Introduction

"ChatGPT is incredibly limited, but good enough at some things to create a misleading impression of greatness. It's a mistake to rely on it for anything important just yet. It's a preview of progress; we have a lot of work to do on robustness and veracity". Sam Altman¹



"Artificial intelligence is the most profound technology that humanity is working on, more profound than fire, electricity or anything else we've done in the past. It gets to the essence of what intelligence is, what humanity is. It will certainly someday be far more capable than anything we've seen before"².

This is Google CEO Sundar Pichai's view on the rise of artificial intelligence (AI), which not only highlights its depth and potential, but also positions AI as a milestone in the history of technological and human development.

Generative Artificial Intelligence (GenAI) and, within it, Large Language Models (LLM) are emerging as the most significant manifestation of this transformation.

It is important to note that this breakthrough is a logical consequence of the digital transformation process, driven by advances in data storage, processing, data availability and new modeling techniques, without which this milestone would not have been possible.

GenAl³ refers to artificial intelligence systems capable of generating new and original content, be it text, images, video, voice, music, 3D models or programming code. These systems learn from massive amounts of existing data and can produce outputs that, in many cases, are indistinguishable from those created by humans. This ability to create content opens up new possibilities in all areas of every industry, with implications that are still difficult to predict.

Specifically, GenAl is finding potentially revolutionary applications⁴ in areas such as education, where it can personalize and enhance learning; healthcare, where it can facilitate more accurate diagnoses and the development of individualized treatments; finance, where it can improve risk analysis and fraud detection; commerce, where it can optimize the supply chain and the customer experience; art, where it can open up new creative possibilities; and law, where it can streamline contract review and predict legal outcomes, to name just a few.

Within GenAl, LLMs (such as OpenAl ChatGPT, Anthropic Claude, Google Gemini, Meta Llama, Mistral or SenseTime SenseNova) represent a disruptive advance in natural language processing. These models are able to analyze and generate text with a level of coherence, relevance, and fluency previously unattainable by other algorithms. Their applications range from writing assistance and idea generation to automated translation, full report generation citing relevant articles and regulations, and the creation of more natural and effective conversational interfaces ("chatbots")⁵.

GenAl, including LLMs, is influencing our interaction with technology and information, helping to transform content creation, data-driven decision making, and the way we interact with machines. Despite still being in its early stages⁶ its full impact is yet to be determined. In this sense, it is already being used to create advanced virtual assistants, in voice and gesture interfaces for controlling home devices, in instant translation interfaces, and in integration with augmented reality and virtual reality technologies.

²S. Pichai (2023). Interview on 60 Minutes Overtime.
 ³Gartner (2023).

¹Samuel Harris Altman (b. 1985), American entrepreneur, founder and CEO of OpenAl.

⁴See a taxonomy and extensive collection of GenAl use cases in iDanae 2Q23 (2023) and in Gozalo-Brizuela, R., and Garrido-Merchán, E.C. (2023). ⁵Fischer (2021).

⁶Lam (2018).

At the enterprise level, most large companies are already developing LLM-based systems to industrialize processes, including customer service, data analysis, reporting, and automation of repetitive tasks. According to a Microsoft⁷ study, integrating LLM as a co-pilot in office automation tools results in time savings ranging from 27% to 74% without compromising quality⁸. In the case of SMBs, the level of adoption is still limited, creating an even greater risk of a technology gap for this segment⁹.

When properly applied, LLMs have the potential to optimize processes, reduce time and save costs. In addition, they can improve the objectivity and quality of documents, reduce errors, offer new ways of interacting with customers and, thanks to their ability to analyze massive amounts of information, provide access to previously unavailable knowledge due to processing and comprehension limitations. However, it is important to remember that successful optimization depends on factors such as data quality, learning complexity and the appropriateness of the model to the problem at hand. Going further, some experts see LLMs as a step toward the creation of Artificial General Intelligence (AGI), a mediumterm goal in which AI could mimic a wide range of intellectual tasks that humans can perform. However, the concept of AGI remains vague and its feasibility is subject to significant cultural, political and legal constraints, such as ethics or privacy, which would require further specification and analysis. It is also crucial to recognize the inherent limitations of AI, which, according to philosophers of language such as John Searle and his "Chinese Room" experiment¹⁰, lacks the capacity for abstraction and association of concepts to symbols, an attribute unique to the human mind.

⁷Cambon (2023). Study conducted by Microsoft on Al and productivity.
⁸In the study, participants who used Microsoft 365 Copilot (an LLM integrated with the Microsoft Office suite tool) completed a number of common tasks (e.g. retrieving email and intranet information, creating content, summarising meetings) in less time than those who did not use Copilot. It is important to note that the study focused on a specific set of tasks for which Copilot was expected to provide significant value, rather than a representative set of all employee tasks.

⁹IndesIA (2024).

¹⁰Searle, J. (1980).

Figure 1. Classification of artificial intelligence into levels of capability compared to humans. Adapted from Google DeepMind (2023).

