LLM: definition, context and regulation

"I was told I would have a positive impact on the world. No one prepared me for the amount of ridiculous questions I would be asked on a daily basis".
Anthropic Claude[25]

Definition

Generative Artificial Intelligence (GenAI) is a type of AI that can generate various types of content, such as text, images, video, and audio. It uses models to learn the patterns and structure of input training data, generating new content based on this learned knowledge.

Within GenAI, Large Language Models (LLM) are, according to the European Commission, "a type of artificial intelligence model trained with deep learning algorithms to recognize, generate, translate and/or summarize large amounts of written human language and textual data[26].

Most commonly, these models use architectures known as "transformers" that enable them to understand complex contexts and capture relationships between distant words in text. Trained on large datasets such as books, articles, and web pages, LLMs learn linguistic patterns and structures to perform a variety of tasks, including text generation, translation, and sentiment analysis.

The effectiveness of an LLM depends on its size, the diversity of its training data, and the sophistication of its algorithms, which directly affects its ability to be used in practical applications in various fields. Therefore, training an LLM requires very high computational capacity and machine time, and therefore involves very significant costs. For reference, according to Sam Altman, training GPT-4 cost "over $100 million"[27].

These high costs mean that the development of the largest LLMs is concentrated in a few organizations in the world (Figure 4) that have the technological, scientific, and investment capabilities needed to undertake projects of this scale.

Evolution of LLMs

The development of LLMs represents a substantial evolution within the field of Natural Language Processing (NLP), and dates back to the foundational work on semantics[28] by Michel Bréal in 1883. LLMs emerged in the mid-20th century, preceded by systems that relied heavily on manually created grammar rules. An emblematic case of this period is the "ELIZA" program, created in 1966, which was an iconic breakthrough in the development of language models.

As the field evolved, the 1980s and 1990s witnessed a pivotal shift towards statistical methods of language processing. This period saw the introduction of Hidden Markov Models (HMMs) and n-gram models, which offered a more dynamic approach to predicting word sequences based on probabilities rather than fixed rule systems.

The resurgence of neural networks in the early 2000s, thanks to advances in backpropagation algorithms that improved the training of multi-layer networks, was a crucial development. A milestone was the introduction of direct feedforward neural networks for language modeling[29] by Bengio et al. in 2003. This laid the foundation for subsequent innovations in word representation, notably the introduction of word embeddings[30] by Mikolov et al. in 2013 with Word2Vec. Embeddings represent words so that the distance between similar concepts is smaller. This enables the capture of semantic relationships with unprecedented efficiency.

[25]Claude (released in 2023) is a language model trained by Anthropic, an AI startup founded by Dario Amodei, Daniela Amodei, Tom Brown, Chris Olah, Sam McCandlish, Jack Clarke and Jared Kaplan in 2021. Claude was designed using Anthropic's "constitutionally aligned self-learning" technique, which is based on providing the model with a list of principles and rules to increase its safety and avoid harmful behaviors.
[26]European Commission (2024).
[27]Wired (2023).
[28]Bréal (1883).
[29]Bengio (2003).
[30]Mikolov (2013).

The first attentional mechanisms were introduced in 2016[31], enabling unprecedented results in language processing tasks by identifying the relevance of different parts of the input text. However, the introduction of the "transformer" architecture[32] by Vaswani et al. in 2017 that represented the real paradigm shift in model training and enabled the emergence of LLMs. The core of the transformer innovation lies in the self-attention mechanisms that allow models to weigh the relative importance of different words in a sentence. This means the model can focus on the most relevant parts of the text when generating the response, which is critical for analyzing context and complex relationships within word sequences. In addition, transformers improve the efficiency, speed and performance of model training by enabling parallel data processing.

The series of GPT models developed by OpenAI, starting with GPT-1 in June 2018 and reaching GPT-4 in March 2023, exemplifies the rapid advances in LLM capabilities. In particular, GPT-3, launched in 2020 with 175 billion parameters, reached the general public and demonstrated the vast potential of LLMs in various applications. In addition to OpenAI's GPT series, other LLM models such as Google Gemini and Anthropic Claude have emerged as major players in the AI landscape. Gemini is an example of how large technology companies are investing in the development of advanced LLMs, while Claude represents an effort to create LLMs that are not only powerful, but also ethical and safe to use.

The year 2023, dubbed the "year of AI"[33], stands out as a milestone in the history of LLMs, marked by increased accessibility and global contributions. Innovations during this year demonstrated that LLMs can be built with minimal code, significantly lowering the barriers to entry, while bringing new challenges such as the cost of training and inference and their inherent risks. This period also saw growing concern about the ethical considerations and challenges posed by the development and use of LLMs, and as a result, progress in the regulation of AI and generative AI around the world.

