Management Solutions
Making things happen

The rise of Large Language Models: from fundamentals to application

Design and Layout
Marketing and Communication Department
Management Solutions

Photographs
Photographic archive of Management Solutions
AdobeStock
Midjourney

Index

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 Altman[1]

"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.

GenAI[3] 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, GenAI is finding potentially revolutionary applications[4] 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 GenAI, LLMs (such as OpenAI 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].

GenAI, 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 stages[6] 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.

[1] Samuel Harris Altman (b. 1985), American entrepreneur, founder and CEO of OpenAI.
[2] S. Pichai (2023). Interview on 60 Minutes Overtime.
[3] Gartner (2023).
[4] See a taxonomy and extensive collection of GenAI use cases in iDanae 2Q23 (2023) and in Gozalo-Brizuela, R., and Garrido-Merchán, E.C. (2023).
[5] Fischer (2021).
[6] Lam (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 Microsoft[7] study, integrating LLM as a co-pilot in office automation tools results in time savings ranging from 27% to 74% without compromising quality[8]. In the case of SMBs, the level of adoption is still limited, creating an even greater risk of a technology gap for this segment[9].

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 medium-term 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" experiment[10], lacks the capacity for abstraction and association of concepts to symbols, an attribute unique to the human mind.

[7] Cambon (2023). Study conducted by Microsoft on AI and productivity.
[8] In 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.
[9] IndesIA (2024).
[10] Searle, 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 AI	**Narrow Non-AI** Calculator software; compiler	**General Non-AI** Human-in-the-loop computing, e.g., Amazon Mechanical Turk
Level 1: Emerging Equal to or somewhat better than an unskilled human	**Emerging Narrow AI** 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 AI** 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 AI** 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 AI** Deep Blue: chess-playing computer developed by IBM that defeated the world champion in 1997. AlphaGo: an AI 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 AI** AlphaFold: predicts protein structures with high accuracy AlphaZero: self-taught AI that masters games like chess, Go, and shogi StockFish: a powerful open-source chess engine	**Artificial Superintelligence (ASI)** Not yet achieved

According to several experts[11], 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., chess-playing AI), Google DeepMind estimates[12] 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, including[13] 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 replacement[16]; 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.

[11] Pew Research Institute (2023).
[12] Google DeepMind (2023).
[13] See iDanae 2Q23 (2023).
[14] Xu (2024).
[15] Weizenbaum, J. (1966).). The "Eliza Effect" is the phenomenon whereby users attribute comprehensibility and emotional responses to a computer program.
[16] Although 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 GenAI will have a positive impact and boost productivity across all industries.

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.

In the face of these challenges, regulators are taking proactive steps at the national and international levels to address the risks and opportunities of AI. Of particular note is the Bletchley Declaration[17], 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 AI.

For its part, the European Union, with the imminent implementation of the Artificial Intelligence Act[18], 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 Order[19], issued on October 30, 2023, and the Blueprint for an Artificial Intelligence Bill of Rights[20] 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 AI strategy (with a particular on GenAI and LLMs), designing their AI adoption plan, and adapting their structures, including the creation of GenAI centers of excellence and the incorporation of new figures such as the Chief AI Officer[21]. Existing management frameworks (model risk, data protection, cybersecurity, etc.) are being adapted accordingly to address AI-specific challenges. This involves adjusting risk appetite, reviewing and updating policies and procedures, and

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

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.

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

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.
4. 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.

[1] The Next Web (2023).

Executive summary

"Artificial intelligence is not a substitute for human intelligence; it is a tool to amplify human creativity and ingenuity".
Fei-Fei Li[22]

LLM: context, definition and regulation

1. Generative Artificial Intelligence (GenAI), and within it Large Language Models (LLM), represents a significant advance in the field of AI, defining a new generation of human-machine interface where communication is done through natural language, with revolutionary applications in all sectors, including education, health, finance, and commerce. However, its development and use also pose significant risks and challenges that must be addressed.

2. LLMs are AI models trained to recognize, generate, translate and summarize large amounts of text. They use architectures such as transformers and are trained on large datasets to learn linguistic patterns and structures. Their effectiveness depends on the size in terms of the number of parameters, structure, diversity of training data and sophistication of their algorithms.

3. LLMs have evolved very rapidly, from the first rule-based models to today's transformer-based models. Important milestones include the introduction of transformer architecture and self-healing mechanisms, and the first commercial LLMs such as GPT. The year 2023 was key, with increased accessibility, global contributions, and the proliferation of open source LLMs.

4. LLMs have numerous applications, such as content creation and enhancement, information analysis and organization, and task interaction and automation. With the emergence of multimodal LLMs, new possibilities are opening up for generating rich audiovisual content and interactive experiences.

5. Regulators are taking steps to address the risks and opportunities of AI, with initiatives such as the EU AI Act, the U.S. AI Bill of Rights and the Bletchley Declaration. Key requirements include transparency, privacy, fairness, security, accountability and human oversight.

LLM: development and deployment

6. LLM development involves several critical components and decisions, such as data selection and preprocessing, tokenization and embedding, pre-training, quantization, and fine-tuning. In particular, the high cost of training often leads to the decision to use a pre-trained model or an open-source model, and to limit fine-tuning to data relative to the application being developed. Implementation requires integration, monitoring, and ethical and legal considerations.

7. Training models is a crucial aspect that influences their effectiveness. Factors such as the quantity and quality of the training data, the model architecture and the learning algorithms used can significantly impact the performance and generalization of an LLM.

8. The most common architecture for LLMs are transformers, which use self-learning mechanisms that allow the model to find relationships between different parts of the text, process them, and generate new text. They have demonstrated exceptional performance in a variety of natural language processing tasks. Variants and extensions aim to improve their efficiency and scalability.

[22] Fei-Fei Li (b. 1976). Co-director of the Stanford Institute for Human-Centered Artificial Intelligence and IT Professor at the Graduate School of Business, known for creating ImageNet and AI4ALL, a non-profit organization working to increase diversity and inclusion in the field of artificial intelligence.

9. LLMOps is a methodology for managing the entire LLM lifecycle, addressing challenges such as managing large volumes of data, scaling computational resources[23], monitoring and maintenance, versioning, and reproducibility.

10. Key challenges for LLMs include biases and hallucinations, lack of explainability and transparency, data quality and accessibility, privacy and security issues, and high resource consumption. There are also challenges of dependency, risk of malicious use, intellectual property issues, and scalability.

LLM: Validation Framework

11. Validation of LLMs is essential to ensure their safe and responsible use, and it is appropriate to take a broad perspective covering the various risks involved. A multi-dimensional validation framework should cover aspects such as model risk, data management, cybersecurity, legal and operational risks, ethics and reputation.

12. LLM validation should be articulated through a combination of quantitative metrics and human judgment techniques. The choice of techniques will depend on the characteristics of the use case, such as level of risk, public exposure, personal data processing and line of business.

13. Emerging trends in LLM validation include explainability[24], the using LLMs to explain other LLMs, attribution scoring, continuous validation, collaborative approaches, prompt engineering, ethical and regulatory alignment, and machine unlearning techniques.

Case study

14. The case study presented illustrates the application of a custom validation framework to a company's internal policy chatbot. The process involved defining the case, designing the validation approach, running quantitative and qualitative tests, and interpreting results.

15. The chatbot validation results showed satisfactory overall performance, with strengths in accuracy, consistency, adaptability and scalability. Areas for improvement were identified in the areas of explainability, bias mitigation and security. It was recommended to proceed with implementation, applying the suggested improvements and establishing a continuous monitoring and improvement plan.

Conclusion

16. In conclusion, LLMs have significant potential to transform multiple sectors, but their development and deployment also pose significant challenges in transparency, fairness, privacy and security. To reap the benefits of LLMs in a responsible way, it is crucial to establish a robust AI governance framework that comprehensively addresses these challenges, including a rigorous, multi-dimensional approach to validation that covers the entire lifecycle of the models. This is the only way to ensure that LLMs are reliable, ethical and aligned with the values and goals of organizations and society at large.

[23] Management Solutions (2022). Auto Machine Learning, towards model automation.
[24] Management Solutions (2023). Explainable Artificial Intelligence (XAI): challenges in model interpretability.

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

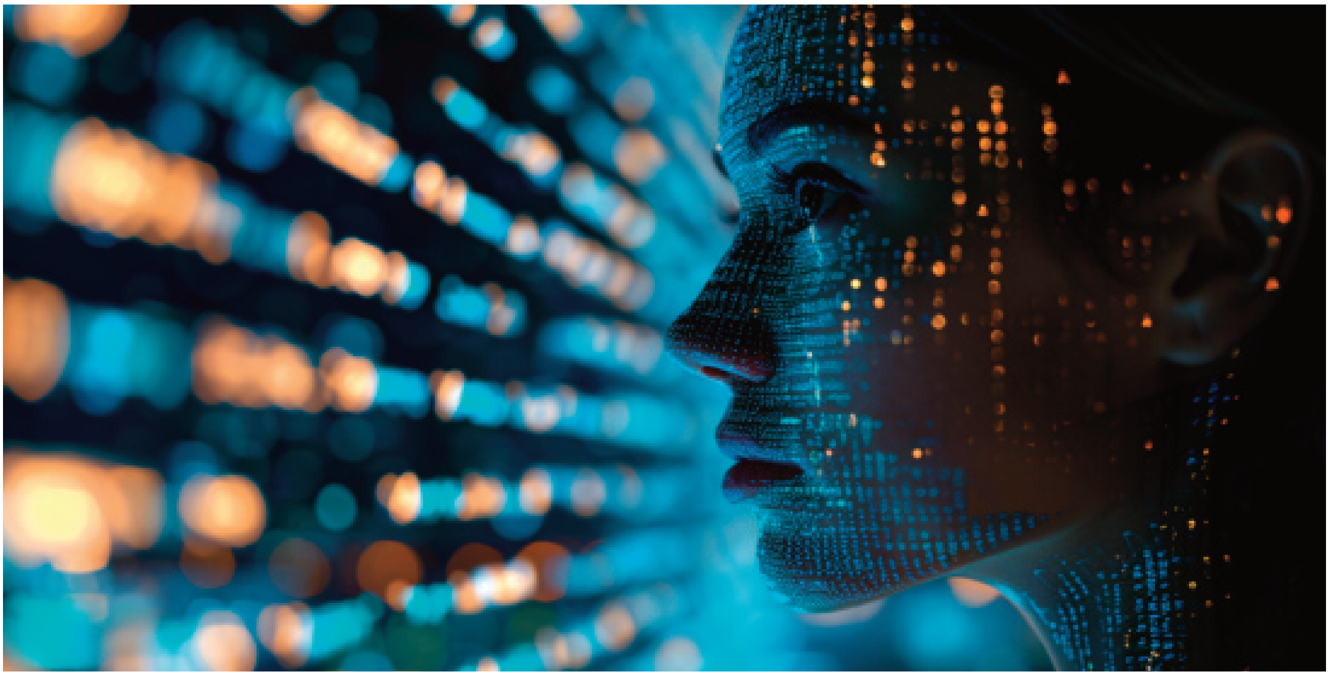
LLMs have evolved beyond simple text prediction to sophisticated applications in different domains, architectures and modalities. This section categorizes LLMs according to various criteria.

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 (LLMs):** 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 (SLMs):** 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.

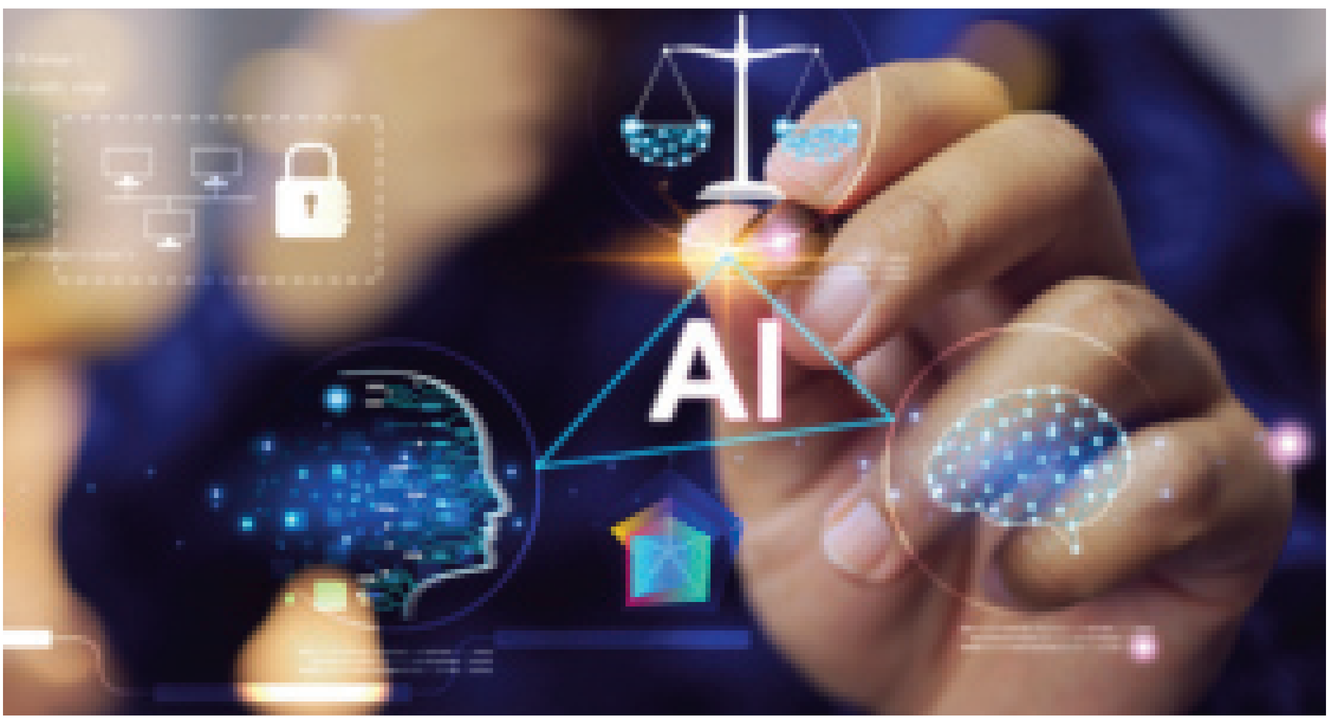
Regulatory requirements

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).

LLM: development and deployment

"Generative AI is the key to solving some of the world's biggest problems, such as climate change, poverty and disease. It has the potential to make the world a better place for everyone".

Mark Zuckerberg[37]

This section discusses the key aspects of the LLM development and deployment process. It examines key components such as data and model architecture, as well as the pre-training, fine-tuning, and implementation phases. It also discusses the key challenges and considerations that must be considered to ensure ethical, robust development aligned with an organization's goals.

Key aspects of LLM development

LLM development is a complex process involving many components and critical decisions. The following is a description of the main components that need to be known about LLM development, and some key aspects about them.

Data

Data are the foundation upon which LLMs are built, and their quality, diversity, and representativeness directly impact the performance and bias of the resulting model. Addressing challenges related to intellectual property, data quality, and preprocessing is essential to developing robust, unbiased, and accurate LLMs. As regulations and best practices in this area evolve, we will llikely see an increased emphasis on responsible and transparent use of data in LLM training.