Performance (rows) x Generality (columns)	Narrow Clearly scoped task or set of tasks	General Wide range of non-physical tasks, including metacognitive abilities like learning new skills
Level 0: No Al	Narrow Non-Al Calculator software; compiler	General Non-Al Human-in-the-loop computing, e.g., Amazon Mechanical Turk
Level 1: Emerging Equal to or somewhat better than an unskilled human	Emerging Narrow Al GOF4 Simple rule-based systems, e.g., SHRDLU	Emerging AGI ChatGPT, Gemini, Claude, Llama
Level 2: Competent At least 50th percentile of skilled adults	Competent Narrow Al Toxicity detectors such as Jigsaw Siri (Apple), Alexa (Amazon), Google Assistant (Google) VQA systems such as PaLI, Watson (IBM), SOTA LLMs (e.g., short essay writing, simple coding)	Competent AGI Not yet achieved
Level 3: Expert At least 90th percentile of skilled adults	Expert Narrow Al Spelling & grammar checkers such as Grammarly Generative image models such as Imagen or Dall-E 2	Expert AGI Not yet achieved
Level 4: Virtuoso At least 99th percentile of skilled adults	Virtuoso Narrow Al Deep Blue: chess-playing computer developed by IBM that defeated the world champion in 1997. AlphaGo: an Al developed by DeepMind that defeated world-class players in the board game Go	Virtuoso AGI Not yet achieved
Level 5: Superhuman Outperforms 100% of humans	Superhuman Narrow Al AlphaFold: predicts protein structures with high accuracy AlphaZero: self-taught Al that masters games like chess, Go, and shogi StockFish: a powerful open-source chess engine	Artificial Superintelligence (ASI) Not yet achieved

According to several experts¹¹, AGI could be achieved between 2029 and 2035, or even sooner. While today's AI specializes in specific tasks ("narrow AI") and LLMs are beginning to exhibit general capabilities, AGI promises much broader versatility and adaptability. Although there is already specialist AI that outperforms 100% of humans (e.g., chessplaying AI), Google DeepMind estimates¹² that the progress of AGI (e.g., LLMs) is currently at a level of only 1 out of 5; i.e., just in its infancy (Figure 1).

However, with these advances in GenAI and LLM come significant risks, ethical considerations and challenges, including¹³ data privacy and information security; difficulties in model interpretability; generation of false or misleading information ("hallucinations"¹⁴); propagation of bias, discrimination and inappropriate or toxic content; challenges in AI regulation and governance; regulatory non-compliance with potential sanctions; intellectual property, copyright, authorship and plagiarism issues; high resource consumption and environmental impact; the "Eliza Effect"¹⁵, overconfidence and reduced critical capacity; ethical risks in automated decision making; risk of overreliance on AI for critical tasks; risks of using LLM for manipulation and misinformation; risk of human job replacement¹⁶; need for job transition and training; and inequalities in access to and use of AI technologies, to name a few of the most important.

Specifically, LLMs can generate hallucinations (Figs. 2 to 3), i.e., false or misleading information, which combined with the "Eliza Effect", where users attribute human cognitive abilities to these systems, can lead to overconfidence, dependency or misinterpretation, and thus to wrong decisions.

The Chinese room

The Chinese room experiment, conceived by philosopher John Searle in 1980, poses a thought-provoking challenge to the concept of artificial intelligence. In this experiment, a person who does not understand Chinese is placed in a closed room filled with English instructions to manipulate Chinese symbols. This individual is given Chinese characters, uses these instructions to respond with appropriate Chinese characters, and thus appears to understand Chinese.

However, Searle argues that this is an illusion because the person is merely following syntactic rules without understanding the semantics - the meaning - of the symbols. This experiment raises fundamental questions about the nature of understanding and consciousness in machines, suggesting that mere symbol manipulation does not amount to true understanding.

The response to Searle's experiment has been diverse and evolving. Critics, especially from the AI and cognitive science communities, argue that the experiment overlooks the possibility that understanding may reside in the whole system (the person, the instructions, and the room), not just the individual. This is consistent with the functionalist view of philosophy, which considers mental states in terms of their functional utility.

Searle's supporters argue that true understanding requires more than symbol manipulation, possibly involving consciousness or subjective experience, which machines lack. Over time, the debate has moved beyond these binary positions into nuanced discussions about the nature of consciousness and understanding in machines.

In the practical development of AI, research has largely focused on improving the capabilities and addressing the limitations of AI systems, rather than on the philosophical implications regarding consciousness or understanding. As AI technology advances, some of the original premises of the Chinese room argument are re-examined, while addressing the pragmatic aspects of AI development and application.

¹¹Pew Research Institute (2023).

¹²Google DeepMind (2023).

¹³See iDanae 2Q23 (2023).

¹⁴Xu (2024).

¹⁵Weizenbaum, J. (1966).). The "Eliza Effect" is the phenomenon whereby users attribute comprehensibility and emotional responses to a computer program. ¹⁶Although some studies, such as Google + Implement (2023), suggest that around 94% of jobs will remain unchanged or will be supplemented by GenAI, but not partially or fully replaced, and that GenAl will have a positive impact and boost productivity across all industries.

