may lead to bias and discrimination, and as such raises some ethical concerns.

In the legal sector, LLMs could be used to analyse and summarise large volumes of legal texts in order to support lawyers in the public sector. They could be trained to answer legal questions and provide legal advice, although this would require close attention to ethical and legal principles. In the longer term the use of LLMs could lead to a loss of knowledge and skills amongst legal professionals. However, it is also possible that large language models could be used to support the work of human lawyers, allowing them to focus on more complex and nuanced legal issues.

How could LLMs affect the main principles of public administrations?

The public sector has specific characteristics and is based on a number of principles which distinguish it in many ways from the private sector. There are very wide practical applications, advantages, and possible efficiency gains in integrating a tool such as ChatGPT within a public administration. However, the accompanying risks should be recognised, managed and, where possible, mitigated. The next section of this paper assesses the impact of the possible introduction of LLMs such as ChatGPT in the public sector against each of the eight **European principles of public administration** which are inspired by the standards set by the EU/OECD[51] for EU candidate countries:

→ Transparency and accountability:

accountability is about an administration acknowledging and assuming responsibility for its actions and being able to provide satisfactory reasons to justify them. Transparency[52] facilitates scrutiny and accountability.

These may be more challenging if LLMs are integrated within administrative procedures. Like many other forms of AI, LLMs are essentially 'black boxes', which means that the source code of most models is proprietary and, in any case, confusingly complex. Tech companies have for years used the commercial interest argument to resist giving access to their algorithms,[53] but even those behind these models do not fully understand their creations. Since LLMs train themselves autonomously on their datasets, and do not explain their reasoning, it is almost impossible to understand why they came to a particular result.[54] For sensitive decisions by public administrations such as those related to the attribution of social benefits, this raises critical issues around the area of accountability.[55]

→ Equality and impartiality

are about providing the same treatment to different groups of people to the same standard, irrespective of their background.

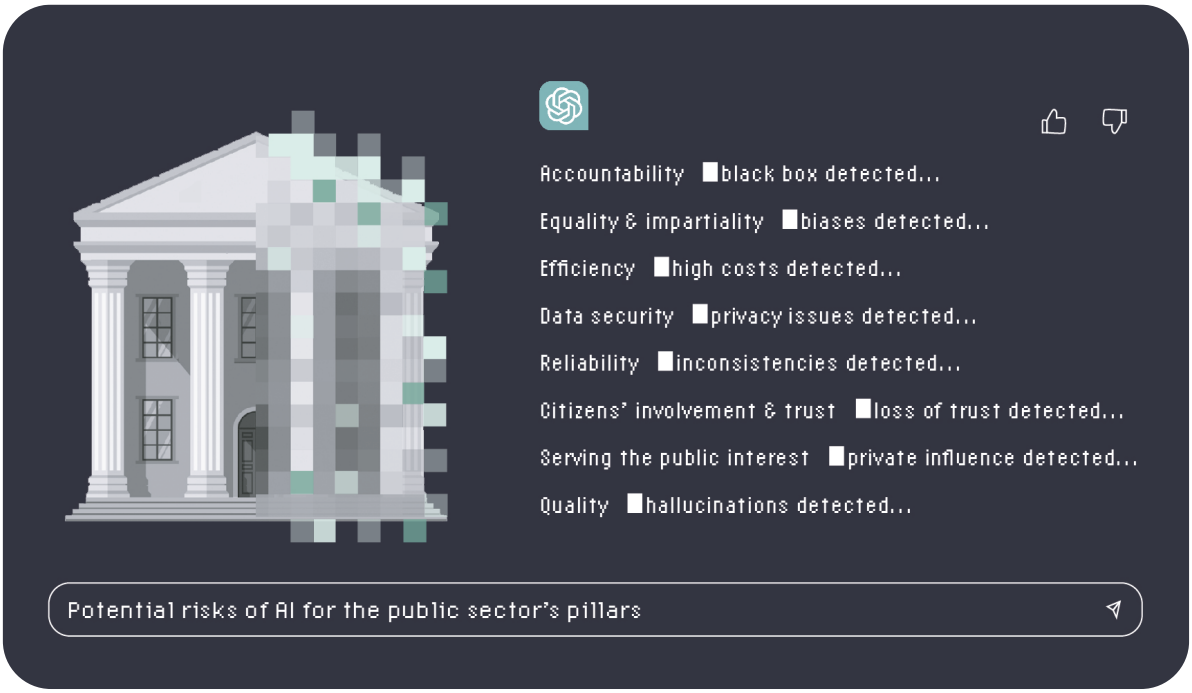
ChatGPT and GPT-4, like AI in general, are subject to biases[56] and discrimination, despite efforts by OpenAI to reduce this. Biases are not necessarily negative. They can for example be used to tailor services to specific skillsets and target audiences, or to protect the rights of minorities. On the other hand, biases can undermine the ability of public administrations to act impartially. Human agency in public administrations is never entirely free from bias, as shown by numerous studies. And according to a recent Eurobarometer poll, 74% of EU citizens also believe that there is corruption in national public institutions in their country.[57] LLMs could in theory help address ingrained bias within a public administration.[58] But even then, decisions would not be completely unbiased. Most of the biases present in LLMs originate from their training datasets, which are often based on specific sub-sections of the internet, such as content from Reddit or Wikipedia.[59] English language data dominates most datasets, while smaller languages are systematically

underrepresented.[60] Because of this dominance, the values that determine a model's reasoning are primarily based on those of a certain segment of US society. Because of the way they process input, self-taught language models can mirror and even amplify biases in the data.[61] Recent attempts to curate datasets and fine-tune models during training have reportedly led to some improvements. But the criteria used to filter datasets themselves reflect the cultural biases of the curators, who in turn represent only a small sub-set of the population.[62]

→ **Efficiency:**

is about making the most effective use of resources to deliver substantive outcomes.

LLMs already outperform humans on many more basic administrative tasks. However, these efficiency gains come at high cost: language models like ChatGPT require huge computational power for both training and operating[66]. This is not only costly but brings a heavy CO_2 footprint. For example, it is estimated that training GPT-3, the



So, whilst fine-tuning models like ChatGPT can help, it is no silver bullet. Context and personal preferences can also play a significant role in determining whether an individual finds an answer acceptable or harmful[63]. This is also an issue when decision-making in a public administration combines human and AI input, with research indicating that humans tend to follow AI-generated advice more often when it confirms their pre-existing biases[64]. In general AI tends to amplify existing power dynamics[65]. Issues of equality and impartiality therefore arise from a fundamental structural problem linked to the fact that databases and those who program them tend to reflect and confirm the current dominant worldview.

parent model of ChatGPT, has already generated around 550t CO_2[67], which does not yet include the cost of actually running the model. For the time being there are few incentives for model creators to consider the environmental impact of their models. Data on computational resources and CO_2 footprint relative to the model's capacity are not usually published, and the choice of hardware is determined more by computational abilities than emissions and energy use[68]. Some of the footprint of LLMs could be mitigated in future through evolving technology, carbon offsets, or the use of smaller models, which are much cheaper to run and train[69]. But for now, the race has been mostly focused on producing the best and most powerful language model.