Some key aspects about LLM training data are:

- Training corpus[38]: LLMs are trained on large corpora of data, often extracted from the internet, containing billions of words and spanning a wide range of domains and genres, such as books, news articles, web pages, social networks and more. These massive corpora enable LLMs to learn patterns and representations of language on a large scale, giving them an unprecedented ability to understand and generate coherent, contextualized text. For example, common corpora for training include BookCorpus[39], Gutenberg[40], Wikipedia[41] or CodeParrot[42].

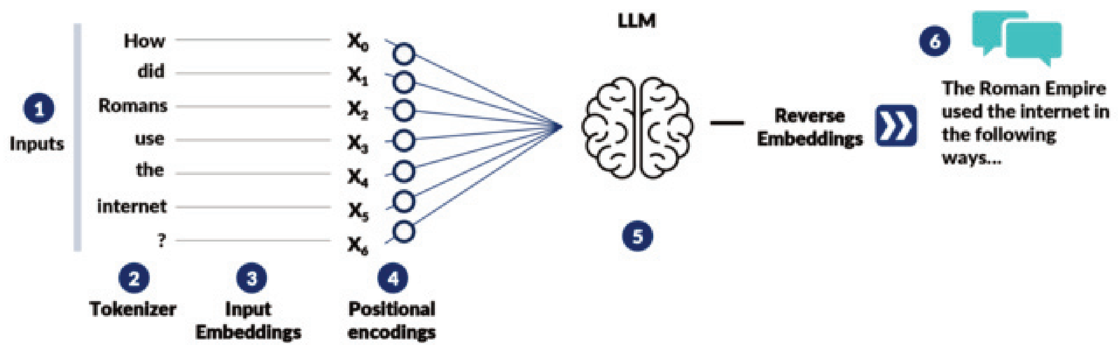
- Intellectual property and copyright[43]: Extracting and using Internet data for LLM training raises challenges related to intellectual property and copyright. Much of this data is protected by copyright, and its use without permission or adequate compensation can be problematic. The AI Act in Europe addresses this issue by imposing new requirements on LLM developers, such as the obligation to disclose the data sources used and to obtain the necessary licenses.

- Data quality and representativeness[44]: Like any model, an LLM is only as good as the data used to train it. If the data is of poor quality, biased or unrepresentative, the model may inherit these problems and produce inaccurate, unfair or inappropriate results. Therefore, it is critical to ensure that training corpora are diverse, balanced, and adequately represent different demographics[45], opinions, and perspectives.

- High quality data initiatives[46]: Some recent initiatives focus on building LLMs with fewer parameters, but higher quality data, such as smaller, but carefully selected and filtered[47] training corpora that include high quality content like books, scientific articles, and respected publications. These filters can be limited, for example, to a single language, or to an industry or subject area, drastically reducing the size of the corpus. This strategy can result in LLMs with better performance and less bias than models trained on massive unfiltered data.

[37] Mark Zuckerberg (n. 1984), co-founder and CEO of Facebook and Meta, one of the world's largest social networking, technology and artificial intelligence companies.
[38] Liu (2024).
[39] Soskek (2019).
[40] Project Gutenberg (2024).
[41] Wikipedia Dumps (2024).
[42] Hugging Face Datasets (2024).
[43] Li (2024), Chu (2023).
[44] Alabdulmohsin (2024).
[45] Yogarajan (2023).
[46] Sachdeva (2024).
[47] Tirumala (2023).

Figure 5. Stages of LLM input data processing.



- Data preprocessing and labeling[48]: Before training or fine-tuning an LLM, the data must be preprocessed and, in some cases such as supervised fine-tuning or using a specific dataset, labeled. Preprocessing involves cleaning and formatting the data[49], removing noise and errors, and applying techniques such as tokenization and normalization (e.g., LayerNorm[50] for Transformers).

Tokenization and encoding

Tokenization refers to the process of breaking down text into smaller units called "tokens", which are the units processed by the LLM during training and response inference. These tokens can be words, parts of words (e.g. lemmas), or characters. For example, one of the simplest ways to generate tokens is to partition the corpus according to the spaces between words. Encoding is the process of representing these text units in numerical form so that the model can process them.

Some key points about tokenization in LLM:

- It is performed on the available text corpus to optimally divide the original text into smaller units. The end result of tokenization is an encoding.

- Encodings have a significant impact on the performance of the LLM[51], as they define the minimum processing unit it will receive and determine the vocabulary the LLM has access to.

- There are several encoding algorithms on the market[52] that differ in the way they divide the text based on words, phrases or sentences, use of spaces, capitalization or formatting, appearance of characters in different languages, or errors present in the text.

- The main encodings[53] used are BytePairEncoding, SentencePieceEncoding and WordPieceEncoding.

The tokenization result is used as a starting point in the embedding model.

Embedding

Embeddings are numerical representations of words, phrases, sentences, or even paragraphs that capture their semantic meaning and the relationships between them. They are based on the LLM input corpus, which is divided into tokens. They are a fundamental component of LLMs and play a crucial role both in the pre-training, fine-tuning, and subsequent use of these models.

[48]Chen (2023).
[49]Wenzek (2019), Penedo (2023).
[50]Zhao (2023).
[51]Rejeleene (2024).
[52]Minaee (2024).
[53]Kudo (2018).

Embeddings in LLMs:

- They are designed to capture semantic relationships between words, so that words with similar meanings have similar vectors. This allows the model to understand the similarity and analogies between words and concepts.
- They are not universal values, but will vary from one model to another, depending on the vector space in which they have been defined.
- They are contextual, meaning that the representation of a word can vary depending on the context in which it appears. This allows nuances of meaning to be captured and polysemous words to be disambiguated. The embeddings are not predefined but are learned from training data based on the LLM embedding model. During pre-training, the model adjusts the embeddings to maximize their ability to predict words in context (e.g. through embedding frameworks such as SentenceTransformers). However, the embeddings alone are already a model that needs to be tuned during the process.

Pre-training

Pretraining is a fundamental stage in LLM development, during which models acquire general and deep language knowledge from large amounts of unlabeled data. Although this process is computationally intensive and costly, it enables model adaptation to a wide range of tasks.

The main goal of pre-training is for the model to acquire a broad and deep knowledge of the language, including its structure, semantics, syntax, and context. During this process, the LLM learns to predict words or text fragments (i.e., tokens) based on the surrounding context, allowing it to capture complex linguistic relationships and patterns. This general knowledge becomes the basis for fine-tuning the model for specific tasks.

There are several popular techniques for LLM pre-training, such as:

- Autoregressive language modeling or unidirectional modeling (e.g., autoregressive modeling[54]), which consists of training the model to predict the next word or text fragment given the previous context. This task allows the model to learn the conditional probabilities of the language and generate coherent text. Examples include the GPT and Claude models.

[54]Devlin (2018), Liu (2022).

Types of embeddings

Embeddings are used in LLMs in order to establish a metric that defines the similarity between word meanings and to incorporate information about the position of words in a sentence. This is crucial, since word order affects meaning. There are three main types of positional embeddings:

- Absolute positional embedding[1]: Assigns to each word - or to each minimal text unit or token - a vector representing its exact position in the sentence (e.g., first, second, third position, etc.).
- Relative positional embedding[2]: Instead of being based on absolute positions, it represents the position of a word relative to the others (e.g. two words before, one word after, etc.).
- Rotary positional embedding[3]: Combines absolute and relative positional information, using trigonometric functions to create more complex vector representations.

In a transformer, a simple positional embedding for a word at a given position can be represented mathematically using sine and cosine functions. Specifically, a positional embedding E for a token i with position P can be represented mathematically in its simplest form as:

$$E\,(P, 2i) = \sin \frac{P}{10000^{\frac{2i}{d}}}$$

$$E\,(P, 2i+1) = \cos \frac{P}{10000^{\frac{2i}{d}}}$$

where P is the position of the token in the input sequence, and d is the dimension of the hidden layers of the transformer.

The choice of positional embedding type can affect LLM performance by determining the amount and type of positional information available to the model during training.

[1]Vaswani (2017).
[2]Shaw (2018).
[3]Su (2021).

- The non-autoregressive model[55], used in models such as Gemini, in which the response is not obtained sequentially word by word, but is transformed and refined as a whole.

- Masked language modeling[56], popularized by models such as BERT, which consists of randomly masking some words in the input text and training the model to predict these masked words based on the surrounding context. This technique allows bidirectional learning and a better understanding of the context. Some LLM architectures (e.g., bidirectional transformers) use this technique.

- Sequence-to-sequence modeling[57] (e.g., seq2seq[58]), where the model is trained to generate text sequences based on other input sequences. This is used in models such as T5, BART or ProphetNET.

- Contrastive pre-training[59], used in models such as CLIP and ALIGN[60], involves training the model to identify text-image pairs that are semantically related , allowing it to learn multimodal representations and transfer knowledge between different modalities[61].

LLM pre-training is a computationally intensive process that requires enormous amounts of data, time and hardware resources. The largest models can have on the order of 1 trillion (10^{12}) parameters and require thousands of high-end GPUs for weeks or months of training. This makes pre-training extremely expensive and affordable for only a few companies and organizations in the world with the necessary resources.

Quantification

During LLM training, neuron weights are adjusted to make more accurate predictions. These weights are typically stored as high-precision numbers, which can result in large and computationally expensive models.

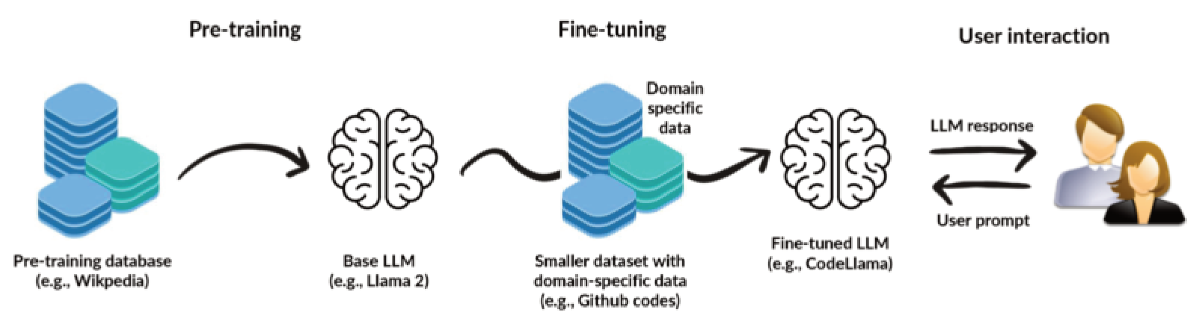
Post-training quantization is a technique[62] that allows the accuracy of model parameters to be reduced without significantly affecting model performance. For example, neural networks that store their parameters in 32-bit floating-point numbers can be switched to using only 16-bit or 8-bit numbers, depending on the type of quantization. This results in smaller and faster models because they require less memory and, with the right hardware, can perform operations more efficiently.

Recently, there has been a trend to develop small language models (SLMs), or even "tiny LLMs"[63], models that maintain high performance despite their much smaller size. These compact models are achieved by combining techniques, including post-training quantization.

By skillfully applying these techniques, SLMs and tiny LLMs can in some cases achieve performance comparable to that of much larger models[64], making them attractive for applications where computational or memory resources are limited.

[55] Xu (2021).
[56] Devlin (2019), Sinha (2021).
[57] Lee (2022).
[58] Sutskever (2014).
[59] Zeng (2023).
[60] Jia (2021).
[61] Cui (2022).
[62] Li (2024).
[63] Tian (2024).
[64] Fu (2024).

Figure 6. LLM fine-tuning.



Fine-tuning, instruction-tuning and RAG

Fine-tuning is the process of adapting a pre-trained LLM to a specific task using a smaller data set. This technique makes it possible to take advantage of the general knowledge acquired during pre-training and specialize it to achieve high performance on the target task.

The main goal of fine-tuning (Figure 6) is to adapt a pre-trained LLM to a specific task, such as sentiment classification, question answering, machine translation, or summary generation. During this process, the model learns to use its general knowledge of the language and apply it effectively to the specific domain and requirements of the task at hand. Commercially available LLMs, whether proprietary or open source, are typically pre-trained (and therefore general-purpose), but have not been fine-tuned to adapt to a specific purpose.

Fine-tuning has several important advantages:

- **Leverages prior knowledge:** By starting from a pre-trained model, fine-tuning allows the vast general knowledge of the language acquired during pre-training to be leveraged, accelerating learning and improving performance on the specific task.
- **Requires less data and resources:** Compared to training from scratch, fine-tuning requires much less labeled data and computational resources, making it more accessible and cost-effective for a wide range of organizations and applications.
- **Enables specialization:** Fine-tuning allows LLMs to be tailored to specific domains and tasks, resulting in highly specialized and effective models for specific applications.
- **Facilitates learning transfer:** Fine-tuned models can receive additional fine-tuning for related tasks, enabling learning transfer and the creation of even more specialized models with relatively little additional data.

Despite its benefits, fine-tuning also presents some challenges:

- **Overspecialization**[65]: If the model is fine-tuned on a data set that is too specific, it may lose some of its generalization ability and perform poorly on unknown or slightly different data.

[65] Wang (2024).

Training LLM: loss functions

LLMs, like other deep learning models, learn by adjusting their parameters to minimize a loss function. This function measures the difference between the model's predictions and the expected outcomes, and guides the model toward better performance.

The choice of loss function depends on the type of task for which the LLM is being trained. For example, for a model that predicts the next word in a sentence (autoregressive language modeling), a common function is cross-entropy. This function compares the probability distribution of the words predicted by the model with the actual distribution observed in the training data.

Mathematically, the cross-entropy loss function for an autoregressive model can be expressed as the sum of the negative logarithms of the probabilities assigned to the correct words at each position in the sequence.

Specifically, given a loss function such as cross-entropy and a training typology such as autoregressive language modeling, the loss function to be minimized can be defined as:

$$f_L(\varphi) = \sum_{i=1}^{N} -\log P(x_i \mid x_i, 1 \ldots N, \varphi)$$

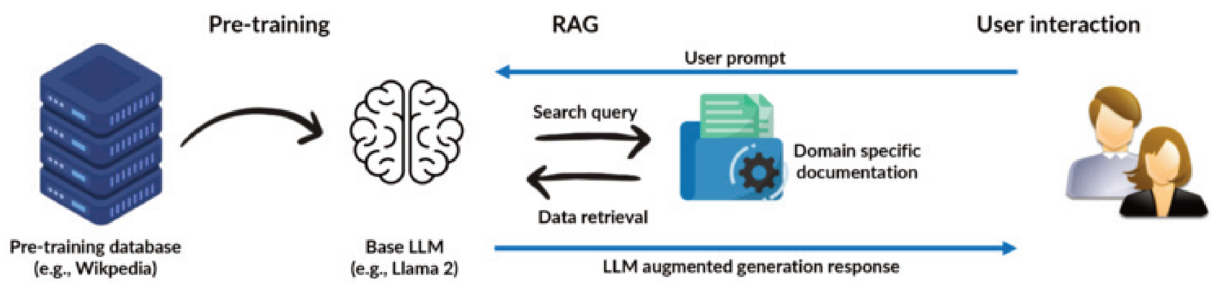
where φ represents the model parameters, i refers to the number of tokens in a given sequence of N tokens, P is the probability of predicting the token i as a function of the sequence x of previous tokens.