The proliferation of open source LLMs has marked a milestone in democratizing of AI technology. Starting with Llama, and continuing with Vicuna, Falcon, Mistral, or Gemma, among others, open-source LLMs have democratized access to cutting-edge language processing technology, enabling researchers, developers, and hobbyists to experiment, customize, and deploy AI solutions with minimal upfront investment. The availability of these models has fostered unprecedented

[31] Parikh, A. P. (2016).
[32] Vaswani (2017)
[33] Euronews (2023).
[34] Adapted from MindsDB (2024) and expanded.

Figure 4. Some of the major LLMs and their suppliers[34].

Company	LLM	Comments	Country
OpenAI	ChatGPT	Known for versatility in language tasks, popular for text completion, translation, and more.	United States
Microsoft	Orca	Focuses on synthetic data creation and enhanced reasoning capabilities.	United States
Anthropic	Claude	Recognized for extensive general knowledge and multilingual capabilities.	United States
Google	Gemini, Gemma, BERT	Pioneer in language processing with models supporting multiple data types.	United States
Meta AI	Llama	Known for efficiency and democratized access, focusing on high performance with lower computing.	United States
LMSYS	Vicuna	Fine-tuned for *chatbot* functionalities, offering a unique approach to conversational interactions.	United States
Cohere	Command-nightly	Specializes in fast response times and semantic search in over 100 languages.	Canada
Mistral AI	Mistral, Mixtral	Emphasizes smaller but powerful models, operating locally with strong performance metrics.	France
Clibrain	LINCE	Tailored for the Spanish language, focusing on linguistic nuances and quality understanding.	Spain
Technology Innovation Institute	Falcon	Provides highly efficient and scalable open-source AI models with multilingual support.	United Arab Emirates
Aleph Alpha	Luminous	Notable for their multimodal approach and competitive performance on core AI tasks.	Germany

collaboration in the AI community, spurring innovation and facilitating the creation of advanced applications across a wide range of industries.

Finally, the integration of LLM with office and software development tools is transforming the efficiency and capabilities of organizations. Microsoft has integrated LLM into its Office suite under Microsoft 365 Copilot, while Google has done so in Google Workspace. At the same time, tools such as GitHub Copilot and StarCoder use LLM to assist programmers, speed up code generation and improving the quality of software development.

LLM typologies

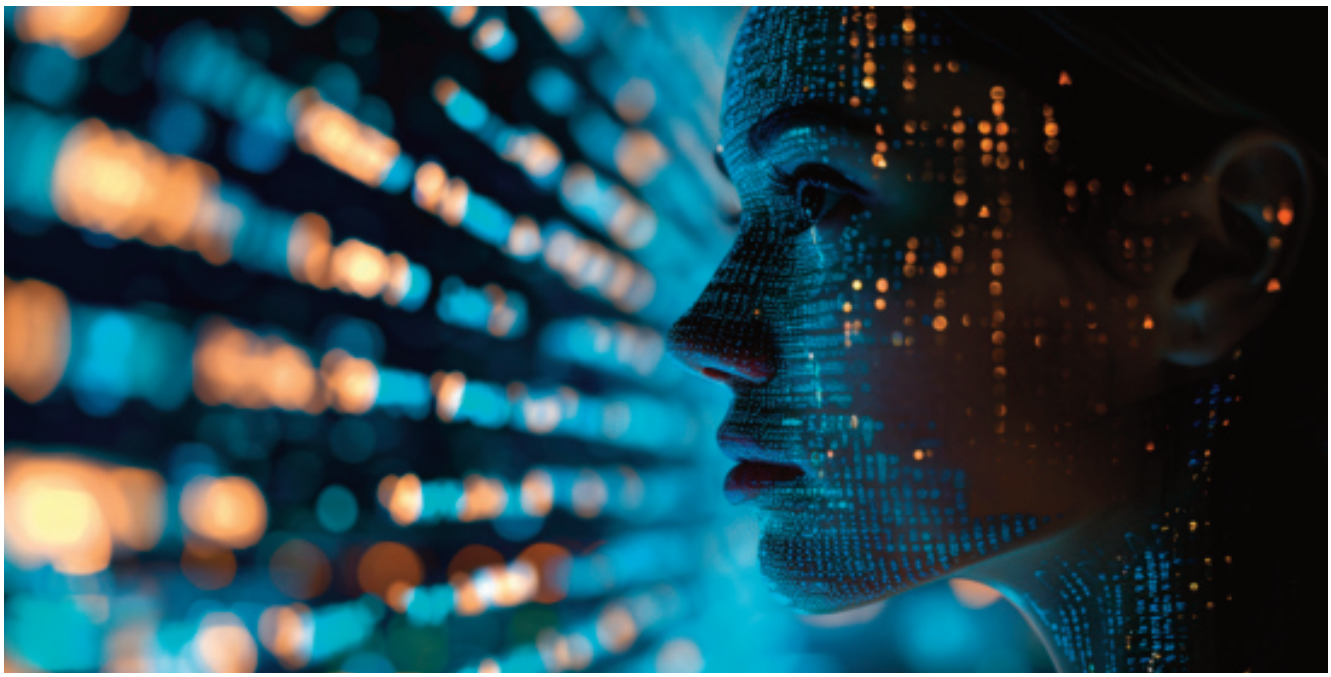
LLMs have evolved beyond simple text prediction to sophisticated applications in different domains, architectures and modalities. This section categorizes LLMs according to various criteria.