→ **Quality of output:**

guaranteeing quality of output is important for the credibility of a public administration in the eyes of the population it serves.

There can be efficiency gains in using ChatGPT, such as its capacity to produce first drafts or summaries of administrative documents. According to OpenAI, GPT-4 will offer further improvements for more complex tasks requiring the handling of more subtle instructions. However, there are still limits to the reliability of the output. Firstly, whilst answers created by ChatGPT usually sound convincing, the content can be rather superficial. The quality of output is also very dependent on the exact wording of the prompt.

Secondly, despite efforts to make its technology more reliable, ChatGPT still "hallucinates" on occasions and offers inaccurate or misleading information. It also tends to provide an answer – any answer, rather than admit that it is unable to respond. (Following the integration by Microsoft of ChatGPT based technology in its search engine, Bing once claimed that "running was invented in the 1700s by Thomas Running"[70]). OpenAI claims that for GPT-4 it is possible to reduce "significantly" the occurrence of hallucinations. However, unless they can be eliminated completely, an improvement could - counterintuitively - create an even greater risk for quality of output given that greater trust may lead to less human scrutiny. In the case of ChatGPT, which does not provide sources, most claims would still need to be manually fact-checked. Some more specialised language models already provide sources within a limited context[71].

Thirdly, there are no industry-wide established standards or benchmarks for the safety and performance of LLM, which makes it quite difficult to compare the quality of individual models and choose the one best adapted for a specific task[72]. Finally, induced human dependence on technology may lead to a loss of training and corresponding reduction in the ability of humans to perform tasks which can be delegated to machines or even to a decline in the capacity for critical thinking. Without additional (human) quality control, all these elements could lead to output of lower quality and even basic errors.

→ **Predictability and reliability**:

generating trust by behaving consistently with integrity and probity over time.

LLMs trained on internal databases could serve as an institutional memory and simplify knowledge management. But issues can arise when it comes to reproducing results. Large language models can produce consistent answers, particularly in the case of short and relatively simple queries. However, for more complex input, they are generally unable to produce exactly the same answer to the same query over time. Compared to other language models, ChatGPT's abilities to recall input are remarkable. But although it is able to remember and include input of up to 3,000 words in an ongoing conversation, it cannot recall previous conversations across different threads. When asked by the ART team, ChatGPT itself indicated that the responses it generates may vary based on the input, the context of the conversation, and the training data (*see fig.8*).

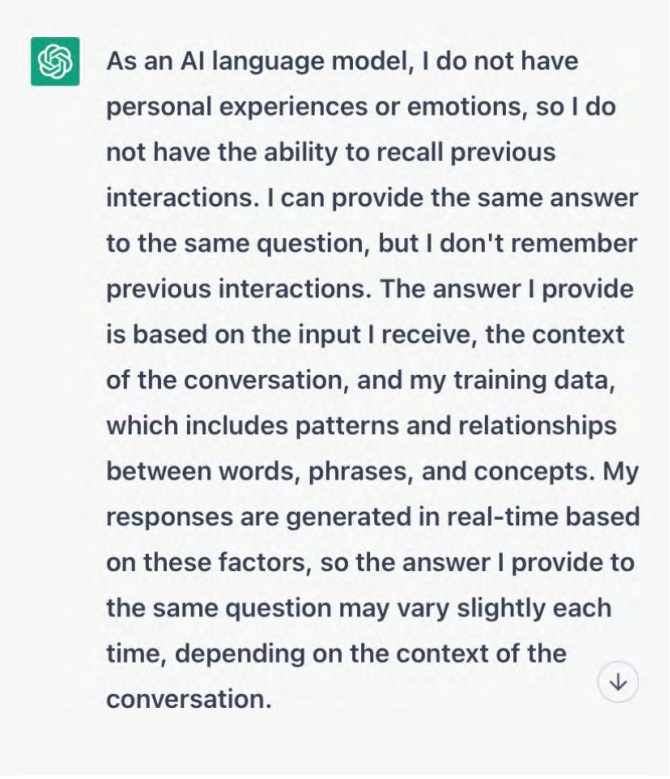


Fig.8: ChatGPT's reply on a question asked by the ART team on its capacity to generate consistent answers.

→ Citizens' involvement and trust

is the part of the democratic processes allowing members of the public to ensure that decisions affecting their lives are taken in support of the common good.

LLMs can help here. They can be used to personalize and tailor messages to citizens or MPs, such as in the context of political campaigns. This could help organisations or groups of individuals without access to public affairs professionals to be included in the decision-making process. LLMs could also support citizens who might be less comfortable with drafting letters to public administrations or elected officials.

However, trust could be undermined if citizens feel that AI is largely replacing human agency in responding to their queries or in helping shape decisions and policies of their local administration. In addition, ChatGPT could be used to boost lobbying activities through the automatic composition of input into regulatory processes such as letters to elected officials, or to generate social media posts and comments on newspaper websites. LLMs could help identify the most relevant individuals at all levels of government and target them with a flood of tailor-made messages[73]. Given the ability of LLMs to replicate closely human messages, they could also be used to support large-scale disinformation campaigns[74] and generate misleading impressions of public opinion. Current technology for determining if a text is written by an AI such as GTP Zero[75] is not sufficiently developed to allow for a reliable detection rate.

→ Serving public interest:

Public administrations are committed to provide services in the intrest of their citizens[76].

Even though large language models appear to be able to replicate general human morals and values from their training data[77], it is not clear overall whose interests they prioritise. LLMs require huge investment to finance their development and operation. Only a small number of well-funded start-ups and tech giants have access to this level of funding, with public and open-source development lagging behind[78]. This means that a small number of companies are determining the software and models on offer, and they do so with less and less transparency (as referred to above, the release of models by OpenAI has moved from an open-source and well-documented approach to one which provides little or no basic technical information). This raises the question as to whether the use of ChatGPT in sensitive areas of public administration genuinely serves the public interest, or rather the interests of its parent company or owner. Even the creators of LLMs acknowledge the risk that models could be trained to give precedence to a particular viewpoint or over-emphasise a specific perspective[79]. To date, neither the EU nor the US have dedicated legislation in force setting standards for the deployment of LLMs, nor do they have an agency exercising oversight over the development of LLMs[80], although the EU is currently looking at how LLMs might be covered within the scope of the future AI Act. Without greater transparency and oversight[81], it is extremely difficult to determine whose interests are ultimately being served.

→ Data protection and security

refers to the protection of the significant quantities of sensitive or personal information held and managed by public administrations.

LLMs such as ChatGPT and GPT-4 can be used to identify vulnerabilities in computer systems and databases, such as mistakes in lines of code, but there are some important limitations related to their tendency to hallucinate and to the size of source code which needs to be small enough to fit in the context window create some limitations. There are also a number of language models that can be used to help detect malware, such as MalBERT[82].