When fine-tuning the model embeddings, specialized loss functions can be used to fine-tune the vector representations of the words. Some popular options are:

- Cosine similarity loss: adjusts embeddings so that similar words have more similar vectors.
- Mean square error loss: minimizes the quadratic difference between predicted and expected embeddings.
- Multiple Negative Ranking Loss: associate embeddings of related words so that they are closer together than those of unrelated words.
- Triplet, Matryoshka or contrastive loss: more advanced variants that consider relationships between trios or groups of embeddings.

Careful selection of the loss function is crucial for training effective and efficient LLMs that can capture the nuances of natural language.

Figure 7. Operation of the RAG.



- **Catastrophic forgetting**[66]: During fine-tuning it is possible for a model to forget previously learned critical knowledge.

- **Instability**[67]: The fine-tuning process can be sensitive to factors such as weight initialization, hyperparameters and data selection, which can lead to inconsistent results or variations in performance.

- **Bias inheritance**[68]: Models that have been fine-tuned may inherit and amplify biases present in both pre-training and fine-tuning data, which requires careful consideration and mitigation.

There are several types of fine-tuning to choose from, depending on how much the initial model needs to be modified to fit a task in a more specific domain. The main methods are:

- **Supervised fine-tuning**[69]: This method require labeled input and response data sets from the LLM that are used to improve its response to specific tasks. A popular method of supervised fine-tuning is called “instruction-tuning”[70], which consists of tuning the model's responses to what is expected by its users through interactions with the model.

- **Reinforcement learning:** These methods are based on reinforcement learning and focus on improving the quality of the LLM response, in this case based on user feedback or reward models (e.g., direct optimization by preference[71]).

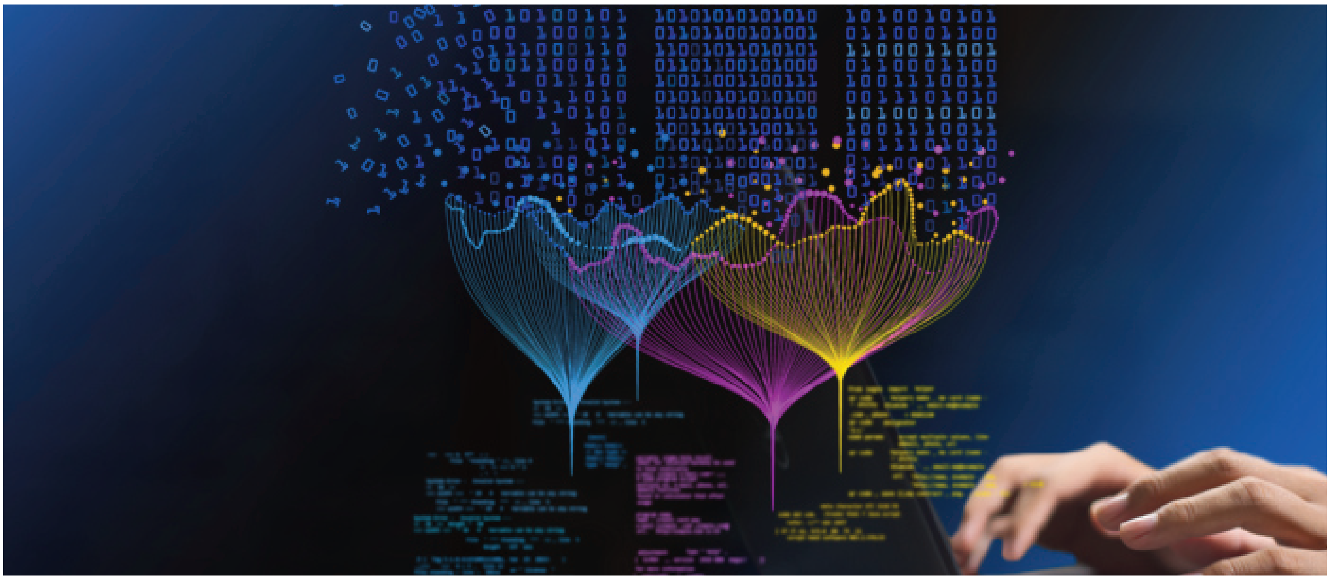
- **Unsupervised fine-tuning**[72]: This is a method that does not require labeled data sets, but relies on retraining the model with the same methods used during pre-training (e.g., predicting the next token).

- **Parameter efficient**[73]: Fine-tuning (PEFT): Other fine-tuning methods aim to increase efficiency and reduce the effort required to retrain the model. For example, techniques based on LoRA[74] (low-rank adaptation), such as QLoRA or LongLoRA[75], allow fine-tuning of the model without changing its weights and store the knowledge learned during the fine-tuning process in additional model parameters.

In many LLM use cases, it is not necessary to use fine-tuning to improve the model´s capabilities in a specific domain. Augmented Retrieval Generation[76](RAG) is a technique that improves LLM performance by using knowledge sources external to the model.

RAG techniques (Figure 7) work by searching a database for documents similar to or related to the input prompt. This search and its results are added to the LLM response generation to enrich it by providing a specific context.

[66]Luo (2024).
[67]Zhang (2024).
[68]Zhang (2024).
[69]Ovadia (2024).
[70]Zhang (2023).
[71]Rafailov (2023).
[72]Zhou (2023).
[73]Xu (2023).
[74]Dettmers (2023).
[75]Chen (2023).
[76]Lewis (2020) and Neelakantan (2022).



Deployment and use

Once trained and validated, the LLM needs to be deployed in a production environment for use in real applications. This involves integrating the model into existing systems and workflows, and creating interfaces and APIs to interact with it.

There are several key aspects to this process, including integration and monitoring.

Integration with systems and workflows

- **Infrastructure**[77]: LLMs are typically large and computationally intensive models that require a robust infrastructure for their implementation. This may include the use of specialized hardware, such as GPUs or TPUs, and cloud computing platforms optimized to perform the inference process efficiently.

- **Interfaces and APIs**[78]: To facilitate the use of the LLM in applications and services, it is necessary to develop interfaces and APIs that allow other systems to interact with the model in an efficient and secure manner. This may include endpoints, client libraries in various programming languages and graphical user interfaces for non-technical users.

- **Integration with other components:** In many cases, LLMs are part of a larger system that includes other components such as databases, natural language processing services and end-user applications. Seamless and efficient integration of the LLM with these components is critical to ensure optimal performance and user experience.

Monitoring and maintenance

- **Performance monitoring**[79]: Once implemented, it is essential to closely monitor LLM performance under real-world conditions. This involves tracking metrics such as latency, throughput, accuracy and resource usage, as well as setting thresholds for resource consumption and cost, and alerts to detect and address any degradation or anomalies.

- **Updating and retraining**[80]: As new data becomes available or areas for improvement are identified, it may be necessary to update or retrain the LLM. This requires a well-defined process to collect and prepare new data, perform fine-tuning, and deploy the updated version of the model without service interruptions.

- **Version management**[81]: With continuous upgrades and enhancements, it is important to maintain strict version control of the LLM and its associated components. This facilitates reproducibility, debugging and the ability to revert to previous versions if necessary.

As can be seen, LLM development and deployment is a complex and multifaceted process that requires careful consideration of multiple aspects, from data selection and preparation to implementation and responsible use of the model. A thorough understanding of the key components, such as pre-training, fine-tuning and embedding, as well as an awareness of the associated challenges and risks, is essential to harnessing the full potential of LLMs in an ethical, sustainable and cost-effective manner that is aligned with each organization's objectives.

[77] Wan (2024).
[78] Abhyankar (2024).
[79] Goyal (2024).
[80] Lester (2021).
[81] Banerjee (2023).

LLM architecture

LLM architecture refers to the structure and organization of the neural networks that make up these models. The choice of architecture and its components significantly impacts the LLM's performance, efficiency and capabilities. This section examines the major architectures used in LLMs and their characteristics, advantages, and limitations.

Transformers: the state of the art in LLMs

Introduced in 2017, transformers have become the dominant architecture for LLMs[82]. Unlike previous architectures based on recurrent neural networks (RNNs) or convolutional neural networks (CNNs), transformers rely solely on attentional mechanisms to process and generate text sequences (Figure 8).

The transformer architecture consists of two main components: the encoder and the decoder, and there are transformers with encoder only, decoder only, or both components. The encoder processes the input sequence and generates a contextual representation for each token, while the decoder generates the output sequence from the encoder representation and previous predictions.

The key to transformers is the attention mechanism, which allows the model to pay attention to different parts of the input sequence (encoder attention) and to previous predictions (decoder attention) to generate the next word or token. This allows long-term dependencies to be captured and coherent sequences to be generated.

Transformers also introduce the concept of multi-head attention, where multiple attention mechanisms operate in parallel, allowing the model to capture different types of relationships and patterns in the data.

The Transformer architecture has demonstrated exceptional performance on a wide range of natural language processing tasks, and has been adopted by most state-of-the-art LLMs.

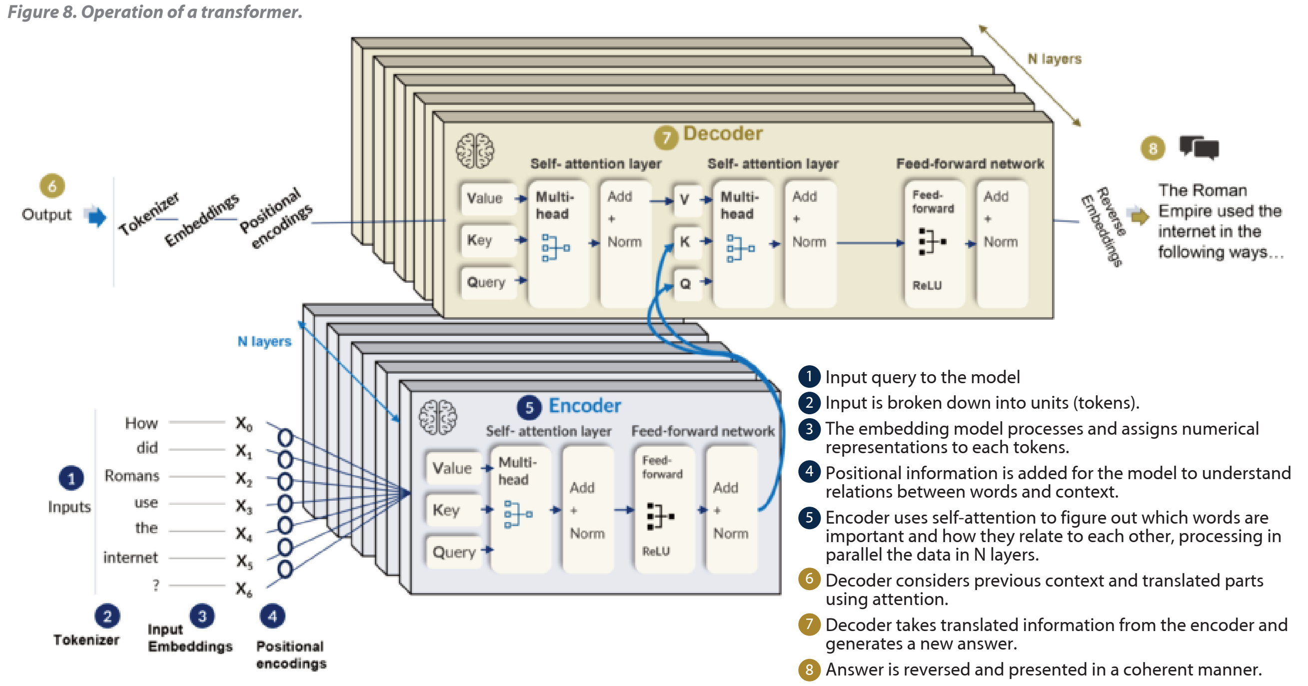
Transformers variants and extensions

Since the introduction of transformers, numerous variants and extensions have been proposed to improve their efficiency, scalability and modeling capabilities.

- One popular variant is the bidirectional transformer, which allows the model to consider each token's left and right context. This is achieved by using a masked language modeling (MLM) pre-training goal, where some tokens are randomly masked and the model must predict them based on the surrounding context.
- Another variant is the Generative Transformer, such as GPT, which uses a one-way language modeling approach. This allows efficient and consistent text generation because the model can only consider the left context of each token.
- Extensions have also been proposed to make transformers more efficient and scalable, such as the sparse transformer, which uses sparse attention to reduce computational complexity, and the compressed transformer, which uses compression techniques to reduce model size.

[82] Vaswani (2017).

Figure 8. Operation of a transformer.



Prompt Engineering in LLMs: Principles and Best Practices

Prompt engineering refers to the process of designing and optimizing prompts to get the best possible results from LLMs. This emerging discipline includes a set of principles and best practices that allow you to take full advantage of the capabilities of these models. Among them are:

- Be clear and specific: The instructions given to the model should explicitly state the format, length, and level of detail expected in the response. For example, instead of simply asking "Analyze the financial situation of company X," it is better to give an instruction such as "Write a 1000-word report on the financial situation of company X, covering its profitability, liquidity, solvency, and future prospects".

- Break down complex tasks: It is useful to break down problems into more manageable subtasks for LLMs. For example, instead of asking "Develop a strategic plan for company Y", subtasks such as "Conduct a SWOT analysis of company Y", "Define the key strategic objectives for company Y", "Propose initiatives to achieve each objective", etc. can be requested.

- Provide illustrative examples (few-shot learning): A few well-chosen examples can go a long way in communicating the desired task. For example, if you want to create value propositions for products, you could give two examples: "Our CRM software enables sales teams to close deals 50% faster" and "Our wellness app helps employees reduce stress and increase their productivity by 25%".

- Ask for step-by-step reasoning: Asking the LLM to verbalize its thought process often leads to more robust results. This is especially useful for business analysis or problem-solving tasks. For example, "Describe step-by-step how you would calculate the ROI of this investment project."

- Ask for references used: Instruct the LLM to provide references to the documents used in its argument, including citations to the original text to which it has access.

- Ask the LLM to adopt a persona: Before the main task, you can first instruct the model to adopt a certain role, tone, or style. For example: "Act as an expert financial analyst and provide an objective assessment of company X". This will help guide its behavior.

- Leverage external knowledge: By providing additional information, the LLM's knowledge base can be supplemented. For example, to answer questions about a specific industry, one could first retrieve relevant industry reports and feed them into the model.

- Iterate and refine systematically: By continuously evaluating model performance, areas for improvement can be identified and prompts adjusted accordingly. Quantitative metrics and qualitative judgments from domain experts can guide this iterative process.

By applying these prompt engineering principles, LLMs are statistically proven to deliver a more accurate and reliable result.

All things considered, a bad prompt for an LLM to write a column on prompt engineering would be, "Write an article on prompt engineering."

And a good prompt for that column would be:

> "Act as an artificial intelligence expert and write a 600-word outreach column on the key principles of prompt engineering to get the best results from LLMs. Structure the column with a brief and engaging introduction, 4-5 paragraphs covering the main points (be specific, break down tasks, give examples...), and a conclusion with the benefits of applying these techniques. Use an informative but rigorous tone, suitable for a business audience. Include concrete examples to illustrate the ideas".

Sources: OpenAI prompt engineering guide[1], Anthropic Claude Opus support and own elaboration.

[1] OpenAI (2024).