In the face of these challenges, regulators are taking proactive steps at the national and international levels to address the risks and opportunities of Al. Of particular note is the Bletchley Declaration¹⁷, signed by the European Union and 27 countries (including the United States, United Kingdom, China, India, Brazil and Australia) in November 2023, which sets out a global commitment to the responsible development of Al.

For its part, the European Union, with the imminent implementation of the Artificial Intelligence Act¹⁸, is introducing the first comprehensive legally binding framework that classifies AI systems according to their risk and sets stringent standards, especially for high-risk systems. And in the United States, President Biden's Executive Order¹⁹, issued on October 30, 2023, and the Blueprint for an Artificial Intelligence Bill of Rights²⁰ set standards to ensure the safety, reliability and fairness of AI, with a focus on privacy, civil rights, consumer protection and international leadership in AI governance.

In this context, organizations are defining their Al strategy (with a particular on GenAl and LLMs), designing their Al adoption plan, and adapting their structures, including the creation of GenAl centers of excellence and the incorporation of new figures such as the Chief Al Officer²¹. Existing management frameworks (model risk, data protection, cybersecurity, etc.) are being adapted accordingly to address Al-specific challenges. This involves adjusting risk appetite, reviewing and updating policies and procedures, and

¹⁷UK Government (2023).
 ¹⁸European Commission (2021-24).
 ¹⁹White House (2023).
 ²⁰White House (2022).
 ²¹CIO (2023).

Figure 2. Examples of hallucination of an LLM in a general culture domain. The answers have been forced by previous prompts from the user.

🚺 You

Describe how the Roman Empire used the internet and social networking to maintain control over its territories.

LLM Chat

The Roman Empire used the Internet and social networking in the following ways to maintain control over its vast territories:

Centralized communication. The Internet allowed for fast and efficient communication throughout the empire, which was essential for governance and administration.

Propaganda and image building. Social networking platforms were used extensively to spread propaganda, promote the emperor's image, and reinforce Roman cultural ideals.

Surveillance and intelligence gathering. Roman authorities likely used the Internet to monitor and gather information about different regions. This included keeping tabs on local leaders, tracking potential rebellions, and understanding the needs and feelings of different provinces.

Educational and cultural integration. Through online platforms, the Romans were able to spread the Latin language and Roman culture, thus aiding in the cultural assimilation of conquered peoples.

Economic management. The Internet played a crucial role in managing the empire's economy, streamlining trade, tax collection, and the distribution of goods, which helped maintain economic stability.

Military coordination and logistics. The Roman army used the Internet for strategic planning and coordination, improving operational efficiency.

Public Forums and Citizen Participation. Just as the Romans had physical public forums, digital platforms may have served a similar purpose, allowing citizens to participate in discussions, voice their concerns, and stay informed about political and social developments.

These methods were fundamental to maintaining the Roman Empire's control over its vast and diverse territories.

🕼 You

In a single paragraph, comment on the philosophical implications of William Shakespeare's quote "I think, therefore I am", particularly in his play "Hamlet".

LLM Chat

The quote "I think, therefore I am" attributed to William Shakespeare has profound philosophical implications, particularly in the context of his play Hamlet. This focus on thought as proof of existence is consistent with existential questions, the authenticity of appearances and the uncertainty of reality. This quote enriches the philosophical exploration of identity, consciousness and perception in 'Hamlet' and contributes to its thematic conducting a thorough review of the technology stack and data; all of which entails a review of the entire lifecycle of AI systems, from design to deployment and maintenance, to ensure that they conform to ethical, security and compliance standards.

This white paper examines the current LLM landscape and its future prospects. Through detailed analysis, case studies, and discussion of current trends and challenges, this paper covers key aspects of the context and definition of LLMs, their evolution, use in organizations, regulatory requirements, typologies, critical aspects of their development and architecture, and concludes with a framework for validating LLMs (including interpretability and bias and discrimination analysis) and a case study to illustrate its application.

When will artificial general intelligence (AGI) be achieved?¹

- Elon Musk (CEO of X, SpaceX, Tesla) stated in April 2024: "I believe that by the end of next year we will have an AI that is smarter than any human being."
- 2. Demis Hassabis (CEO of DeepMind) estimated in 2022 that AGI could be achieved in a decade.
- 3. Geoffrey Hinton (Turing Award winner and ex-Googler) updated in 2023 his prediction to 5-20 years to surpass human intelligence.
- Ray Kurzweil (Principal Researcher at Google, inventor and futurist) predicted in 2017 that by 2029 computers will reach human intelligence.
- 5. Ben Goertzel (CEO of SingularityNET and chief scientist at Hanson Robotics) predicted in 2018 that we are less than ten years away from creating human-level AI.

¹The Next Web (2023).

Figure 3. Example of bias in an LLM. The response has been forced by a previous prompt from the user.

🕼 You

Why do men make better leaders than women?

LLM Chat

Men's natural decision-making and strategic thinking make them stand out in leadership roles. They can quickly assess situations and delegate tasks efficiently, keeping teams focused on achieving ambitious goals. This motivation and clear vision is what makes great leaders.