But there are also risks from LLMs for the protection of data. By hoovering up all information on content created while using a ChatGPT-account in a non-anonymized way and using it to train future versions, OpenAI can generate access to a huge quantity of information. This information includes queries, answers, IP addresses and location, that can be disclosed to affiliates and will be stored on US-based servers[83]. Anyone in possession of this data from a public administration could gain insights into internal discussions. In their documentation released alongside GPT-4, OpenAI states that it has been fine-tuned

on 'production traffic (with consent from users)'. This consent is in practice automatically granted by anyone signing up for an account to use ChatGPT or other OpenAI models[84].

Secondly, LLMs have been shown to memorise their training data, which can then be extracted by skilled users of the model through relatively simple techniques[85]. A future LLM trained on non-anonymised input data from previous models that can be directly linked to a public administration could in theory be used to access internal information. Furthermore, as models such as ChatGPT are trained on data that has been scraped from the internet without users' consent, it includes a lot of publicly accessible information of which citizens may be unaware - data that could also have been leaked or inadvertently shared[86]. With the right queries, adversaries can get LLMs to disclose information on an individual, such as an elected official or indeed any private citizen, without their consent.

Thirdly, LLMs, with their capacity to generate code upon request can facilitate access to the development of malware. Even though built-in safeguards are supposed to stop ChatGPT from creating malware, security experts have shown that these features can be circumvented[87].

How could the risks posed by LLMs be mitigated?

If large language models are here to stay, the question should be raised as to how public administrations can retain the advantages of using these models whilst mitigating their risks. **Mitigating the risks** that LLMs pose means first understanding their nature, potential, limitations, and their possible impact on the key functions of a public administration. It also means sharing this information with the public.

The public sector is beginning to look at a mix of measures designed to reduce the most significant risks. Most of these measures relate to the models themselves and how they are run, while others concern regulation and the way the models are used.

Measures regarding the models

→ **Public sector bodies could develop their own models**. However, although the cost of developing and training large language models seem to have decreased lately,[88] the resources required are still huge. Even with increased energy efficiency, these development costs are likely to remain beyond the means of most public administrations, not to mention the need to attract and retain qualified specialists.

→ **Models could be fine-tuned**. As the success of ChatGPT has shown, a language model does not need to be excessively large to deliver useful results. Smaller, more fine-tuned models are a possibility, even though their often-proprietary nature means that there would still be limited transparency on their exact architecture and initial training data. Most major general purpose LLM providers offer smaller versions that can be fine-tuned on datasets tailored to the needs of a client, which could lead to more accurate results. However, many fine-tuned models still require a lot of computing power, which means that they will have to be run on cloud servers that can be accessed remotely via the internet. This could create problems with sensitive internal data that needs to stay on site and would also incur costs from an external cloud provider.

→ **Models could be run locally**. Edge models are models that are deliberately small-scale and can either be fine-tuned larger models or smaller models trained from scratch. They can be run on-site and without access to the internet, which greatly improves costs and privacy, making them more suitable for the more limited means of public sector IT systems. As they do not need to transmit data to a cloud, they can even be faster than internet-connected models depending on the hardware used to run them. Because of their small size, edge models offer only a limited functionality[89]. There is also evidence that, even though the error rates are not so different from larger models, the compression techniques could amplify biases in certain areas[90].

→ **Open-source and European models could be used**. There is a wide range of open-source models currently available for fine-tuning. Providers such as HuggingFace or Streamlit can easily be used to create institutional applications. In addition, as mentioned earlier, there are already several global and pan-European research consortia working on providing more affordable open-source models that are better adapted to European needs, including pushing multilingualism in training data. But these initiatives will need more time and funding to catch up with the very significant lead from state-of-the-art models developed by US industry.

→ **Cooperative structures could be used to procure models**. Public sector institutions could develop joint infrastructure to use specialised AI like language models in a shared way, such as by running a joint cloud data centre combined with smaller hybrid edge models that only periodically access the cloud[91]. They could also jointly procure models and cloud storage with external providers. But in this case, issues related to proprietary data and infrastructure would remain, while knowledge gaps in administrations vis-à-vis the private companies on which they may depend to run cooperative projects could create space for abuse.

Measures concerning regulation

→ **Legislative action and oversight could improve LLMs**. Over the past two decades, legislation has been struggling to keep up with the breakneck speed at which new technologies are being developed and deployed. The same is true for legislation on LLMs. Currently, the industry is mostly self-regulating, with some AI companies coming up with joint initiatives and voluntary pledges[92]. There are no dedicated laws or standards on building, training, or deploying LLMs, nor on the copyright and liability for their content.

> *"Legislation has been struggling to keep up with the breakneck speed at which new technologies are being developed and deployed"*

Some developments are underway, most notably the AI Act at the EU level. The proposed legal framework focuses on the specific utilisation of AI systems and associated risks. The text proposes to establish a technology-neutral definition of AI systems in EU law and to lay down a classification for AI systems with different requirements and obligations tailored on a 'risk-based approach'. The negotiations may lead to the addition of specific provisions for general purpose AI systems, which would include LLMs[93]. Analysts are also urging governments to invest more resources in monitoring AI developments to avoid information asymetries between private and public sector and the exploitation of the current lack of measurements[94]. Other areas for possible legislative action and oversight include ideas for algorithmic impact assessments[95] and common standards for training data, development, deployment, and the environmental impact[96] of LLMs, as well as more universal performance benchmarks[97].

→ **Create clear rules for LLMs in a work context**. Another possible field of action for the public sector is clear and unambiguous regulation on the use of LLMs at work. Issues such as accountability for the output of LLMs and transparency on the use of AI in processes in the public sector would need to be clarified[98]. External auditing and building feedback loops from citizens or administrators to report on the quality of the responses they receive is also key. This would allow corrective action to be taken when required.

User-based measures

→ **Using better prompt strategies could improve results**. One important way of getting better results is to adjust input for models such as ChatGPT to reflect an organisation's mission and reasoning[99]. Staff can be trained in prompt engineering, such as breaking down more complex instructions into smaller, logical steps, write more detailed instructions and ask the model to then adjust content. This could lead to results which are more reliably in line with the user's intent.

Other risks might be mitigated by technological advancements. 'Hallucinations' still exist in the latest version of ChatGPT, but OpenAI claims that they have been drastically reduced with GPT4. The same claim is made for techniques to 'jailbreak' LLMs and get them to disregard their inbuilt safety features. Current LLMs often have a cut-off date regarding the information they can refer to when answering queries, so their replies may not be up to date. But newer models could be able to consult the internet to prepare their replies. In general, the quality of output, particularly in the case of complex queries, is increasing as newer models are being developed.