Comparison to previous architectures

Before transformers, the dominant architectures for sequence modeling were recurrent neural networks (RNN), such as long short-term memory (LSTM) and gated recurrent unit (GRU), and convolutional neural networks (CNN).

- RNNs can capture long-term dependencies in sequences, but suffer from problems such as gradient vanishing and difficulty in parallelizing training. In addition, RNNs have difficulty capturing very long dependencies due to their sequential nature and the use of constant range recurrence.
- CNNs can capture local patterns in sequences and are computationally efficient, but have difficulty modeling long-term dependencies and require a fixed context size.

In contrast, transformers overcome these limitations by using attention mechanisms that can efficiently capture long-term dependencies in parallel. In addition, transformers are more flexible in handling variable-length sequences and can be pre-trained on large amounts of unlabeled data.

The transformer architecture has revolutionized the field of LLM and has enabled significant advances in a wide range of natural language processing tasks. However, challenges such as the scalability, interpretability, and efficiency of these models remain. As research continues, new architectures and techniques are likely to emerge that will overcome these limitations and take LLMs to new heights of performance and capability.

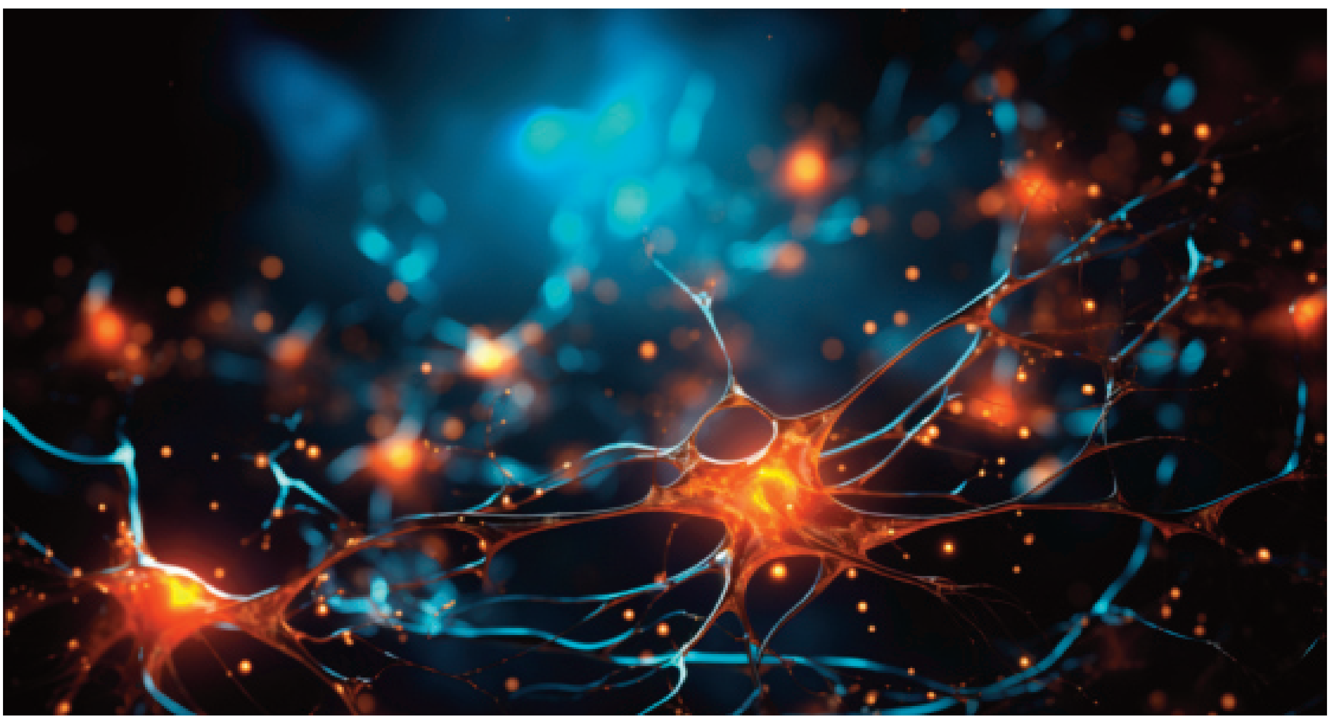
LLMOps

Machine Learning Operations (MLOps) is a methodology and set of practices designed to manage the complete lifecycle of machine learning models, from development and training to deployment and maintenance in production.

In recent years, an adaptation of the MLOps methodology specifically for LLMs has emerged, known as LLMOps (Large Language Model Operations). This discipline focuses on efficiently managing the entire LLM lifecycle, from development and training to deployment and maintenance in production environments.

LLMOps integrates traditional software development processes with tools and techniques designed to address the unique challenges of large language models. These challenges include:

- **Managing large amounts of data:** LLMs require massive amounts of training data, which implies the need for scalable and efficient storage and processing infrastructures.
- **Scaling of computational resources:** LLM training and inference require massive computational resources, which calls for the use of parallelization and distribution techniques, as well as optimizing the use of specialized hardware such as GPUs and TPUs.
- **Monitoring and maintenance:** Once deployed in production, LLMs must be closely monitored to detect and correct performance issues, biases, risks such as hallucinations, and model degradation over time.



- **Versioning and reproducibility:** Given the size and complexity of LLMs, it is critical to maintain strict version control and maximize the reproducibility of experiments and results.

To address these challenges, LLMOps relies on a number of specific tools and frameworks, such as MLFlow[83], CometML[84] and Weights & Biases[85]. These platforms provide capabilities for experiment tracking, model management, performance monitoring, and cross-team collaboration.

In addition, LLMOps promotes practices such as process automation, continuous testing, comprehensive documentation and model governance. This not only improves the efficiency and quality of LLM development, but also ensures its ethical and responsible use.

Challenges

The development and deployment of LLMs presents a number of significant challenges that must be addressed to ensure their responsible, ethical, and secure use. This section explores some of the key challenges that organizations face in deploying and using LLM.

Biases, hallucinations and reliability

One of the biggest challenges of LLMs is the presence of biases and hallucinations in their results and predictions. Biases can arise from several sources, such as biased training data, limitations of model architectures, or human biases implicit in annotation and evaluation tasks. On the other hand, hallucinations refer to the generation of information or content that appears plausible but is not based on facts or knowledge acquired during training.

Biases in LLMs can manifest themselves in a variety of ways, such as perpetuating gender, race, or age stereotypes, discriminating in classification tasks, or generating offensive or inappropriate content. These biases can have serious consequences, especially when LLMs are used in sensitive legal, financial or medical applications. In turn, hallucinations can lead to the dissemination of incorrect or misleading information, which can have a negative impact on user confidence and the credibility of LLM-based applications.

To address the challenge of bias, it is necessary to develop robust techniques to detect, measure, and mitigate its presence in LLMs. This includes the creation of bias-specific evaluation datasets, the use of fairness metrics, and the application of bias elimination (debiasing) techniques in both pre-training and fine-tuning. In addition, it is critical to establish ongoing auditing and monitoring processes to ensure that LLMs remain unbiased over time.

To address hallucinations in LLMs, several methods are being developed that focus on improving training data, applying robust regularization techniques, and using human feedback to tune model responses. In addition, architectural changes to the models are being investigated to make them inherently less prone to hallucination. Text generation methods and input context can also be optimized to reduce hallucinations. Human supervision and rigorous evaluation are essential to detect and correct inaccurate information. Also, the development of specific tools, such as hallucination assessment models and obfuscation techniques, can help improve the accuracy of LLMs.

[83] Zaharia (2018).
[84] CommetML: https://www.comet.com/
[85] Weights and biases: https://wandb.ai/site

Explainability and accountability

Another major challenge with LLMs is their opacity and lack of explainability. Due to their complexity and the nature of their architectures, it is difficult to understand how these models arrive at their results.

This lack of transparency raises accountability issues, especially when LLMs are used in highly sensitive contexts where decisions significantly impact individuals (e.g., the use of LLMs in medicine, pharmaceutical research, critical infrastructure, or access to the labor market). Without a clear understanding of how these models work, it is difficult to determine liability in the event of errors or unintended behavior.

To address this challenge, it is necessary to develop techniques and tools that allow for greater interpretability and explainability of LLMs. This includes methods for visualizing and analyzing internal attention mechanisms, attribution techniques for identifying the most relevant parts of the input, and approaches for generating natural language explanations of model predictions.

In addition, it is important to establish clear accountability frameworks that define the responsibilities of LLM developers, implementers and users, as proposed in Europe by the AI Act. This may involve the creation of standards and guidelines for the ethical development of LLMs, external monitoring and auditing mechanisms, and channels for stakeholders to raise concerns.

Confidentiality and information protection

LLMs are often trained with large amounts of data that may contain personal, sensitive or confidential information. In addition, when used in real-world applications, these models may be exposed to user input, which may include private data. This poses significant privacy and security challenges, as LLMs may memorize and reproduce sensitive information from their training data, or be vulnerable to attacks that attempt to extract private data through carefully crafted queries.

To address this challenge, it is necessary to develop privacy preserving techniques in LLM training and deployment (e.g., Digger[86] to detect protected information, the use of dummy data[87] during training to detect copyrighted material).

In addition, it is crucial to establish robust security and access control protocols to protect LLMs and their associated data from unauthorized access or malicious use. This may involve the use using authentication and authorization techniques, security monitoring and anomaly detection.

Rational use of resources

LLM training and deployment requires massive amounts of computational resources, storage and power. With models reaching hundreds of billions or even trillions of parameters, the financial and environmental cost of developing and operating these systems can be very significant[88].

This high resource consumption poses efficiency, scalability and sustainability challenges. As the demand for larger and more powerful LLMs continues to grow, ways must be found to optimize their performance and reduce their resource footprint.

To address this challenge, several research directions are being explored. One is the design of more efficient model architectures, such as using sparse attention mechanisms or compression techniques that reduce the size and computational complexity of LLMs without significantly compromising their performance.

[86] Li (2024).
[87] Meeus (2024).
[88] iDanae 1T24 (2024).

Research is also underway to improve continuous pre-training techniques[89] and continuous fine-tuning[90], which seek to integrate the ability to use information from diverse domains without relying on extensive and costly retraining with specific new data. This aims to integrate the ability to use information from different domains without relying on extensive and costly retraining with specific new data. Progress is also being made in using innovative systems and designing green AI algorithms that address the computational and environmental costs associated with AI (e.g., Qsimov Quantum Computing's GreenLightningAI system[91] develops incremental retraining and provides straightforward interpretability).

Another direction is the development of more sustainable computing infrastructures and platforms, such as using specialized low-power hardware, more efficient cooling systems and renewable energy sources to power the data centers where LLMs are trained and deployed.

In addition, it is important to promote practices of rational and shared use of resources, such as reusing and adapting pre-trained models instead of training new models from scratch for each task, and the sharing of resources and knowledge between organizations and research communities.

Other challenges

Among the many additional challenges that organizations face in developing, implementing, and using LLMs, the following are worthy of brief mention because of their importance:

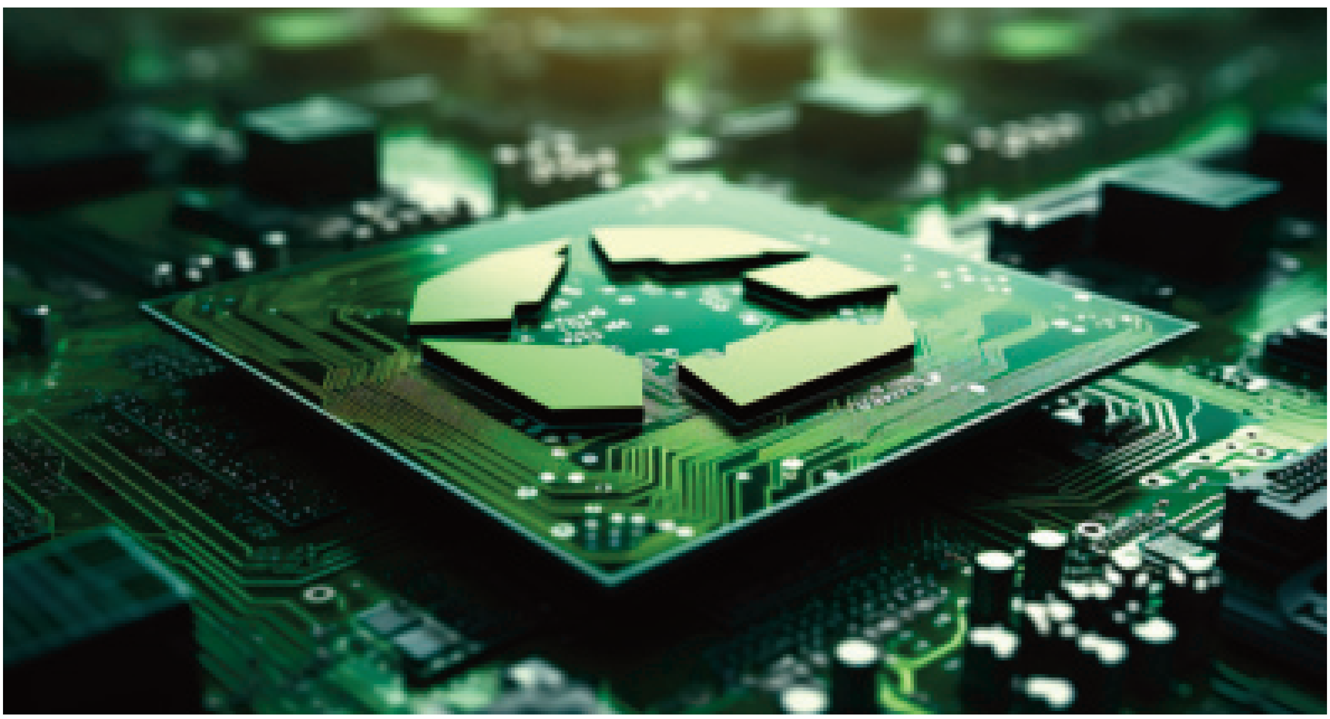
- **Dependency and lock-in:** Organizations that rely on LLMs provided by third parties may face dependency and lock-in risks, especially if the models are based on proprietary data or infrastructure. It is important to consider diversification strategies and contingency plans.

- **Security risks and malicious use**[92]: LLMs can be vulnerable to adversarial attacks, such as poisoned data injection or reverse engineering. They can also be used maliciously to generate misinformation, spam, or misleading content. It is essential to implement robust security measures and design models with safeguards against misuse.

- **Intellectual property and licensing issues:** The use of LLM raises questions about intellectual property and licensing of training data, models and generated results. Additionally, there is a risk of theft of information or personal data from users launching queries to LLM deployed in third-party clouds. Regulatory compliance and ethical frameworks are necessary to balance the rights of creators, users and the public interest, and, for organizations, to avoid legal and compliance risks.

- **Scalability of LLM architecture**[93]: An additional challenge is the scalability of transformers as the size of sequences and models increases. Attention mechanisms have quadratic complexity concerning sequence length, which limits their applicability to very long sequences.

[89] Yıldız (2024).
[90] Mehta (2023).
[91] iDanae 1T24 (2024).
[92] Pankajakshan (2024).
[93] Rae (2021).