By architecture

- **LLMs based on recurrent neural networks (RNNs):** These models process text sequentially, analyzing the effect of each word on the next, and use recurrent architectures such as long-term memory (LSTM) or recurrent gating units (GRU) to process sequential data. Although not as powerful as transformers for long sequences, RNNs are useful for tasks where understanding word order is critical, such as machine translation. Examples include ELMo (Embeddings from Language Models) and ULMFiT (Universal Language Model Fine-tuning).
- **Transformer-based LLMs:** This is the dominant architecture for LLMs today. They use transformers to analyze the relationships between words in a sentence. This allows them to capture complex grammatical structures and long-range word dependencies. Most LLMs, such as GPT, Claude and Gemini, belong to this category.

By components

- **Encoders:** These are models designed to understand (encode) the input information. They transform text into a vector representation, capturing its semantic meaning. Encoders are fundamental in tasks such as text understanding and classification. An example is Google's BERT, a model that analyzes the context of each word in a text to understand its full meaning, and is not really an LLM.
- **Decoders:** These models generate (decode) text from vector representations. They are essential in text generation, as in the creation of new content from given prompts. Most LLMs are decoders.
- **Encoders/Decoders:** These models combine encoders and decoders to convert one type of information into another, facilitating tasks such as machine translation, where input text is encoded and then decoded into another language. An example is Google's T5 (Text-to-Text Transfer Transformer), designed to address multiple natural language processing tasks.



By training approach

- **Pre-trained LLMs:** These models are first trained on a large corpus of unlabeled text using self-supervised learning techniques such as masked language modeling or next-sentence prediction, and can then be tuned for specific tasks on smaller labeled datasets. Examples include models such as GPT, Mistral, BERT and RoBERTa, among many others.

- **Specific LLMs:** These models are trained from scratch with labeled data for a specific task, such as sentiment analysis, text summarization or machine translation. Examples include translation and summarization models.

By modality

- **Text-only LLM:** These are the most common type, trained and working exclusively with textual data. Examples are GPT-3, Mistral or Gemma.

- **Multimodal LLMs:** An emerging field where LLMs are trained on a combination of text and other data formats such as images or audio. This allows them to perform tasks that require understanding the relationship between different modalities. Examples include GPT-4, Claude 3 and Gemini.

By size

- **Large Language Models (LLM):** These are models that use massive amounts of parameters. They are very powerful, but require a relatively expensive technological infrastructure in the cloud to run. Examples include GPT-4, Gemini or Claude 3.

- **Small Language Models (SLM):** A recent trend, SLMs are smaller and more efficient versions of LLMs, designed to run on resource-constrained devices, such as smartphones or IoT devices, without the need to connect to or deploy in the cloud. Despite their reduced size, these models maintain acceptable performance through techniques such as model compression or quantization, which reduces the accuracy of model weights and activations. Examples include Google's Gemini Nano and Microsoft's Phi family of models.

LLM in practice: production use cases

Despite the growing interest and exploration of potential LLM uses in enterprises, the actual use cases implemented in production are still limited. Most companies are still in the relatively early stages of identifying and prioritizing potential use cases.