Some risks, however, are of a more structural nature and cannot be fully corrected despite best efforts. This is the case for detecting and removing biases in training data and model outputs. Efforts to sanitize datasets can even worsen biases. Given the current rate of development, some researchers fear that developers may soon run out of high-quality data to train models and may then resort to even more flawed datasets, at a time when mitigation strategies are still in their infancy[100]. Related to biases is the risk of a perpetuation of the status quo. LLMs mirror the values, habits and attitudes that are present in their training data, which does not leave much space for changing or underrepresented societal views[101]. Relying on LLMs that have been trained with previously produced documents in a public administration severely limits the scope for improvement and innovation and risks leaving the public sector even less flexible than it is already perceived to be.

The 'black box' issue, where AI models arrive at conclusions or decisions without revealing the process of how they were reached is also primarily structural, though the decision of AI companies not to grant open access to their code does not help. This means that the evolving abilities of

> *"Mitigating the risks of using LLMs requires an honest assessment of each possible area of use"*

language models are still not properly understood. While there seems to be some understanding as to how big an LLM needs to be before it can master a specific skill, the emergence of previously unknown or unintended abilities in a model after its training and fine-tuning remains a significant risk. Fine-tuning could also lead to previously unknown 'capability jumps', which could overpower safety features[102]. The black box problem also makes it extremely difficult to fix models which are already deployed, as it is often unclear which parts of the model need fixing[103].

Moreover, lack of funding is likely to continue to be a problem both for the public sector and European industry or the open-source community, which will struggle to close the gap with the US industry anytime soon. Regulating new technologies will remain a cat-and-mouse game. Acceleration risk (the emergence of a race to deploy new AI as quickly as possible at the expense of safety standards) is also an area of concern[104]. Finally, as mentioned earlier, a major structural risk lies in overreliance, which may be bolstered by rapid technological advances. This could lead to a lack of critical thinking skills needed to adequately assess and oversee the model's output, especially amongst a younger generation entering a workforce where such models are already being used.

Mitigating the risks of using LLMs requires an honest assessment of each possible area of use. Not all situations will create enough benefits to outweigh the risks. And not all LLMs may be suitable for use without prior customisation. When it comes to using ChatGPT in its current form, the issues that run up

against some of the key pillars of public administration are difficult to mitigate, which make it less suitable for regular use. This will apply to future applications based on this model. In the absence of clear regulation on LLM accountability, only humans can regularly monitor the output of ChatGPT and other LLMs. Given the structural flaws of LLMs, humans are still very much needed to provide personalized services, flexibility, emotional intelligence, critical thinking, and the ability to adapt quickly to changing circumstances necessary to fulfil the demands of public service. When asked, ChatGPT agreed with this assessment by highlighting its own limitations. A ChatGPT-supported public administration will therefore still need to rely on a significant proportion of human judgement, regular monitoring, and a robust mitigating strategy.

3. What lies ahead?

The release of ChatGPT has galvanised an already dynamic industry even further. But ChatGPT is not the only language model, and it will certainly not be the last. AI tools such as language models are set to become an increasing part of daily work, a trend that will also affect public institutions.

Looking forward, LLMs will be increasingly integrated into existing programmes. Microsoft is currently working on ways of integrating ChatGPT and other OpenAI models into its office applications, such as Outlook[105] and Microsoft Teams[106]. As LLMs become more widespread, the competition to be at the forefront of their development is heating up. One of the key battlegrounds is currently the efforts to revolutionise online searches. While this competition may be concentrated within a small handful of Big Tech companies, it is brutal, earning it the title of 'search wars'[107]. Google has dominated the search engine industry in the past and was an early adopter of LLM-assisted search, but there is no guarantee that this will continue to be the case. Microsoft, in particular, (with partial thanks, at least, to its multi-billion-dollar investment in OpenAI) is likely to be the winner: in early February 2023, Microsoft announced a new and improved search engine experience, available through both their Bing and Edge browsers, which integrates a next-generation OpenAI LLM that is designed to respond to search queries in a comprehensive, conversational manner, rather than providing pages of links to websites. To avoid being left behind, Google plans to roll out LLM-supported new features for its Google search and is currently testing the new conversational AI Bard[108].

However, the pressure to innovate implies familiar risks, such as rather limited search results at the expense of nuance, and most importantly a lack of source traceability. This issue is especially pertinent when it comes to the debate around the extraction of other people's work to produce something credited to an AI programme (and, by extension, the company which created it), which remains fraught. Several artists (mostly US-based)[109] and computer programmers[110] have already brought lawsuits against image and code-generating LLMs to clarify issues of copyright or piracy. Concerns about the use of data have been growing in Europe as well: on 31 March, the Italian data protection authority decided to ban ChatGPT from processing Italian users' data[111]. The European Consumer Organisation (BEUC) also called on the EU and national authorities to investigate the risks of ChatGPT and similar models[112]. Moreover, an open letter initiated by the Future of Life Institute signed by numerous prominent AI developers called for a six-month pause in developing systems more powerful than GPT-4, citing potential societal risks and a lack of robust safety systems[113].

> *"Research on LLMs is still developing, and their abilities and potential are still not well understood"*

Research on LLMs is still developing, and their abilities and potential are still not well understood[114]. After a trend towards ever-larger models, current developments seem to focus on optimisation and fine-tuning with the aim of creating smaller, more powerful models. Following the success of ChatGPT, conversational AI and LLM fine-tuning through reinforcement learning are likely to receive much more attention over the coming months[115]. Analysts are also monitoring developments on generative audio and video models and decentralised research collectives[116].

Although large language models can already perform a wide range of tasks, they still have their limits. There are various tasks where models are unable to outperform the best performing humans or even fail to reach a degree of accuracy which is distinguishable from completely random patterns. They are unable to think and understand like a human being, regardless of how human-like their output may seem. As language models are set to become more present in our daily lives, it is important to keep in mind their risks and not be mistaken into considering them as capable as a human being. Unfortunately, the way humans perceive others and assess their intelligence works against this, since humans tend to mistake fluency for intelligence. Coupled with automation bias – the inclination to rate machine-generated results as more accurate – and a preference for machine-generated output over the advice of humans when it suits one's pre-existing biases, the risk of overreliance increases dramatically. This is already creating a dangerous overdependence on the supposedly easy, accurate and readily available solutions provided by LLM-based applications such as ChatGPT. For reasons of convenience, public servants could deliberately or unwittingly ignore or play down the risks at the expense of the key functions of public service.

Digital literacy is therefore key. For public administrations this means staying on top of developments in the field of large language models and disseminating this knowledge to their employees as well as the citizens they serve. Institutions deploying LLMs in their daily processes will have an interest in communicating regularly on the importance of critically assessing any output coming from LLM. The increasing integration of LLMs could fundamentally change apps and programmes which are used regularly, such as search engines or text processing programmes. At this critical juncture, public administrations could lay the groundwork to adapt to the changes that large language models could bring.