LLM: validation framework

"The consequences of AI going wrong are serious, so we need to be proactive rather than reactive".
Elon Musk[94]

Framework

Large Language Models (LLMs) have great potential to transform various industries and applications, but they also pose significant risks that must be addressed. These risks include the generation of misinformation or hallucinations, perpetuation of biases, difficulty in forgetting learned information, ethical and fairness concerns, privacy issues due to misuse, difficulty in interpreting results, and the potential creation of malicious content, among others.

Given the potential impact of these risks, LLMs must be thoroughly validated before deployment in production environments. Validation of LLMs is not only a best practice, but also a regulatory requirement in many jurisdictions. In Europe, the proposed AI Act requires risk assessment and mitigation of AI systems[95]. At the same time, in the United States, the NIST AI Risk Management Framework[96] and the AI Bill of Rights highlight the importance of understanding and addressing the risks inherent in these systems.

Validation of LLMs can be based on the principles established in the discipline of model risk, which focuses[97] on assessing and mitigating the risks arising from errors, poor implementation or misuse of models. However, in the case of AI, and particularly LLMs, a broader perspective needs to be taken that encompasses the other risks involved. A comprehensive approach to validation is essential to ensure the safe and responsible use of LLMs.

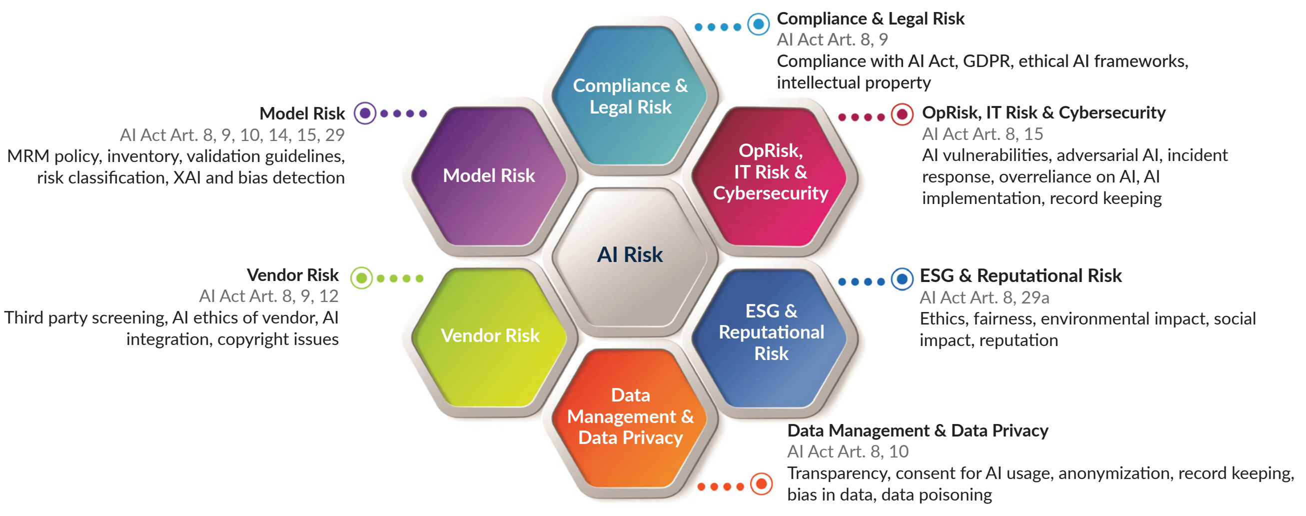
This holistic approach is embodied in a multidimensional validation framework for LLMs that covers key aspects (Figure 9) such as model risk, data and privacy management, cybersecurity, legal and compliance risks, operational and technology risks, ethics and reputation, and vendor risk, among others. By systematically addressing all of these issues, organizations can proactively identify and mitigate the risks associated with LLMs and lay the foundation for unlocking their potential in a safe and responsible manner.

In LLMs, this risk assessment can be anchored in the following dimensions used in the model risk discipline, adapting the tests according to the nature and use of the LLM:

- **Input data:** text comprehension[98], data quality[99].
- **Conceptual soundness and model design:** selection of the model and its components (e.g., fine-tuning methodologies, database connections, RAG[100]), and comparison with other models[101].

[94]Elon Musk (n. 1971), CEO of X, SpaceX, Tesla. South African-American entrepreneur, known for founding or co-founding companies such as Tesla, SpaceX and PayPal, owner of X (formerly Twitter), a social network that has its own LLM, called Grok.
[95]European Parliament (2024) AI Act Art. 9: "A risk management system shall be established, implemented, documented and maintained in relation to high-risk AI systems. The risk management system [...] shall [...] comprise [...] the estimation and evaluation of risks that may arise when the high-risk AI system is used in accordance with its intended purpose, and under reasonably foreseeable conditions of misuse".
[96]NIST (2023): "The decision to commission or deploy an AI system should be based on a contextual assessment of reliability characteristics and relative risks, impacts, costs, and benefits, and should be informed by a broad set of stakeholders".
[97]Management Solutions (2014). Model Risk Management: Quantitative and Qualitative Aspects.
[98]Imperial et al. (2023).
[99]Wettig et al (2024).
[100]RAG (Retrieval-Augmented Generation) is an advanced technique in which a language model searches for relevant information from an external source before generating text. This enriches answers with accurate and current knowledge by intelligently combining information search and text generation. By integrating data from external sources, RAG models, such as the RAG-Token and RAG-Sequence models proposed by Lewis et al. (2020), provide more informed and consistent responses, minimizing the risk of generating inaccurate content or 'hallucinations'. This advance represents a significant step towards more reliable and evidence-based artificial intelligence models.
[101]Khang (2024).

Figure 9. AI Risks and Regulatory References in the AI Act.



- **Model evaluation and analysis of results:** privacy and security of the results[102], model accuracy[103], consistency[104], robustness[105], adaptability[106], interpretability (XAI)[107], ethics, bias and fairness[108], toxicity[109], comparison against challenger models.

- **Implementation and use:** human review in use (including monitoring for misuse), error resolution, scalability and efficiency, user acceptance.

- **Governance**[110] **and ethics**[111]: governance framework for generative AI, including LLMs.

- **Documentation**[112]: completeness of the model documentation.

- **Regulatory compliance**[113]: assessment of regulatory requirements (e.g., AI Act).

To ensure the effective and safe use of language models, it is essential to perform a risk assessment that considers both the model itself and its specific use. This will ensure that the model, regardless of its origin (in-house or from a vendor) or customization (fine-tuning), will function properly in its context of use and meet the necessary security, ethical, and regulatory standards.

Validation techniques

When an organization is considering implementing an LLM for a specific use case, it may be beneficial to take a holistic approach that encompasses the key dimensions of the model's lifecycle: data, design, assessment, implementation and use. It is also necessary to assess compliance with applicable regulations, such as the AI Act in the European Union, in a cross-cutting manner.

In each of these dimensions, two sets of complementary techniques allow for a more complete validation (Figure 10):

- Quantitative evaluation metrics (tests): These standardized quantitative tests measure the model's performance on specific tasks. They are predefined benchmarks and metrics for evaluating various LLM performance aspects after pre-training or during the fine-tuning or instruction tuning (i.e., reinforcement learning techniques), optimization, prompt engineering, or information retrieval and generation phases. Examples include summarization accuracy, robustness to adversarial attacks, or consistency of responses to similar prompts.

- Human evaluation: involves qualitative judgment by experts and end users, such as a human review of a specific sample of LLM prompts and responses to identify errors.

The validation of a specific use of an LLM is therefore carried out by a combination of quantitative (tests) and qualitative (human evaluation) techniques. For each specific use case, it is necessary to design a tailor-made validation approach consisting of a selection of some of these techniques.

[102]Nasr (2023).
[103]Liang (2023).
[104]Elazar (2021).
[105]Liu (2023).
[106]Dun (2024).
[107]Singh (2024).d
[108]NIST (2023), Oneto (2020), Zhou (2021).
[109]Shaikh (2023).
[110]Management Solutions (2014). Model Risk Management.
[111]Oneto (2020).
[112]NIST (2023).
[113]European Parliament (2024). AI Act.

Figure 10. LLM evaluation tests.

Dimensions	Validated aspects	Description	Validation metrics (examples)	Human evaluation (examples)
1. Input data	**1.1 Data quality**	Degree of quality of modeling or application data.	• Flesch-Kinkaid Grade	• Case-by-case review
2. Model design	**2.1 Model design**	Choice of appropriate models and methodology	• Review of LLM elements: RAG, input or output filters, prompts definition, finetuning, optimization... • Comparison with other LLMs	• A/B Testing
3. Model evaluation	**3.1 Privacy and security**	Respect confidentiality and do not regurgitate personal information.	• Data leakage • PII tests, K-anonymity	• Registrations • Ethical hacking
	3.2 Accuracy	Correctness and relevance of model responses	• Q&A: SummaQA, Word error rate • Information retrieval: SSA, nDCG • Summary: ROUGE • Translation: BLEU, Ruby, ROUGE-L • Others: QA systems, level of overrides, level of hallucinations... • Benchmarks: XSUM, LogiQA, WikiData...	• Backtesting of overrides • Case-by-case review
	3.3 Consistency	Correctness and relevance of model responses	• Cosine similarity • Jaccard similarity index	• Case-by-case review • A/B Testing
	3.4 Robustness	Resilience to adverse or misleading informationa	• Adversarial text generation (TextFooler), Regex patterns • Benchmarks of adversarial attacks (PromptBench), number of refusals	• Ethical hacking • Incident drills
	3.5.Adaptability	Ability to learn or adapt to new contexts	• LLM performance on new data by Zero/One/Few-shot learning	• A/B Testing • Case-by-case review
	3.6 Explainability	Understanding the decision making process	• SHAP • Explainability scores	• UX tracking • Focus groups
	3.7 Biases and fairness	Responses without demographic bias	• AI Fairness 360 toolkit • WEAT score, demographic parity, word associations... • Benchmarks of biases (BBQ...)	• Ethical hacking • Focus groups
	3.8 Toxicity	Propensity to generate harmful content.	• Perspective API, Hatebase API • Toxicity benchmarks (RealToxicityPrompts, BOLD, etc.)	• Ethical hacking • Focus groups
4.Implementation and use	**4.1 Human review and safety of use**	Avoid harmful or illegal suggestions and include a 'human-in-the-loop' review.	• Risk protocols, safety assessments • Human control	• Ethical hacking • Focus groups
	4.2 Recovery and error handling	Ability to recover from errors and handle unexpected inputs	• System recovery tests • Error processing metrics	• Incident drills
	4.3 Scalability	Maintain performance with more data or users	• Stress testing of the system, Apache Jmeter... • Scalability benchmarks	• Incident drills • A/B Testing
	4.4 Efficiency	Resource utilization and speed of response	• Time-to-first-byte (TTFB), GPU/CPU utilization, broadcast inference, memory, latency	• Incident drills
	4.5 User acceptance	User acceptance testing.	• User requirements checklist, user opt-out • User Satisfaction (Net Promoter Score, CSAT)	• UX tracking • A/B Testing

The exact selection of techniques will depend on the particular characteristics of the use case; and, in particular, several important factors to consider when deciding on the most appropriate techniques are:

- The level of risk and criticality of the tasks to be entrusted to the LLM.
- Whether the LLM is open to the public (n which case ethical hacking becomes particularly relevant) or its use is limited to the internal scope of the organization.
- Whether the LLM processes personal data.
- The line of business or service the LLM will be used for.

Careful analysis of these factors will allow the construction of a robust validation framework tailored to the needs of each LLM application.

Quantitative evaluation metrics

Although this is an emerging field of study, there is a wide range of quantitative metrics that can be used to evaluate LLM performance. Some of these metrics are adaptations of those used in traditional machine learning models, such as accuracy, recall, F1 score, or area under the ROC curve (AUC-ROC). Other metrics are specifically designed to evaluate unique aspects of LLMs, such as the coherence of the generated text, factual fidelity, or language diversity.

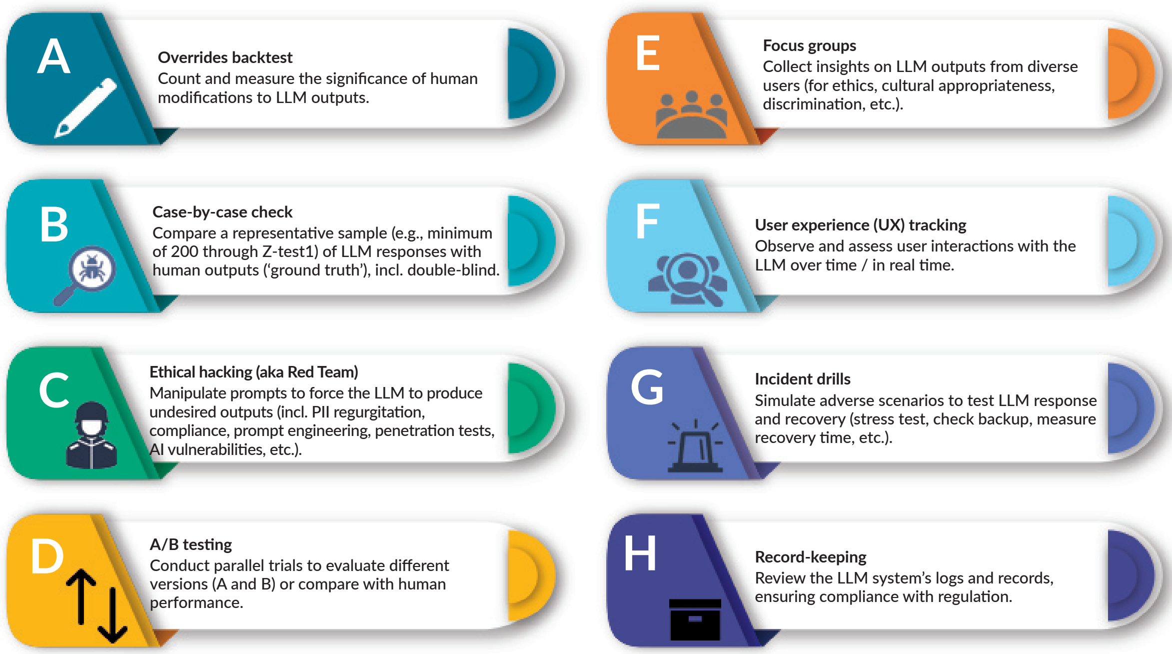
In this context, holistic quantitative LLM testing frameworks already exist in Python programming environments, which facilitate the implementation of many of the quantitative validation metrics, such as:

- **LLM Comparator**[114]: a tool developed by Google researchers for automatically evaluating and comparing LLMs, which checks the quality of LLM answers.
- **HELM**[115]: Holistic Evaluation of Language Models, which compiles evaluation metrics along seven dimensions (accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency) for a set of predefined scenarios.
- **ReLM**[116]: LLM validation and query system using language usage, including evaluation of linguistic models, memorization, bias, toxicity and language comprehension.

At present, certain validation techniques, such as SHAP-based explainability methods (XAI), some metrics such as ROUGE[117] or fairness analyses using demographic parity, do not yet have widely accepted predefined thresholds. In these cases, it is the task of the scientific community and the industry to continue research to establish clear criteria for robust and standardized validation.

[114] Kahng (2024).
[115] Liang (2023).
[116] Kuchnik (2023).
[117] Duan (2023).

Figure 11. Some LLM human evaluation techniques.