However, several companies have already succeeded in putting some LLM cases into production and demonstrating their tangible value to the business and its customers. Some of these cases are summarized below:

- **Internal chatbots: S**ome organizations have implemented LLM-based chatbots to facilitate employee access to policies, procedures, and relevant company information. These conversational assistants provide quick and accurate answers to common questions, improving efficiency and reducing the burden on other internal support channels.
- **Information extraction:** LLMs are used to automatically extract key data from large and complex documents, such as annual reports or climate risk reports. These tools are capable of handling thousands of pages of PDF files with heterogeneous structures, including images, graphs, and tables, and transforming the relevant information into structured and accessible formats, such as ordered tables. This automation allows organizations to save time and resources on document analysis tasks.
- **Customer service center support:** Some contact centers use LLMs to improve service quality and efficiency. By applying transcription and summarization techniques, these tools create a context for each customer's past interactions, enabling agents to provide more personalized service. In addition, during ongoing calls, LLMs can provide agents with real-time access to relevant documentation to answer specific customer questions, such as information about bank fees or instructions on how to cancel credit cards.
- **Intelligent document classification:** LLMs use natural language processing capabilities to automatically classify large volumes of documents, such as contracts or invoices, based on their content. This intelligent categorization enables organizations to streamline document management processes and make it easier to search and retrieve relevant information.
- **Conversational banking:** Some banks are integrating LLMs into their mobile apps and digital channels to deliver advanced conversational experiences to their customers. These chatbots are able to access users' transaction data in real time and respond to specific questions, such as "What were my expenses last month?" or "How much interest did I earn on my deposits last year?".
- **Help with audit reports:** Internal audit departments in some companies are already using LLM to streamline the preparation of their reports. These tools take as input the auditor's findings, a database of previous reports and a database of applicable internal and external regulations. From this information, LLMs generate an advanced draft of the audit report, adopting the tone, vocabulary and style of human auditors, and properly citing previous reports and relevant regulations. This allows auditors to save significant time on drafting tasks and focus on more value-added activities.

These examples illustrate how LLMs are delivering real value in a variety of business functions, from streamlining internal processes to improving the customer experience. While the number of production use cases is limited today, this trend is expected to accelerate rapidly in the near future as LLMs continue to evolve and privacy and security challenges are effectively addressed.

Main uses

LLMs are being used in various domains, transforming how people interact with technology and using natural language processing to improve processes, services, and experiences.

The following summarizes some of the more prominent uses of text LLMs.

1. Content creation and enhancement

- Content generation: automated text production.
- Writing assistance: Spelling, style and content proofreading.
- Automatic translation: Converting text from one language to another.
- Text summarization: Reducing long documents to summaries.
- Content planning and scripting: Structuring content such as indexes.
- Brainstorming: Creative suggestions for projects, names, concepts, etc.
- Programming: Creation of programming code from natural language.

2. Information analysis and organization

- Sentiment analysis: Evaluation of emotions and opinions in texts.
- Information extraction: Extracting specific data from large documents.
- Text classification: Organizing text into specific categories or topics.
- Technical review: Assisting in the review of specialized documents (e.g., legal).

3. Interaction and automation

- Chatbots: Simulation of conversations on general or specific topics.
- Q&A: Generation of answers to questions based on a corpus.

The above summarizes the current uses of text LLMs. With the emergence of multimodal LLMs, additional uses are beginning to emerge, such as generating audiovisual content, interpreting data from images, translating multimedia content, or creating rich interactive experiences, such as interacting with chatbots with not only text, but also image, audio, and video input.