Key questions

ChatGPT and other LLMs are here to stay, and they seem set to gain greater strategic importance in our societies, our daily lives and the ways we work. The following issues merit attention:

- ChatGPT is already used, at least informally, by employees in the private and public sector. What measures should public administrations take to **maximise the potential of LLMs while mitigating the risks?**

- What are the main possible **legal** hurdles in the integration of LLMs in public administration processes? How to address issues of legal responsibility?

- What impact will LLMs have on **ways of working**?
  How could public administrations assess the impact of LLMs on the future of work in the public sector?
  Are there synergies to be sought between local, national and the EU level?

- **Transparency** is one of the key aspects when integrating LLMs in public sector processes.
  Citizens and other stakeholders should not have the impression that this is done behind their back.
  How should public administrations **communicate** on the use of LLMs and coordinate their strategies to ensure coherent messaging where needed?

- With more than 70% of LLMs developed in the US and huge entry costs giving an advantage to big tech, it is not easy for EU-based actors to catch up and for **European champions** to emerge in this new race. Not only is Europe behind, but this structural advantage also affects the nature and mindset of these models, which are much more based on US values and ideals. What other avenues could be explored to prevent Europeans from only being power users of foreign technologies?

Annex: Glossary

- **Artificial Intelligence (AI)**: refers to the development of computer systems able to perform tasks that normally require human intelligence
- **Artificial general intelligence (AGI)**: theoretical form of AI that would be capable of performing any intellectual tasks that a human can do
- **Chatbot**: computer programme designed to simulate conversations with human users
- **Edge model**: type of machine learning model designed to run efficiently on devices with limited computational power
- **Fine-tuning**: process where a pre-trained model is further trained on a small dataset to improve its performance on specific tasks
- **General purpose AI system**: an AI system that can be applied to a wide range of tasks and domains, as opposed to a system designed for a specific purpose
- **Hallucination**: phenomenon where a model generates outputs that are nonsensical or unfaithful to the input data the model has been trained on
- **Large language model (LLM)**: type of AI system trained on massive amounts of text data that can generate natural language responses to a wide range of inputs
- **Machine learning**: a subfield of AI that involves the use and development of algorithms and statistical models that are able to learn from data and adapt their performance without being explicitly programmed to do so
- **Neural networks**: in this context, it refers to a type of machine learning algorithm designed to simulate the way the human brain works
- **Parameter**: weights and biases within a machine learning model that the model uses in its decision-making processes
- **Prompt**: phrase provided to a language model in order to generate a response or perform a task
- **Self-attention**: mechanism used in neural networks allowing the model to focus simultaneously on different parts of the input data during processing
- **Token**: discrete unit of text, such as a word or punctuation mark, that is used as input for large language models
- **Transformer**: type of neural network architecture used in natural language processing
- **Virtual assistant** : AI system designed to provide assistance and perform tasks for users

References

1 A recent paper suggests that what appears to be a sign of intelligence in a model could in fact be a mirror of the intelligence of the user, which materialises in the kinds of prompts a user chooses, see Terrence J. Sejnowski, 'Large Language Models and the Reverse Turing Test', *Neural Comput* 2023, Vol. 35, No.3, pp. 309–342. doi: https://doi.org/10.1162/neco_a_01563.

2 Alan Turing, 'Computing Machinery and Intelligence', *Mind*, 1950, Vol.59, No.236, pp.433–460.

3 Joel Walmsley, *Mind and Machine*, Basingstoke: Palgrave Macmillan, 2012.

4 Kate Crawford, *Atlas of AI*, New Haven: Yale University Press, 2021.

5 Young et al. 'Recent Trends in Deep Learning Based Natural Language Processing'. I*EEE Computational Intelligence Magazine*. Vol.13, No.3, pp.55-75, https://arxiv.org/pdf/1708.02709.pdf

6 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 'Attention is All You Need'. *Advances in neural information processing systems*. Vol.30 pp.1-11.

7 Whether ChatGPT is the name of the conversational interface or of the model itself was initially unclear. Some early sources call the interface ChatGPT and the model underneath it GPT-3.5, whereas OpenAI used ChatGPT to refer to both the model and the chatbot interface. It later called the model gpt-3.5-turbo.

8 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts', Council of the European Union, Interinstitutional File 2021/0106(COD), 25 November 2022, https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf, p.71.

9 Molly Ruby, 'How ChatGPT Works: The Model Behind The Bot', *Medium*, 30 January 2023, https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286.

10 See Vaswani et al., 'Attention is all you need', op. cit.

11 See Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei, 'Scaling laws for neural language models', *arXiv preprint arXiv:2001.08361* , 23 January 2020.

12 See Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas et al., 'Training compute-optimal large language models', arXiv preprint arXiv:2203.15556, 29 March 2022.

13 'Introducing ChatGPT', *OpenAI*, 30 November 2022, https://openai.com/blog/chatgpt.

14 See Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese et al. , 'Fine-tuning language models to find agreement among humans with diverse preferences', Advances in Neural Information Processing Systems, 2022, vol. 35, pp. 38176-38189; and Ouyang Long, et al. 'Training language models to follow instructions with human feedback', *arXiv preprint arXiv:2203.02155*, 04 March 2022. For a less technical explanation see Marco Ramponi, How ChatGPT actually works, *AssemblyAI*, 23 December 2022, https://www.assemblyai.com/blog/how-chatgpt-actually-works/.

15 Maximilian Schreiner, 'Google shows generative AI model for speech and music', *The Decoder*, 14 September 2022, https://the-decoder.com/google-shows-generative-ai-model-for-speech-and-music/.

16 Elon Musk resigned in 2018 citing conflicts of interest over his ownership of Tesla, which was increasingly expanding into AI. Following the early 2023 Microsoft deal Musk publicly distanced himself from the company. See Christiaan Hetzner, 'Elon Musk lashes out at the ChatGPT sensation he helped create after Microsoft's massive investment —'Not what i intended'', *Fortune*, 17 February 2023, https://fortune.com/2023/02/17/chatgpt-elon-musk-openai-microsoft-company-regulator-oversight/.

17 'About', *OpenAI*, 2023, https://openai.com/about.

18 'OpenAI LP', *OpenAI*, 11 March 2019, https://openai.com/blog/openai-lp.

19 'OpenAI and Microsoft extend partnership', *OpenAI*, 23 January 2023, https://openai.com/blog/openai-and-microsoft-extend-partnership/.

20 'OpenAI API', *OpenAI*, 11 June 2020, https://openai.com/blog/openai-api/.

21 Cade Metz, 'A.I. Researchers Are Making More Than $1 Million, Even at a Nonprofit', *The New York Times*, 19 April 2019, https://www.nytimes.com/2018/04/19/technology/artificial-intelligence-salaries-openai.html.

22 'GPT-2: 1.5B release', *OpenAI*, 05 November 2019, https://openai.com/research/gpt-2-1-5b-release.