Human evaluation techniques

While quantitative assessment metrics are more directly implementable due to the multitude of online resources and publications in recent years, human assessment techniques[118] are varied and must be constructed based on the specific task[119] being performed by the LLM, and include (Figure 11):

- **User override backtesting:** counting and measuring the importance of human modifications to LLM results (e.g., how many times a sales manager must manually modify customer call summaries generated by an LLM).
- **Case-by-case review:** comparing a representative sample of LLM responses to user expectations ("ground truth").
- **Ethical hacking (Red Team):** manipulating prompts to force the LLM to produce undesired results (e.g., regurgitation of personal information, illegal content, penetration testing, vulnerability exploitation).
- **A/B testing:** comparison to evaluate two versions of the LLM (A and B), or an LLM against a human being.
- **Focus groups:** gathering opinions from various users on LLM behavior, e.g., ethics, cultural appropriateness, discrimination, etc.
- **User experience (UX tracking):** observing and evaluating user interactions with the LLM over time or in real time.
- **Incident drills:** simulating adverse scenarios to test LLM response (e.g., stress test, backup check, recovery time measurement, etc.).
- **Record keeping:** reviewing LLM system logs and records to ensure compliance with regulations and the audit trail.

[118] Datta, Dickerson (2023).
[119] Guzmán (2015).

Benchmarks for LLM Evaluation

Most generative artificial intelligence models, including LLMs, are tested against public benchmarks to evaluate their performance on a variety of tasks related to natural language understanding and usage. These tests are used to measure how well the LLM handles specific tasks and mirrors human understanding. Some of these benchmarks include:

- GLUE/SuperGLUE: assesses language comprehension through tasks that measure a model's ability to understand text.
- Eleuther AI Language Model Evaluation Harness: performs "few-shot" model evaluation, that is, evaluates model accuracy with very few training examples.
- ARC (AI2 Reasoning Challenge): tests the model's ability to answer scientific questions that require reasoning.
- HellaSwag: evaluates the model's common sense through tasks that require predicting a coherent story ending.
- MMLU (Massive Multitask Language Understanding): tests the model's accuracy on a variety of tasks to assess its understanding of multitasking.
- TruthfulQA: challenges the model to distinguish between true and false information, assessing its ability to handle truthful data.
- Winogrande: another tool to assess common sense, similar to HEllaSwag, but with different methods and emphasis.
- GSM8K: uses mathematical problems designed for students to assess the model's logical-mathematical capability.

New trends

The field of LLM validation is constantly evolving, driven by rapid advances developing these models and a growing awareness of the importance of ensuring their reliability, fairness and alignment with ethics and regulation.

Below are some of the key emerging trends in this area:

- **Explainability of LLMs:** As LLMs become more complex and opaque, there is a growing need for mechanisms to understand and explain their inner workings. XAI (eXplainable AI) techniques such as SHAP, LIME, or assigning importance to input tokens are gaining importance in LLM validation. Although a variety of post-hoc techniques for understanding the operation of models at the local and global level are available for traditional models[120] (e.g., Anchors, PDP, ICE), and the definition and implementation of inherently interpretable models by construction has proliferated, the implementation of these principles for LLMs is still unresolved.

- **Using LLMs to explain LLMs:** An emerging trend is to use one LLM to generate explanations for the behavior or responses of another LLM. In other words, one language model is used to interpret and communicate the underlying reasoning of another model in a more understandable way. To enrich these explanations, tools are being developed[121] that also incorporate post-hoc analysis techniques.

- **Post-hoc interpretability techniques:** These techniques are based on the interpretability of the results at the post-training or fine-tuning stage, and allow to identify which parts of the input have most influenced the model response (feature importance), to find similar examples in the training data set (similarity based on embeddings) or to design specific prompts that guide the model towards more informative explanations (prompting strategies).

- **Attribution scores:** As part of post-hoc interpretability[122], techniques are being developed to identify which parts of the input text have the greatest influence on the response generated by an LLM. They help to understand which words or phrases are most important for the model. There are different methods for calculating these scores:

  - Gradient-based methods: Analyze how the gradients (a measure of sensitivity) change for each word as it moves back through the neural network.

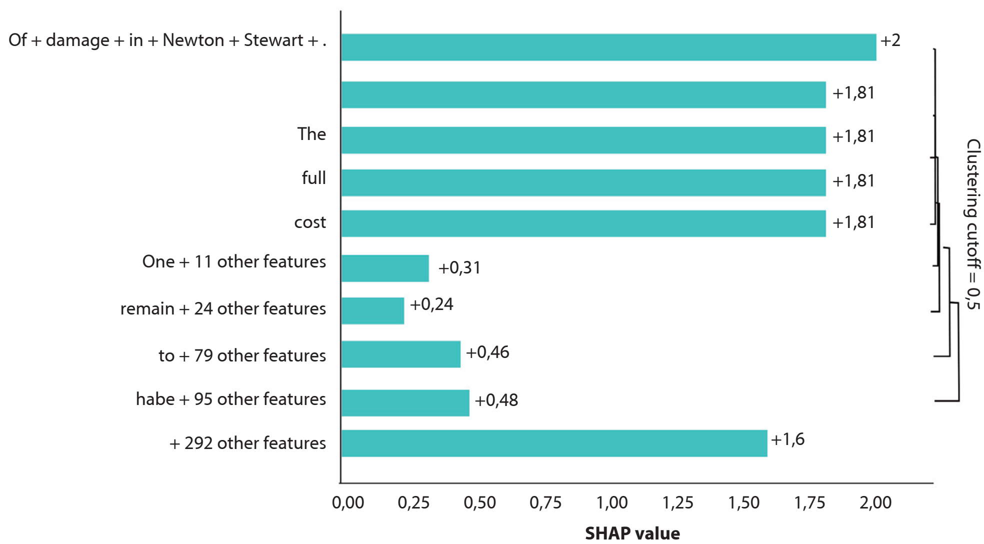
  - Perturbation-based methods: Slightly modify the input text and observe how the model response changes.

  - Interpretation of internal metrics: Use metrics calculated by the model itself, such as attention weights in transformers, to determine the importance of each word.

[120] Management Solutions (2023). Explainable Artificial Intelligence.
[121] Wang (2024).
[122] Sarti (2023).

Figure 12. Implementation of SHAP values for text summarization.

Output summary: "The full cost of damage in Newton Stewart, one of the areas worst affected, is still being assessed . First Minister Nicola Sturgeon visited the area to inspect the damage. Labour Party 's deputy Scottish leader Alex Row ley was in Haw ick on Monday to see the situation first hand. He said it was important to get the flood protection plan right"



An example of attribution scoring is the use of the SHAP technique to provide a quantitative measure of the importance of each word to the LLM output, which facilitates its interpretation and understanding (Figure 12).

- **Continuous validation and monitoring in production:** In addition to pre-deployment evaluation, the practice of continuously monitoring the behavior of LLMs in production, as is done with traditional models, is growing. This makes it possible to detect possible deviations or degradations in their performance over time, and identify biases or risks that were not initially anticipated.

- **Collaborative and participatory validation:** Greater involvement of different stakeholders in the validation process is encouraged, including not only technical experts but also end users, regulators, external auditors and representatives of civil society. This plural participation allows for the inclusion of different perspectives and promotes transparency and accountability.

- **Ethical and regulatory-aligned validation:** In addition to performance metrics, it is becoming increasingly important to assess whether LLM behavior is ethical and in line with human values and regulations. This involves analyzing issues such as fairness, privacy, security, transparency, or the social impact of these systems.

- **Machine unlearning:** This is an emerging technique[123] that allows unlearning "known information from a LLM without retraining it from scratch. This is achieved, for example, by adapting the hyperparameters of the model to the data to be unlearned. The same principle can be used to remove identified biases. The result is a model that retains its general knowledge but has problematic biases removed, improving its fairness and ethical orientation in an efficient and selective way. Several machine unlearning methods are currently being explored, such as gradient ascent[124], the use of fine-tuning[125] or selective modification of certain weights, layers or neurons of the model[126].

[123] Liu (2024).
[124] Jang (2022).
[125] Yu (2023).
[126] Wu (2023)

SHAP (SHapley Additive exPlanations) applied to an LLM

SHAP is a post-hoc explainability method based on cooperative game theory. It assigns each feature (token) an importance value (Shapley value) that represents its contribution to the model prediction.

Formally, let x = (x_1,…,x_n) be a sequence of input tokens. The prediction of the model is denoted by f(x). The Shapley value φ value for the token x_i is defined as:

$$\phi_i = \sum_{\{S \subseteq N_i\}} \frac{\{|S|!\,(n-|S|-1)!\}}{\{n!\}} [f(S \cup \{i\}) - f(S)]$$

where N is the set of all tokens, S is a subset of tokens, and f(S) is the model prediction for subset S.

Intuitively, the Shapley value φi captures the average impact of token xi on the model prediction, considering all possible subsets of tokens.

Example: Consider an LLM trained to classify corporate emails as "important" or "unimportant". Given a vector of input tokens:

x = [The, Q2, financial, report, shows, significant, increase, in, revenue, and, profitability].

The model classifies the mail as "important" with = 0.85.

Using SHAP, the following Shapley values are obtained:

φ_1 = 0.01 (The)

φ_2 = 0.2 (report)

φ_3 = 0.15 (financial)

φ_4 = 0.02 (from)

φ_5 = 0.1 (Q2)

φ_6 = 0.05 (show)

φ_7 = 0.01 (a)

φ_8 = 0.15 (increase)

φ_9 = 0.1 (significant)

φ_{10} = 0.01 (in)

φ_{11} = 0.02 (th)

φ_{12} = 0.12 (income)

φ_{13} = 0.01 (and)

φ_{14} = 0.02 (the)

φ_{15} = 0.08 (profitability)

Interpretation: The tokens "report" (0.2), "financial" (0.15), "increase" (0.15) and "revenue" (0.12) have the highest contribution to the classification of the mail as "important". This suggests that the LLM has learned to associate these terms with the importance of the message in a business context.

Case study: validation of a policy chatbot

"Artificial intelligence will reach human levels by 2029".
Ray Kurzweil[127]

"I think we will have an AI that is smarter than any human being probably by the end of 2025".
Perplexity[128]

To illustrate the application of the LLM validation techniques described above, this section presents a case study of the validation of a company's internal policy chatbot.

Case definition

The company has developed a chatbot based on an open source LLM to answer questions and provide information about its internal policies. The main objective of this chatbot is to facilitate employee access to company policies.

The chatbot has been built using a cloud infrastructure and has been fed with all of the company's policies, which comprise approximately 1,000 pages of documentation. To improve its responsiveness, Retrieval-Augmented Generation (RAG) techniques have been applied, which allow the model to retrieve relevant information from its knowledge base before generating a response. Initially, the possibility of applying fine-tuning to the model was considered, but after initial testing it was concluded that the combination of the base LLM with RAG was sufficient to achieve satisfactory results.

Prior to its final implementation, the company has decided to conduct a thorough validation process to assess the chatbot's accuracy, security and suitability in the specific context of its intended use. This validation process aims to identify potential areas for improvement and to ensure that the chatbot meets the Company's quality standards and expectations.

Validation of the policy chatbot will be conducted using a combination of quantitative metrics and human evaluation techniques, following the multidimensional validation framework described in the previous section. The results of this process will be used to make informed decisions about the implementation of the chatbot and to establish a continuous improvement plan.

Design of the validation approach

In order to comprehensively validate the policy chatbot, a tailored validation approach was designed following the validation framework presented in the previous section, covering the key dimensions of the model lifecycle: data, design, evaluation, implementation, and usage. This approach combines quantitative metrics and human evaluation techniques, with the goal of obtaining a complete picture of the chatbot's performance and suitability in the company's specific context.

The tests and techniques selected for each dimension are summarized below:

Data

- Metrics: The Flesch-Kincaid scale will be used to evaluate the readability and complexity of the policies that feed the chatbot.
- Human evaluation: A representative sample of policies will be reviewed to identify possible inconsistencies, errors or ambiguities.

Model design

- Metrics: Specific elements of the LLM will be modified in the development code (e.g., the RAG technique and its hyperparameters, such as the size or the chunking strategy[129]) that may change its response performance, and the results will be compared against the original model.

[127] Ray Kurzweil (n. 1948). Director of Engineering at Google, computer scientist, inventor and futurist, known for the invention of OCR and for his contributions in AI.

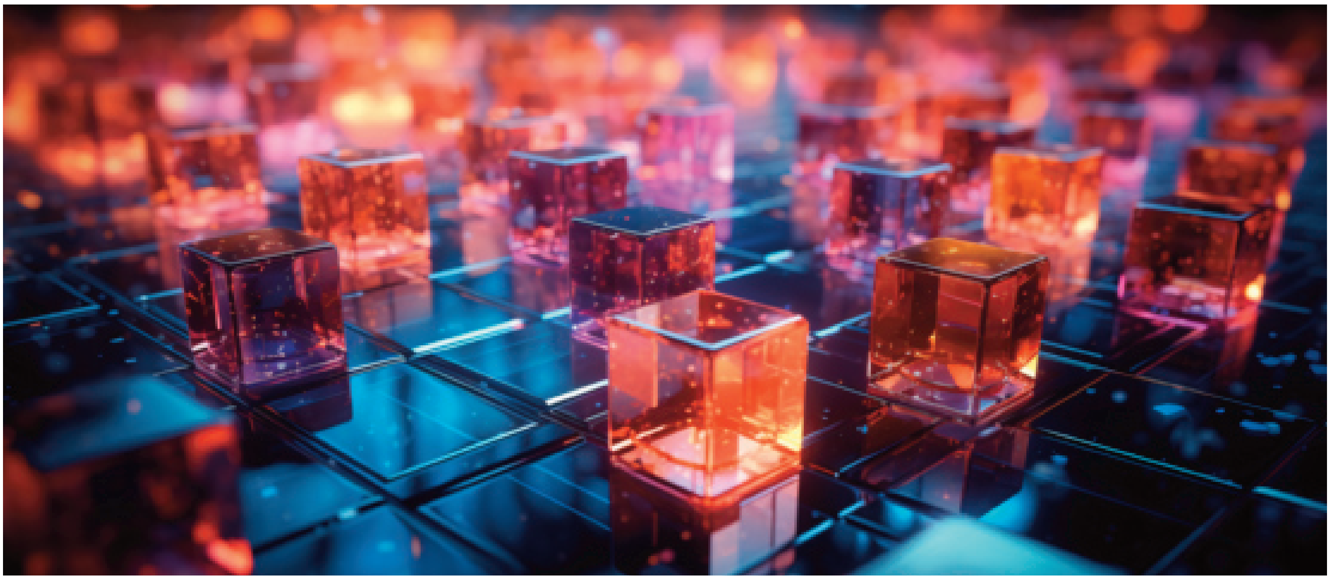
[128] Elon Musk (n. 1971), CEO of X, SpaceX, Tesla. South African-American entrepreneur, known for founding or co-founding companies such as Tesla, SpaceX and PayPal, owner of X (formerly Twitter), a social network that has its own LLM, called Grok.

[129] The chunking strategy refers to the process of dividing the input text to an LLM into smaller, more manageable units ("chunks") during use or implementation.

- Human evaluation: A thorough review of the chatbot components will be performed, including RAG configuration, input and output filters, prompt definition, and hyperparameter optimization. In addition, A/B testing will be conducted to compare the chatbot's performance with other LLMs available in the market.