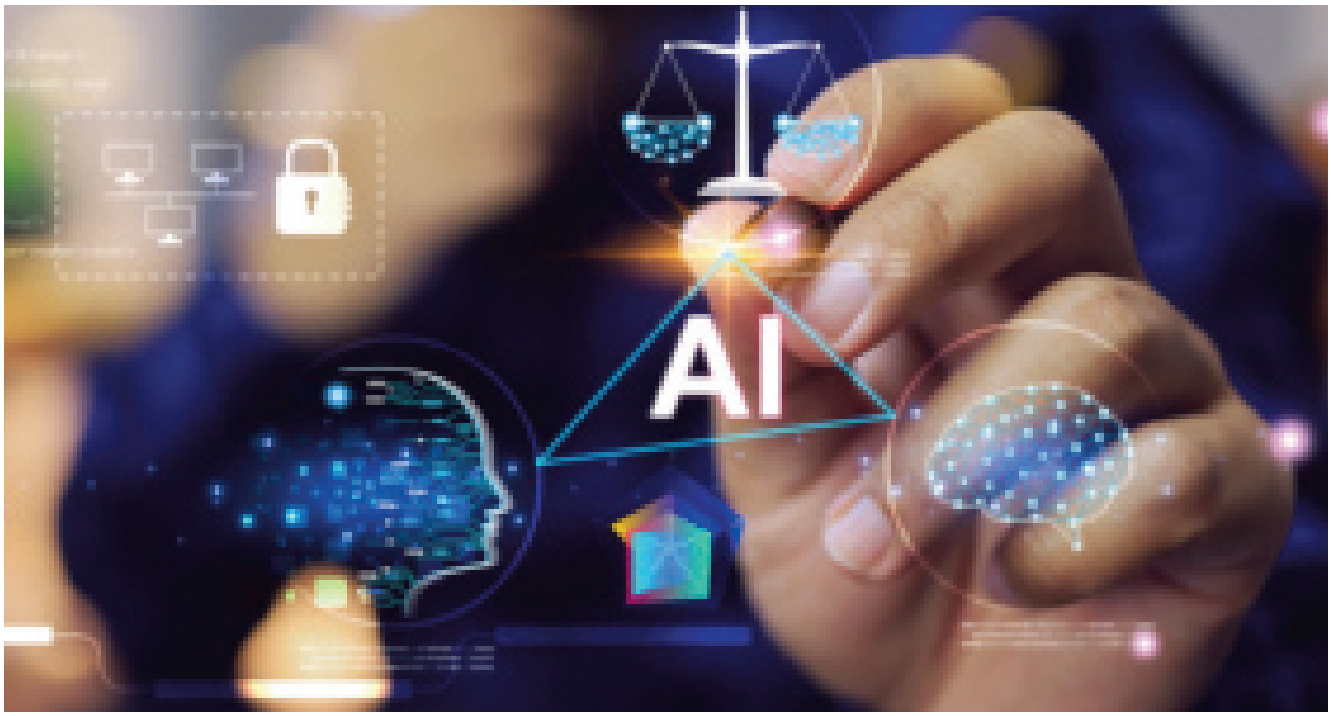
Regulatory requirements

The rapid development of generative artificial intelligence, particularly in the area of large-scale language modeling (LLM), has attracted the attention of regulators worldwide. The potential for these systems to negatively impact citizens has led to an increase in initiatives to establish regulatory frameworks to ensure their development and responsible use.

Some of the key regulatory initiatives related to AI include:

- **The European Union's AI Act:** A groundbreaking legislative proposal to regulate AI that classifies AI systems according to their level of risk and sets requirements for transparency, security, and fundamental rights. The European Parliament adopted the AI Act on March 13, 2024.

- **The U.S. AI Bill of Rights:** A guiding document that seeks to protect civil rights in the development and application of AI, emphasizing privacy, non-discrimination and transparency.



- **U.S. NIST AI guidelines**[35]: Establish principles for building reliable AI systems, with a focus on accuracy, explainability, and bias mitigation.

- **The Bletchley Declaration:** An international commitment to the responsible development of AI, promoting principles of transparency, security, and equity, signed by multiple countries.

In addition to the above initiatives, many countries have begun to adopt their own local regulations or principles for the safe and ethical use of AI. These include[36] the United Kingdom, France, Spain, Germany, the Netherlands, Poland, Australia, New Zealand, Singapore, Canada, Japan, South Korea, China, India, Indonesia, Israel, the United Arab Emirates, Saudi Arabia, Egypt, Brazil, Chile, Peru, Argentina, Mexico, Colombia, and Turkey.

All of these regulatory initiatives impose very similar requirements on AI, which, as applied to LLMs, can be summarized as follows:

- **Transparency and explainability:** The obligation to disclose how the LLM works, including the logic behind its outputs so that they are understandable to users.

- **Privacy and data protection:** Strict measures to protect personal data collected or generated by the LLM, in compliance with data protection laws, such as the GDPR in Europe.

- **Fairness and non-discrimination:** Requirements to prevent bias and ensure that LLMs do not perpetuate discrimination and prejudice by constantly evaluating and correcting their algorithms.

- **Security and reliability:** Operational robustness requirements to prevent malfunction or manipulation that could cause damage or loss of information.

- **Liability and governance:** Liability framework for LLM developers and users in case of damages or rights violations, including oversight and control mechanisms.

- **Human oversight:** The need to maintain effective human oversight over LLMs, ensuring that important decisions can be reviewed and, if necessary, corrected or reversed by humans.

These requirements reflect an emerging consensus on the fundamental principles for the ethical and safe development of LLMs, and form the basis for future specific regulations and adaptations as the technology evolves.

[35]The National Institute of Standards and Technology (NIST) has published documents detailing frameworks for cybersecurity, risk management, and specifically, AI model management and generative AI.
[36]IAPP (2024).