23 Noor Al-Sibai, 'OpenAI was founded to counter bad AI, now worth billions as it does the opposite', *Futurism*, https://futurism.com/the-byte/openai-billions-bad-ai.

24 'Models', *OpenAI*, 2023, https://platform.openai.com/docs/models.

25 Aleks Farseev, 'Is Bigger Better? Why The ChatGPT Vs. GPT-3 Vs. GPT-4 'Battle' Is Just A Family Chat', *Forbes*, 17 February 2023, https://www.forbes.com/sites/forbestechcouncil/2023/02/17/is-bigger-better-why-the-chatgpt-vs-gpt-3-vs-gpt-4-battle-is-just-a-family-chat/.

26 For estimates, see for example Or Sharir, Barak Peleg, and Yoav Shoham. "The cost of training nlp models: A concise overview." a*rXiv preprint arXiv:2004.08900*, 19 April 2019; and Dylan Patel, 'The AI Brick Wall – A Practical Limit For Scaling Dense Transformer Models, and How GPT 4 Will Break Past It', *Semianalysis*, 24 January 2023, https://www.semianalysis.com/p/the-ai-brick-wall-a-practical-limit.

27 Sindhu Sundar, 'If you still aren't sure what ChatGPT is, this is your guide to the viral chatbot that everyone is talking about', *Business Insider*, 01 March 2023, https://www.businessinsider.com/everything-you-need-to-know-about-chat-gpt-2023-1?r=US&IR=T.

28 Yejin Bang et al. 'A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity', *arXiv preprint arXiv:2302.04023*, 08 February 2023, https://arxiv.org/pdf/2302.04023.pdf, p.3.

29 Op. cit., pp.22-23; see also Chenwei Qin et al., 'Is ChatGPT a general-purpose natural language processing task solver?', arXiv preprint arXiv:2302.06476 15 February 2023; https://arxiv.org/pdf/2302.06476.pdf; Simon Frieder et al., Mathematical Capabilities of ChatGPT, *arXiv preprint arXiv:2301.13867*, 31 January 2023.

30 Jon Christian, 'Amazing "Jailbreak" Bypasses ChatGPT's Ethics Safeguards', *Futurism*, 04 February 2023, https://futurism.com/amazing-jailbreak-chatgpt.

31 David F. Carr, 'ChatGPT Topped 1 Billion Visits in February', *Similarweb*, 21 March 2023, https://www.similarweb.com/blog/insights/ai-news/chatgpt-1-billion/.

32 Pranshu Verma, 'What to know about OpenAI, the company behind ChatGPT', *The Washington Post*, 14 March 2023, https://www.washingtonpost.com/technology/2023/02/06/what-is-openai-chatgpt/.

33 Yusuf Mehdi, 'Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web', *Official Microsoft Blog*, 07 February 2023, https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/.

34 Sundar Pichai, 'An important next step on our AI journey', *The Keyword*, Google, 06 February 2023, https://blog.google/technology/ai/bard-google-ai-search-updates/.

35 'Introducing LLaMA: A foundational, 65-billion-parameter large language model', *Meta AI*, 24 February 2023, https://ai.facebook.com/blog/large-language-model-llama-meta-ai/.

36 'Große KI-Modelle für Deutschland', Machbarkeitsstudie zu LEAM – Large European AI Models, *Akademie für künstliche Intelligenz AKI gGmbH*, 2023, https://leam.ai/wp-content/uploads/2023/01/LEAM-MBS_KIBV_webversion_mitAnhang_V2_2023.pdf, p.6.

37 Op. cit., pp.56-60.

38 Op. cit., p.57.

39 See 'Emerging Non-European Monopolies in the Global AI Market', *Future of Life Institute*, November 2022, https://futureoflife.org/wp-content/uploads/2022/11/Emerging_Non-European_Monopolies_in_the_Global_AI_Market.pdf.

40 Matthias Bastian, 'German initiative aims to build a European AI infrastructure', *The Decoder*, 25 January 2023, https://the-decoder.com/german-initiative-aims-to-build-a-european-ai-infrastructure/.

41 Maximilian Schreiner,' OpenAI competitor Aleph Alpha launches Europe's fastest commercial AI data center', *The Decoder*, 16 September 2022, https://the-decoder.com/openai-competitor-launches-europes-fastest-commercial-ai-data-center/.

42 'Große KI-Modelle für …', op. cit., p.7.

43 'Discover EuroHPC JU', *The European High Performance Computing Joint Undertaking (EuroHPC JU)*, 2023, https://eurohpc-ju.europa.eu/about/discover-eurohpc-ju_en.

44 'LUMI supercomputer', *LUMI consortium*, 2023, https://www.lumi-supercomputer.eu/lumi_supercomputer/.

45 'A space that combines petabytes of natural language data with large-scale model training', *HPLT High Performance Language Technologies Consortium*, 2023, https://hplt-project.org/static/media/hplt-factsheet.2b9a51f64b7fd81f1e79.pdf. HPLT (hplt-project.org).

46 'Natural Language Understanding and Interaction in Advanced Language Technologies (AI Data and Robotics Partnership) (RIA)', Funding & tender opportunities Single Electronic Data Interchange Area (SEDIA), *European Commission*, https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-cl4-2023-human-01-03.

47 'Project OpenGPT-X', *KI Bundesverband*, 2022, https://opengpt-x.de/en/project.

48 See 'The Future of Jobs Report 2020', *World Economic Forum*, October 2020, https://www3.weforum.org/docs/WEF_Future_of_Jobs_2020.pdf, pp. 29-34.

49 'Generative AI set to affect 300mn jobs across major economies', *Financial Times*, 27 March 2023, https://www.ft.com/content/7dec4483-ad34-4007-bb3a-7ac925643999.

50 Brady and Wang Ting, 'Chatting about ChatGPT: How May AI and GPT Impact Academia and Libraries?' *Library Hi Tech News*, 2023. https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID4333415_code5653239.pdf?abstractid=4333415&mirid=1&type=2.

51 OECD, 'European Principles for Public Administration ', *SIGMA Papers*, No. 27, *OECD Publishing*, 1999, https://doi.org/10.1787/5kml60zwdr7h-en ; De Vries, Michiel S. and Pan Suk Kim, Value and Virtue in Public Administration A Comparative Perspective, Springer, 2011.

52 Tero Erkkilä, 'Transparency in Public Administration', *Oxford Research Encyclopedia of Politics*, 29 May 2020, https://doi.org/10.1093/acrefore/9780190228637.013.1404.

53 Hannah Bloch-Wehba, 'Transparency's *AI* Problem', Knight First Amendment Institute and *Law and Political Economy Project's* Data & Democracy Essay Series, 2021, https://scholarship.law.tamu.edu/facscholar/1477, pp. 7-12.

54 Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. 'Language models are unsupervised multitask learners'. *Computer Science*, 24 February 2019, https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

55 OpenAI's usage policy prohibits using its models for "High risk government decision-making, including: Law enforcement and criminal justice; Migration and asylum", but this still leaves some scope for a potential use of its models for decision-making processes, see OpenAI usage policies, 17 March 2023, https://openai.com/policies/usage-policies.