Evaluation of the model

- Privacy and security
  - Metrics: K-anonymity tests will be applied to evaluate the protection of personal data in chatbot responses, and PII (Personal Identifiable Information) tests will be applied to identify sensitive attributes in the data, using PIIfilter.
  - Human assessment: Ethical hacking tests will be performed to identify potential vulnerabilities and detailed logs of chatbot interactions will be maintained.
- Accuracy
  - Metrics: Word Error Rate (WER) and ROUGE metrics will be used to assess the accuracy of chatbot responses compared to the original policies. Domain-specific benchmarks, such as a set of questions and answers designed by the company's policy experts, will also be used.
  - Human evaluation: A case-by-case review of a representative sample of chatbot interactions will be performed to identify possible errors or inaccuracies.
- Consistency
  - Metrics: Cosine Similarity and Jaccard Index will be used to assess the consistency of chatbot responses to similar queries.
  - Human evaluation: A/B tests will be conducted to compare chatbot responses in different scenarios and a case-by-case review will be performed to identify possible inconsistencies.
- Robustness
  - Metrics: Tools such as TextFooler will be used to generate adversarial text and evaluate the chatbot's resilience to misleading information. In addition, the number of chatbot rejections to malicious prompts will be counted.
  - Human evaluation: Ethical hacking tests and mock incidents will be conducted to evaluate the chatbot's ability to handle adverse situations.
- Adaptability
  - Metrics: The chatbot's performance will be evaluated against new policies or updates using few-shot learning techniques. The chatbot's response to languages not used in the policies or requests for translations into languages not included in the RAG (e.g., Polish) will be evaluated.
  - Human evaluation: A/B testing and case-by-case reviews will be conducted to evaluate the chatbot's ability to adapt to new scenarios.
- Explainability
  - Metrics: Explainability techniques, such as SHAP, will be used to understand the chatbot's decision-making process. The chatbot's intrinsic interpretability module, which provides an explanation of the origin of the information in the response to the user, will be evaluated.
  - Human evaluation: The user experience (UX) will be monitored and a focus group will be conducted to evaluate users' perceptions of the chatbot's transparency and explainability.
- Biases and fairness
  - Metrics: The AI Fairness 360 toolkit will be used to assess potential demographic bias in chatbot responses. Specific benchmarks, such as the Bias Benchmark for QA (BBQ), will also be used to measure fairness in the context of company policies.
  - Human evaluation: Ethical hacking tests and a focus group will be conducted to identify potential bias or discrimination in the chatbot's responses.
- Toxicity
  - Metrics: Perspective API and Hatebase API tools will be used to assess the presence of toxic or inappropriate language in chatbot responses. In addition, specific benchmarks, such as RealToxicityPrompts, will be used to measure toxicity in the context of corporate policy.
  - Human evaluation: Ethical hacking tests will be conducted to identify potential instances of offensive or inappropriate language in chatbot interactions.



Implementation and use

- Scalability
  - Metrics: System stress tests will be performed using Apache JMeter to evaluate the chatbot's performance under heavy workloads.
  - Human evaluation: Simulations will be conducted to evaluate the chatbot's ability to handle an unforeseen increase in the number of users or queries.
- Efficiency
  - Metrics: Response time (Time-to-First-Byte, TTFB), resource usage (GPU/CPU, memory) and latency will be measured to evaluate chatbot efficiency.
- User acceptance
  - Metrics: A checklist of user requirements will be created and user satisfaction will be measured using indicators such as Net Promoter Score (NPS) and Customer Satisfaction Score (CSAT).
  - Human evaluation: User experience (UX) tracking will be conducted to evaluate user acceptance and satisfaction with the chatbot.

This customized validation approach will enable the company to obtain a comprehensive evaluation of the policy chatbot, identify areas for improvement and ensure its suitability for its intended use. The results of these tests and evaluations will be used to make informed decisions about the implementation and the chatbot's ongoing refinement.

Results

After applying the customized validation approach to the policy chatbot, promising results were obtained, demonstrating its overall suitability for the company's intended use (Figure 13). The chatbot achieved satisfactory performance in most evaluated dimensions, meeting quality standards and established expectations.

With respect to the quality of input data, the policies that fed the chatbot were generally found to be of sufficient readability and complexity to be understood by users. In addition, the human review did not identify any significant inconsistencies or errors in the content of the policies.

The model design also proved appropriate for the use case, with optimal configuration of the chatbot components and superior performance compared to other LLMs available on the market.

In terms of model evaluation, the chatbot achieved positive results in most of the metrics and tests applied. The high accuracy of the responses, the consistency in handling similar queries and the ability to adapt to new scenarios stand out. However, some areas for improvement were identified in aspects such as explainability, bias detection, and the response to very specific questions where further model refinement of the model is required. In the area of cybersecurity, a more detailed analysis of the specific vulnerabilities of the open-source LLMs used is required to mitigate this risk in production.

In terms of implementation and use, the chatbot demonstrated good scalability and efficiency in handling high workloads. In addition, user satisfaction was high, indicating a good acceptance of the tool in the company context.

Figure 13. Summary of results of policy chatbot human evaluation metrics and techniques.

Dimension	Test	Result	Interpretation
Datas	**Flesch-Kincaid**	Adequate legibility (grade 8)	The policies are understandable to most users.
	Human Review	No significant inconsistencies	The policies are consistent and free of material misstatement.
Model design	**Challenger models**	Parameter improvements identified	Adapting RAG parameters to the policy context (i.e., chunk size) is required to improve information capture on very specific questions.
	Component overhaul	Optimum configuration	Chatbot design is appropriate for the use case.
	A/B testing	Superior performance compared to other LLMs	Chatbot outperforms other models available on the market
Model Evaluation	**K-anonimato**	Adequate protection of personal data	Chatbot does not reveal sensitive information in its responses.
	Ethical hacking	Identified minor vulnerabilities	Adjustments required to strengthen chatbot security
	Word Error Rate (WER)	WER < 5%	Chatbot responses are highly accurate
	ROUGE	ROUGE-L > 0.8	Chatbot responses adequately capture the content of the policies
	Cosine similarity / Jaccard index	Similarity > 0.9	Chatbot provides consistent responses to similar queries
	TextFooler	Resiliencia moderada ante texto adversario	Chatbot is moderately robust to misleading information
	Few-shot learning	Satisfactory adaptability	Chatbot can adapt to new policies or updates with minimal training, but it is required to monitor and add those new documents to the RAG periodically.
	SHAP	Satisfactory adaptability	Improvements are required in the chatbot's ability to explain its decisions , although the RAG component has been built in such a way that the LLM gives a self-explanatory answer.
	AI Fairness 360 / BBQ	Identified minor demographic biases	The chatbot presents some biases that need to be mitigated
	Perspective API / RealToxicityPrompts	Low toxicity (< 5%)	Chatbot responses rarely contain toxic or inappropriate language
Implementation and use	**Apache JMeter**	Satisfactory scalability (up to 1000 users)	Chatbot can handle high workloads without significant performance degradation
	TTFB / / Resource usag / Latencia	Adequate efficiency (TTFB < 1s, moderate use)	Chatbot responds quickly and uses resources efficiently
	NPS / CSAT	High satisfaction (NPS > 60, CSAT > 80%)	Users are highly satisfied with the chatbot and would recommend it to others

These results indicate that the policy chatbot is well on its way to being implemented in the company, although some specific areas were identified that require further improvement. The following section presents the main conclusions and recommendations derived from this validation process.

Main conclusions

The policy chatbot validation process has shown that this LLM-based system can be a valuable tool for facilitating employee access to relevant corporate information. The results of the various tests and evaluations indicate that the chatbot largely meets the quality, security and efficiency requirements set by the organization.

Strengths identified included the accuracy and consistency of the chatbot's responses, its ability to adapt to new scenarios, and its scalability to handle large workloads. In addition, user satisfaction with the tool is high, indicating good acceptance and adoption by employees.

However, the validation process has also revealed some areas for improvement that need to be addressed before the final implementation of the chatbot. In particular, the following recommendations are made:

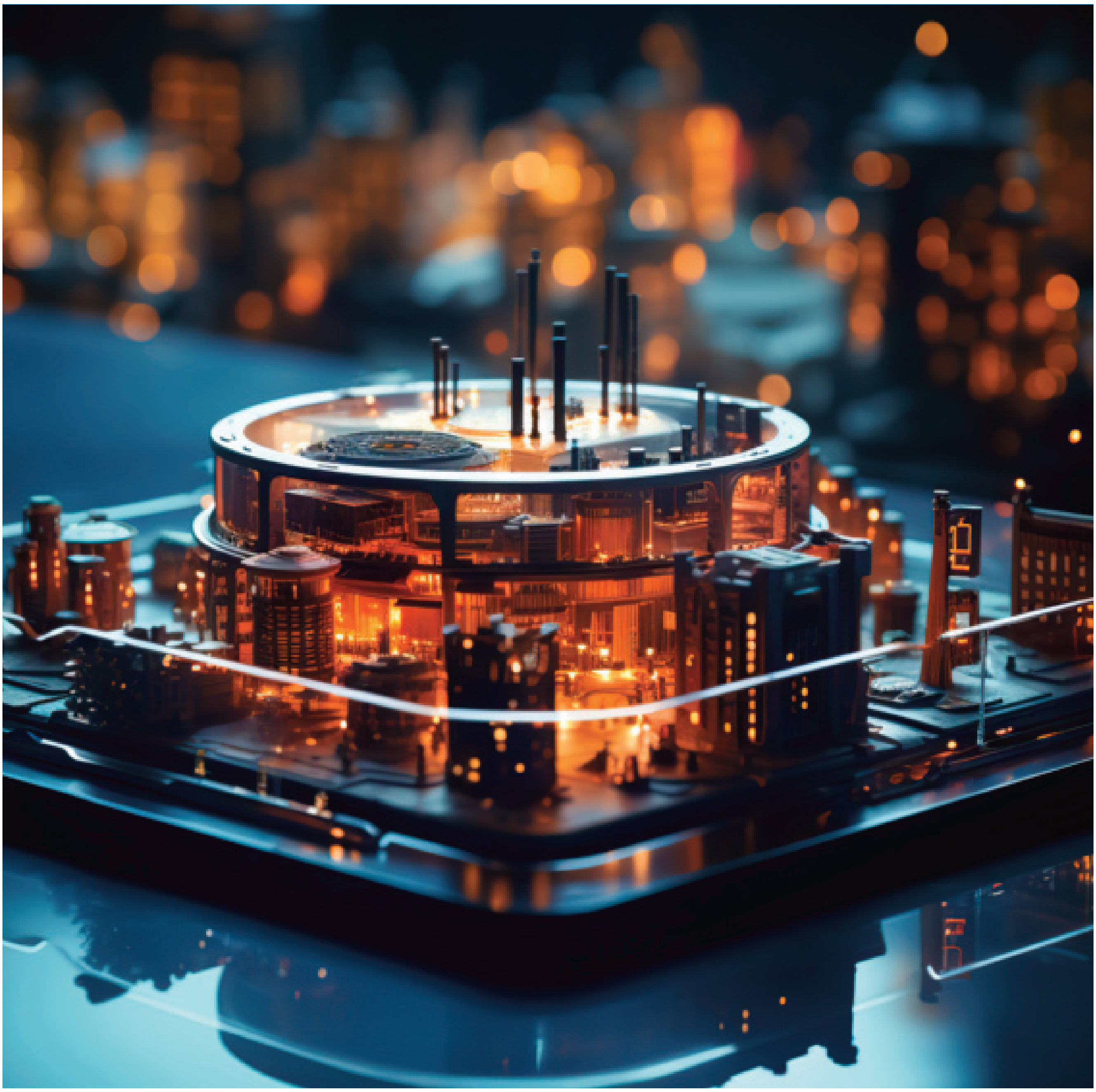
1. **Improve the explainability of the model:** It is necessary to develop more advanced techniques so that the chatbot can provide clear and understandable explanations of its decision-making process. This will increase transparency and user confidence in the tool. While the RAG component has been built in such a way that the LLM gives a self-explanatory answer and refers to the corresponding policy, this explanation is not entirely clear for very specific questions.
2. **Mitigate identified biases:** Although the identified biases are small, it is advisable to apply debiasing techniques to ensure that chatbot responses are fair and non-discriminatory. Periodic review of biases and implementation of corrective measures where necessary is suggested.
3. **Strengthen security and privacy:** While the chatbot meets basic personal data protection standards, additional and recurring ethical hacking tests and more robust security measures are recommended to prevent potential vulnerabilities
4. **Establish a monitoring and continuous improvement plan:** It is essential to define a process for regularly monitoring and evaluating the chatbot's performance in order to identify opportunities for improvement and ensure its optimal performance in the long term. This plan should include collecting feedback from users, regularly updating policies and including them in the chatbot database, monitoring to improve the parameters used in the RAG and updating them, and incorporating new techniques and technologies as they become available.

In conclusion, the policy chatbot has shown potential to improve the efficiency and accessibility of information in the company. With the implementation of the suggested improvements and a focus on continuous improvement, this LLM-based system can become a strategic tool for organizational success. The final recommendation has been to proceed with the implementation of the chatbot, taking into account the observations and recommendations derived from this validation process.

Conclusions

"LLMs are the only people who can write a novel, translate it into ten languages, and still not understand the plot.".

Perplexity[130]



Large Language Models (LLMs) represent a significant advance in the field of artificial intelligence and are revolutionizing the way we interact with technology and leverage natural language processing. Their ability to process and generate coherent, contextualized text opens up a wide range of applications in a variety of industries, from content creation and sentiment analysis to task automation and improved user experience.

However, there are a number of important challenges and considerations in developing and deploying LLM. The presence of biases and hallucinations in their results, the lack of transparency and explainability of their decisions, the challenges of privacy and information security, and the high consumption of computational resources are some of the key challenges that must be addressed to ensure a responsible and ethical use of these systems.

To address these challenges, it is critical to establish a robust AI governance framework, especially in the area of generative AI and LLMs. This framework must encompass all key aspects, including strategy, risk appetite, governance, organization, control framework (policies and procedures), data, systems and reporting. Only a comprehensive and well-structured approach will ensure these technologies' responsible development and usea.

Validation plays a crucial role within this governance framework. Adopting a multidimensional approach that covers all stages of the LLM lifecycle is desirable, from the quality of the input data and robustness of model design to thorough evaluation of results and appropriate implementation and use. This validation process should combine standardized quantitative metrics with human evaluation techniques tailored to the specific context of each use case.

In addition, it is necessary to keep abreast of the latest trends and advances in the field of LLM validation, such as the development of more advanced explainability techniques, the use of LLMs to explain the behavior of other LLMs, continuous validation and monitoring in production, and alignment with ethical principles and regulatory requirements.

The case study presented in this white paper illustrates how the application of a customized validation framework can help organizations identify strengths and areas for improvement in their LLM-based systems, and make informed decisions about their implementation and continuous improvement.

In short, LLMs have great potential to transform the way businesses and society at large benefit from artificial intelligence. However, to realize their full potential in a safe and responsible manner, it is imperative to establish a robust AI governance framework that addresses the challenges associated with their development and deployment, and includes a rigorous, multidimensional approach to validation. This is the only way to ensure that these systems are reliable, fair, and aligned with the values and goals of organizations and society.