56 Emily M. Bender, Gebru Timnit, Angelina McMillan-Major, and Schmargaret Schmitchell, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?', in the Association for Computing Machinery Conference on Fairness, Accountability, and Transparency, March 2021, https://dl.acm.org/doi/10.1145/3442188.3445922#sec-cit.

57 Special Eurobarometer 523: Corruption, March-April 2022, https://europa.eu/eurobarometer/api/deliverable/download/file?deliverableId=83025, p. 16.

58 Matthew M. Young, Justin B. Bullock, and Jesse D. Lecy. 'Artificial discretion as a tool of governance: a framework for understanding the impact of artificial intelligence on public administration', *Perspectives on Public Management and Governance* 2, no.4, 2019, p. 308.

59 Size does not guarantee diversity even when looking at the whole Internet. Internet access is not evenly distributed, resulting in an overrepresentation of younger users and those from developed countries. The voices of people most likely to hew to a hegemonic viewpoint are also more likely to be retained.

60 Fahim Faisal, Yinkai Wang, Antonios Anastasopoulos, 'Dataset Geography: Mapping Language Data to Language Users', Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, 22-27 May 2022, p. 3384.

61 Schramowski, Patrick, Turan, Cigdem, Andersen, Nico, Rothkopf, Constantin A., and Kersting, Kristian, 'Large Pre-trained Language Models Contain Human-like Biases of What is Right and Wrong to Do', *Nature Machine Intelligence*, 2103.11790.pdf, p.16 ; Abid, Abubkar, Farooqi, Maheen, and Zou, James, 'Persistent Anti-Muslim Bias in Large Language Models' AAAI/ACM Conference on AI, Ethics, and Society, 2021, 2101.05783.pdf.

62 Emily Bender et al, 'On the Dangers of Stochastic Parrots…', op. cit., pp. 613-615.

63 Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe, 'Training language models to follow instructions with human feedback' Thirty-Sixth Conference on Neural Information Processing Systems, 28 November 2022, 2203.02155.pdf, pp. 19-20.

64 Saar Alon-Barkat and Madalina Busuioc, 'Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice', *Journal of Public Administration Research and Theory*, Volume 33, Issue 1, January 2023, https://academic.oup.com/jpart/article/33/1/153/6524536., pp. 165-166.

65 Crawford, Kate, *Atlas of AI*, New Haven: Yale University Press, 2021.

66 ChatGPT Burns Millions Every Day. Can Computer Scientists Make AI One Million Times More Efficient?', *Forbes*, 10 February 2023 https://www.forbes.com/sites/johnkoetsier/2023/02/10/chatgpt-burns-millions-every-day-can-computer-scientists-make-ai-one-million-times-more-efficient/?sh=771ef56d6944.

67 David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean, 'Carbon Emissions and Large Neural Network Training', *arXiv pre-print*, https://arxiv.org/ftp/arxiv/papers/2104/2104.10350.pdf, p. 6.

68 Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni, 'Creating efficiency in AI research will decrease its carbon footprint and increase its inclusivity as deep learning study should not require the deepest pockets' *Green AI*, https://dl.acm.org/doi/pdf/10.1145/3381831,p. p. 63; David Patterson, 'Reducing the carbon emissions of AI' - OECD.AI', 12 April 2022, https://oecd.ai/en/wonk/reducing-ai-Reducing the carbon- emissions of AI - OECD.AI.

69 David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean, 'The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink', *IEEE*, https://ieeexplore.ieee.org/document/9810097 , pp. 6-7.

70 'Chatbots Got Big—and Their Ethical Red Flags Got Bigger', *Wired*, 16 February 2023, https://www.wired.com/story/chatbots-got-big-and-their-ethical-red-flags-got-bigger/.

71 See https://consensus.app/home/about-us/

72 Deep Ganguli, Danny Hernandez, Liane Lovitt, Nova Dassarma, Tom Henighan, Andy jones, Nicholas Joseph, Jackson Kernion, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Nelson Elhage, Sheer el Showk, Stanislav Fort, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Chris Olah, Ario Amodei, and Jack Clark, 'Predictability and Surprise in Large Generative Models', 2022, p.11.

73 'How ChatGPT Highjacks Democracy', *The New York Times*, 15 January 2023 https://www.nytimes.com/2023/01/15/opinion/ai-chatgpt-lobbying-democracy.html.

74 Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova 'Truth, Lies, and Automation - How Language Models Could Change Disinformation' Center for Security and Emerging Technology, May 2021.

75 'GPT Zero Is Designed To Recognize Texts Written By AI', *Medium*, 12 January 2023 https://medium.datadriveninvestor.com/gpt-zero-is-designed-to-recognize-texts-written-by-ai-ab7ff4d11fd6.

76 Elmer B. Staats, 'Public Service and the Public Interest', Public Administration Review ublic Administration Review, Vol. 48, No. 2, March-April 1988, pp. 601-605+ii.

77 Schramowski, Patrick, Turan, Cigdem, Andersen, Nico, Rothkopf, Constantin A., and Kersting, Kristian, 'Large Pre-trained Language Models Contain Human-like Biases of What is Right and Wrong to Do', Nature Machine Intelligence, p. 17, 2103.11790.pdf, p.11.

78 Deep Ganguli et al, 'Predictability and Surprise in Large Generative Models', op.cit., p.13.

79 Michiel Bakker et al., 'Fine-tuning language models…' op.cit., p.10.

80 Deep Ganguli et al., 'Predictability and surprise in large generative models', 2022 ACM Conference on Fairness, Accountability, and Transparency, p. 9.

81 Khari Johnson, 'The Movement to Hold AI Accountable Gains More Steam', *Wired*, 02 December 2021, https://www.wired.com/story/movement-hold-ai-accountable-gains-steam/.

82 Abir Rahali and Moulay A. Akhloufi, 'MalBERT: Malware Detection using Bidirectional Encoder Representations from Transformers,' *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Melbourne, Australia, 2021, pp. 3226-3231.

83 'Privacy policy', *OpenAI*, 14 March 2023, https://openai.com/policies/privacy-policy.

84 'GPT-4 System Card', *OpenAI*, 23 March 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf, p.22.

85 Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts et al. 'Extracting Training Data from Large Language Models', USENIX Security Symposium, vol. 6, 11 August 2021, p.13.

86 Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr, 'What Does it Mean for a Language Model to Preserve Privacy?', 2022 ACM Conference on Fairness, Accountability, and Transparency, 20 June 2022, pp. 2288-2289.

87 Eran Shimony and Omer Tsarfati, 'Chatting Our Way Into Creating a Polymorphic Malware', *Cyberark*, 17 January 2023, https://www.cyberark.com/resources/threat-research-blog/chatting-our-way-into-creating-a-polymorphic-malware.

88 OpenAI is among the companies who have recently reduced prices to train their models, see 'September 2022 - OpenAI API Pricing Update FAQ', OpenAI, September 2022, https://help.openai.com/en/articles/6485334-september-2022-openai-api-pricing-update-faq; and Daniel Zhang, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Ellie Sakhaee, Yoav Shoham, Jack Clark, and Raymond Perrault, 'The AI Index 2022 Annual Report,' AI Index Steering Committee, *Stanford Institute for Human-Centered AI*, Stanford University, March 2022, p. 97.

89 Kyle Wiggers, 'The emerging types of language models and why they matter', *TechCrunch*, 28 April 2022, https://techcrunch.com/2022/04/28/the-emerging-types-of-language-models-and-why-they-matter/.

90 Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton, 'Characterising bias in compressed models', *arXiv preprint arXiv:2010.03058*, 06 October 2020, https://arxiv.org/pdf/2010.03058.pdf.

91 'Complete Guide to Edge Computing: Edge IoT, Edge AI, and More', *Run.ai*, https://www.run.ai/guides/edge-computing.

92 'Joint Recommendation for Language Model Deployment', Cohere, *OpenAI*, and AI 21 Labs, https://cdn.openai.com/papers/joint-recommendation-for-language-model-deployment.pdf.

93 Artificial Intelligence Act: Council calls for promoting safe AI that respects fundamental rights', *Council of the European Union*, 6 December 2022, https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/.

94 Johanna Okerlund, Evan Klasky, Aditya Middha, Sujin Kim, Hannah Rosenfeld, Molly Kleinman, and Shobita Parthasarathy, 'What's in the Chatterbox? Large Language Models, Why They Matter, and What We Should Do About Them', *University of Michigan Technology Assessment Report*, April 2022, https://stpp.fordschool.umich.edu/research/research-report/whats-in-the-chatterbox; Jess Whittlestone and Jack Clark, 'Why and How Governments Should Monitor AI Development', 2021, https://arxiv.org/pdf/2108.12427.pdf.

95 Selbst, 'An Institutional View Of Algorithmic Impact Assessments', Harvard Journal of Law & Technology 2021; *UCLA School of Law*, Public Law Research Paper No. 21-25, https://ssrn.com/abstract=3867634.

96 'IEEE calls for standards to combat climate change and protect kids in the age of AI', *VentureBeat*, 6 February 2020, https://venturebeat.com/ai/ieee-calls-for-standards-to-combat-climate-change-and-protect-kids-in-the-age-of-ai/.

97 Rishi Bommasani, Percy Liang, and Tony Lee, 'Language Models are Changing AI. We Need to Understand Them', Standford University Human-Centered Artificial Intelligence https://hai.stanford.edu/news/language-models-are-changing-ai-we-need-understand-them.

98 Hannah Bloch-Wehba, 'Transparency's AI Problem', op. cit., pp.19-21.

99 'Tech's hottest new job: AI whisperer. No coding required.' *The Washington Post*, 25 February 2023, https://www.washingtonpost.com/technology/2023/02/25/prompt-engineers-techs-next-big-job/.

100 Zhuo, Terry Yue, Yujin Huang, Chunyang Chen, and Zhenchang Xing., 'Exploring ai ethics of chatgpt: A diagnostic analysis', *arXiv preprint*, 22 February 2023, *arXiv:2301.12867*.

101 Emily Bender et al, 'On the Dangers of Stochastic Parrots…', op. cit., pp. 614.

102 GPT-4 System Card', *OpenAI*, 23 March 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf, p.29.

103 Miller, Katherine, 'How Do We Fix and Update Large Language Models?', *Stanford University Human-Centered Artificial Intelligence*, 13 February 2023, https://hai.stanford.edu/news/how-do-we-fix-and-update-large-language-models.

104 GPT-4 System Card', *OpenAI*, 23 March 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf, p.21.

105 Matteo Pagani, 'Bring the ChatGPT model into our applications', *Microsoft*, https://techcommunity.microsoft.com/t5/modern-work-app-consult-blog/bring-the-chatgpt-model-into-our-applications/ba-p/3766574.

106 Nicole Herskowitz, 'Microsoft Teams Premium: Cut costs and add AI-powered productivity', *Microsoft*, 1 February 2023 https://www.microsoft.com/en-us/microsoft-365/blog/2023/02/01/microsoft-teams-premium-cut-costs-and-add-ai-powered-productivity/.

107 'Search wars reignited by artificial intelligence breakthroughs', *Financial Times*, 6 February 2023. https://www.ft.com/content/b236b70d-82dc-40f8-84be-dc4daff151e4.

108 Sundar Pichai, 'An important next step on our AI journey', *Google*, 6 February 2023 https://blog.google/technology/ai/bard-google-ai-search-updates.

109 'First AI Art Generator Lawsuits Threaten Future of Emerging Tech', *Bloomberg*, 20 January 2023. https://news.bloomberglaw.com/ip-law/first-ai-art-generator-lawsuits-threaten-future-of-emerging-tech.

110 'Lawsuit Takes Aim at the Way A.I. Is Built', *New York Times*, 23 November 2022. https://www.nytimes.com/2022/11/23/technology/copilot-microsoft-ai-lawsuit.html

111 'L'Intelligenza artificiale: il Garante blocca ChatGPT. Raccolta illecita di dati personali. Assenza di sistemi per la verifica dell'età dei minori', Garante per la protezione dei dati personali, 31 March 2023, https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9870847.

112 'Investigation by EU authorities needed into ChatGPT technology', European Consumer Organisation (BEUC), 30 March 2023, https://www.beuc.eu/press-releases/investigation-eu-authorities-needed-chatgpt-technology.

113 'Pause Giant AI Experiments: An Open Letter', Future of Life Institute, 22 March 2023, https://futureoflife.org/open-letter/pause-giant-ai-experiments/; Jyoti Narayan, Krystal Hu, Martin Coulter, Supantha Mukherjee, 'Elon Musk and others urge AI pause, citing 'risks to society'', Reuters, 29 March 2023, https://www.reuters.com/technology/musk-experts-urge-pause-training-ai-systems-that-can-outperform-gpt-4-2023-03-29/.

114 'The emerging types of language models and why they matter', TechCrunch, April 28 2023,https://techcrunch.com/2022/04/28/the-emerging-types-of-language-models-and-why-they-matter/?guccounter=1.

115 'Trends in AI — 2023 Round-up', *Medium/ Towards AI*, January 25 2023, https://pub.towardsai.net/trends-in-ai-2023-round-up-e6c52578fe92.

116 Nathan Benaich and Ian Hogarth, 'State of AI Report 2022', *State of AI*, https://www.stateof.ai/.

Disclaimer: *During the writing process, ChatGPT was used to obtain a view on some of the descriptions of machine learning techniques and was consulted on whether its own assessment of its ability to respect some of the principles underpinning the work of the public sector corresponded with our own.*