[130] Perplexity AI is an LLM-based conversation and research search engine founded in 2022 by Andy Konwinski, Denis Yarats, Johnny Ho, and Aravind Srinivas (formerly OpenAI) that answers queries using natural language predictive text.

Glossary

AGI (Artificial General Intelligence): Hypothetical future artificial intelligence that would equal or surpass human intelligence in any intellectual domain, capable of performing any intellectual task that a human can do.

Hallucinations: The generation of information or content by an LLM that appears plausible but is not based on actual facts or knowledge acquired during training, leading to inaccuracies or inventions in the model's responses.

CNN (Convolutional Neural Network): A type of neural network specialized in processing data with a grid topology, such as images or time series. CNNs use convolution layers to automatically extract local and abstract features from data, and are widely used in computer vision and signal processing tasks.

Quantization: A technique used to reduce the size and speed up the inference of LLMs, which involves reducing the numerical precision of the model weights by moving from floating-point numbers to lower precision representations, such as integers or fixed-point numbers.

Training data: A set of examples used to train a machine learning model, including the inputs (features) and, in the case of supervised learning, the labels or expected responses. The quality and diversity of this data is crucial for model performance and generalization.

Eliza Effect: A psychological phenomenon whereby users tend to attribute human-like cognitive and emotional capabilities to AI-based conversational systems, despite these systems possessing no real understanding of language or general intelligence.

Embeddings: Dense, continuous representations of discrete elements (such as words, phrases or documents) in a high-dimensional vector space, where similar elements have close representations. They are used in LLMs to capture semantic and syntactic relationships between language elements.

AI ethics: The discipline that studies the moral principles, values and guidelines that should guide the development, deployment and use of artificial intelligence systems, with the aim of ensuring that they are beneficial, fair, transparent and aligned with human values.

Human evaluation: The process of qualitative review and assessment of the behavior and results of an AI system by experts and users, which complements quantitative metrics and allows the detection of errors, biases or undesired behaviors that might go unnoticed in a purely automatic evaluation.

Explainability (XAI, eXplainable AI): The property of an AI model that refers to its ability to provide human-understandable explanations of its inner workings, the reasoning behind its predictions, and the factors that influence its decisions.

Few-shot learning: The ability of a machine learning model, especially LLMs, to learn to perform a new task from a few examples (from one to a few tens), leveraging prior knowledge acquired during pre-training on large amounts of data.

Fine-tuning: A technique for adapting a pre-trained language model to a specific task, through additional training with a smaller data set specialized in that task. It allows taking advantage of the general knowledge of the model and adjusting it to obtain high performance in specific applications.

Ethical hacking: The practice of testing and challenging an AI system in a controlled and permissioned manner, with the goal of identifying vulnerabilities, flaws, biases or undesired behaviors, and then correcting them to improve the security and robustness of the system.

Instruction tuning: A fine tuning technique for LLM that consists of providing the model with instructions, questions and examples of expected responses, with the objective of aligning its behavior with the expectations and preferences of users in a specific domain.

Artificial Intelligence (AI): A field of computer science and engineering dedicated to the development of systems capable of performing tasks that normally require human intelligence, such as learning, reasoning, perception, natural language interaction and problem solving.

Generative Artificial Intelligence (GenAI): A subfield of AI that focuses on the creation of models and algorithms capable of generating new and original content, such as text, images, video, audio, source code or 3D designs, by learning patterns and features from a training data set.

Large Language Models (LLM): Deep learning models specialized in natural language processing and generation, trained on huge amounts of text and with a large number of parameters (from millions to billions), capable of performing various linguistic tasks with a high level of comprehension and coherence.

LLMOps (Large Language Model Operations): A set of practices, tools and processes to efficiently and scalably manage the complete LLM lifecycle in production environments, covering training, deployment, monitoring, updating and governance of these models.

Machine learning: Branch of artificial intelligence that focuses on the development of algorithms and models that allow systems to learn and improve automatically through experience, without being explicitly programmed to do so.

Machine unlearning: A set of techniques to selectively remove or "unlearn" certain information or unwanted biases from an already trained machine learning model, without the need to retrain it from scratch, allowing compliance with privacy requirements or correct unwanted behaviors.

Quantitative metrics: Standardized numerical measures used to objectively and consistently evaluate the performance of an AI model on specific tasks, such as precision, completeness, accuracy or efficiency.

Generative model: A type of machine learning model designed to learn the underlying probability distribution of a data set and generate new samples that are similar to the training data and can create new and realistic content.

Pre-training: The initial stage of LLM training in which a large corpus of unstructured and unlabeled text is used for the model to learn general representations and language patterns, acquiring a broad and robust knowledge that can then be adapted to specific tasks by fine-tuning.

Differential privacy: A cryptographic technique used to share aggregated information about a dataset, while protecting the privacy of the individuals present in that data, by introducing random noise that makes it difficult to identify individual entries from the analysis results.

Prompt engineering: Discipline that focuses on designing, optimizing and adapting prompts (text inputs) to obtain the best possible results from LLMs in specific tasks, taking advantage of techniques such as the inclusion of examples, the specification of formats or step-by-step guidance.

A/B testing: An experimental method used to compare the performance of two different versions of an AI system (A and B) or between an AI system and an alternative approach (such as a human or a base model), in order to determine which performs better according to predefined metrics.

AI regulation: The set of laws, regulations, standards and guidelines established by governments and organizations to ensure that the development, deployment and use of artificial intelligence systems is conducted responsibly, safely, ethically and in line with society's fundamental values and rights.

Retrieval-Augmented Generation (RAG): a technique used in LLMs that consists of retrieving relevant information from an external knowledge base before generating a response, thus combining the ability to access structured information with the generation of coherent and fluent natural language.

RNN (Recurrent Neural Network): A type of neural network designed to process sequences of data, such as text or time series. Unlike feedforward neural networks, RNNs have recurrent connections that allow them to maintain internal state and capture temporal dependencies. Variants such as LSTM and GRU have been widely used in natural language processing tasks before the rise of transformers.

AI safety: The discipline that focuses on identifying, preventing and mitigating potential risks associated with the development and use of advanced AI systems, both in the short and long term, including security risks, biases, errors, misuse or unintended consequences.

Bias: Systematic tendency of a machine learning model to produce results that unfairly favor or disadvantage certain groups or individuals, due to sensitive characteristics such as gender, ethnicity, age or sexual orientation, and usually resulting from biases present in the training data or suboptimal decisions during model development.

Token: A discrete unit into which a text is divided for processing by a language model. Tokens can be words, subwords or characters, and are the basic input for LLM training and inference.

Tokenization: The process of converting a text into a sequence of tokens. The choice of tokenization strategy has a significant impact on the performance and efficiency of the model.

Transformers: A deep neural network architecture that uses attention mechanisms to process and generate sequences in parallel, rather than sequentially like RNNs. It allows capturing long-term and contextual dependencies, being the dominant architecture for LLMs and setting the state of the art in various natural language processing tasks.

Validation: A comprehensive and multidisciplinary process to evaluate an AI system, especially LLM, in terms of performance, robustness, safety, security, fairness, explainability and alignment with ethical and social requirements and values, combining quantitative metrics and qualitative assessment by experts and users.

References


Österreich
España
France
English
日本
Ai
文
United Kingdom
Belgium English
Danmark
United States
Magyarország

Abhyankar, R. et al. (2024). APIServe: Efficient API Support for Large-Language Model Inferencing. https://arxiv.org/abs/2402.01869. arXiv:2402.01869v1.

Alabdulmohsin, I. et al. (2024). CLIP the Bias: How Useful is Balancing Data in Multimodal Learning? https://arxiv.org/html/2403.04547v1. arXiv:2403.04547v1.

Banerjee, I., et al. (2023). MLOps with enhanced performance control and observability. https://arxiv.org/abs/2302.01061. arXiv:2302.01061v1.

Bengio, Y. et al. (2003). A Neural Probabilistic Language Model. https://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

Bréal, M. (1883). Les lois intellectuelles du langage fragment de sémantique. Annuaire de l'Association pour l'encouragement des études grecques en France. Vol. 17 (1883), pp. 132-142. https://www.jstor.org/stable/44253893.

Cambon, A. et al. (2023). Early LLM-based Tools for Enterprise Information Workers Likely to Provide Meaningful Boosts to Productivity. A first update from Microsoft's research initiative on AI and Productivity.

Chen, D. et al. (2023). Data-Juicer: A One-Stop Data Processing System for Large Language Models. https://arxiv.org/abs/2309.02033. arXiv:2309.02033v3.

Chen, Y. et al. (2023). LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models. https://arxiv.org/abs/2309.12307. arXiv:2309.12307v3.

Chiang, C. et al. (2023). Can Large Language Models Be an Alternative to Human Evaluations? https://arxiv.org/abs/2305.01937. arXiv:2305.01937v1.

Chu, T., Song, Z., Yang, C. (2023). How to Protect Copyright Data in Optimization of Large Language Models? https://arxiv.org/abs/2308.12247. arXiv:2308.12247v1.

CIO (2023). Chief AI Officer: What it takes to land the C-suite's hottest new job. https://www.cio.com/article/657977/chief-ai-officer-what-it-takes-to-land-the-c-suites-hottest-new-job.html

Cui, Q. et al. (2022). Contrastive Vision-Language Pre-training with Limited Resources. https://arxiv.org/abs/2112.09331. arXiv:2112.09331v3.

CommetML. https://www.comet.com/site/.

Datta, T. et al. (2023). Who's Thinking? A Push for Human-Centered Evaluation of LLMs using the XAI Playbook. https://arxiv.org/abs/2303.06223. arXiv:2303.06223v1.

Dettmers, T. et al. (2023). QLoRA: Efficient Finetuning of Quantized LLMs https://arxiv.org/abs/2305.14314. arXiv:2305.14314v1

Devlin, J. et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805. arXiv:1810.04805v2.

Duan, J. et al. (2023). Shifting attention to relevance: towards the uncertainty estimation of large language models. https://arxiv.org/abs/2307.01379. arXiv:2307.01379v2.

Dun, C. et al. (2024). Sweeping Heterogeneity with Smart MoPs: Mixture of Prompts for LLM Task Adaptation. https://arxiv.org/abs/2310.02842. arXiv:2310.02842v2.

Elazar, Y. et al. (2021). Measuring and Improving Consistency in Pretrained Language Models. https://aclanthology.org/2021.tacl-1.60/.

Euronews (2023). 2023 was the year AI went mainstream. It was also the year we started to panic about it. https://www.euronews.com/next/2023/12/27/2023-was-the-year-ai-went-mainstream-it-was-also-the-year-we-started-to-panic-about-it

European Parliament (2024). Artificial Intelligence Act / European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 - C9-0146/2021 - 2021/0106(COD)). https://artificialintelligenceact.eu/; https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

European Commission (2024). Knowledge Center on Interpretation. https://knowledge-centre-interpretation.education.ec.europa.eu/en/news/what-large-language-model

Fisher, M., Campagna, G., Choi, E., Lam, M. S., Freund, S. N., Yahav, E.,. (2021). DIY Assistant: A Multi-modal End-User Programmable Virtual Assistant. https://dl.acm.org/doi/10.1145/3453483.3454046.

Gartner (2023). What is generative AI? https://www.gartner.com/en/topics/generative-ai

Google DeepMind (2023). Levels of AGI: Operationalizing Progress on the Path to AGI. Meredith Ringel Morris; Jascha Sohl-Dickstein; Noah Fiedel; Tris Warkentin; Allan Dafoe; Aleksandra Faust; Clement Farabet; and Shane Legg. arXiv:2311.02462v1

Google + Implement (2023). The economic opportunity of generative AI in D9+. An Implement Consulting Group study commissioned by Google.

Gozalo-Brizuela, R., and Garrido-Merchán, E.C. (2023). A survey of Generative AI Applications. https://ar5iv.labs.arxiv.org/html/2306.02781.

Guo, Z. et al. (2023). Evaluating Large Language Models: A Comprehensive Survey. https://arxiv.org/pdf/2310.19736.pdf. arXiv:2310.19736v3.

Guzman, F. et al. (2015). How do Humans Evaluate Machine Translation. https://aclanthology.org/W15-3059.pdf.

Fu, HY. et al. (2023). Estimating Large Language Model Capabilities without Labeled Test Data. https://arxiv.org/abs/2305.14802. arXiv:2305.14802v2.

Fu, X. et al (2024). Tiny Titans: Can Smaller Large Language Models Punch Above Their Weight in the Real World for Meeting Summarization? https://arxiv.org/abs/2402.00841. arXiv:2402.00841.

Goyal, S. et al (2024). LLMGuard: Guarding Against Unsafe LLM Behavior. https://arxiv.org/abs/2403.00826. arXiv:2403.00826v1.

Hendrycks, D. et al (2021). Measuring Massive Multitask Language Understanding. https://arxiv.org/abs/2009.03300. arXiv:2009.03300v3.

Huang, L. et al. (2023). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. https://arxiv.org/abs/2311.05232. arXiv:2311.05232v1.

Hugging Face Datasets (2024). CodeParrot. https://huggingface.co/codeparrot.

IAPP (2024). Global AI Law and Policy Tracker. https://iapp.org/resources/article/global-ai-legislation-tracker/

iDanae 2Q23 (2023): Large Language Models: a new era in artificial intelligence. iDanae Chair. Quarterly Newsletter 2Q23. http://www.idanae-stem.com/

iDanae 1Q24 (2024): Towards a sustainable artificial intelligence. iDanae Chair. Quarterly Newsletter 1Q24. http://www.idanae-stem.com/

Imperial, JM., et al. (2023). Flesch or Fumble? Evaluating Readability Standard Alignment of Instruction-Tuned Language Models. https://arxiv.org/abs/2309.05454. arXiv:2309.05454v2.

IndesIA (2024). Barometer of artificial intelligence adoption in Spanish SMEs. https://www.indesia.org/wp-content/uploads/2024/04/IndesIA.-Barometro-de-adopcion-de-la-inteligencia-artificial-en-las-pymes-espanolas-Edicion-2024.pdf

Jang et al. (2022). Knowledge unlearning for mitigating privacy risks in language models. https://arxiv.org/abs/2210.01504. arXiv:2210.01504.

Jia, C. et al (2021). Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. https://arxiv.org/abs/2102.05918. arXiv:2102.05918v2.

Kahng, M. et al. (2024). LLM Comparator: Visual Analytics for Side-by-Side Evaluation of Large Language Models. https://arxiv.org/abs/2402.10524. arXiv:2402.10524v1.

Kuchnik, M. et al. (2023). Validating Large Language Models with Realm. https://arxiv.org/abs/2211.15458. arXiv:2211.15458v2.

Kudo, T., Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. https://arxiv.org/abs/1808.06226. arXiv:1808.06226v1.

Lam, M. (2018). https://profiles.stanford.edu/monica-lam?tab=publications. Keeping the Internet Open with an Open-Source Virtual Assistant.

Lee, C. et al (2024). OrchestraLLLM: Efficient Orchestration of Language Models for Dialogue State Tracking. https://arxiv.org/html/2311.09758v2. arXiv:2311.09758v2.

Lee, J. et al. (2022). Seq2Seq-SC: End-to-End Semantic Communication Systems with Pre-trained Language Model. https://arxiv.org/abs/2210.15237. arXiv:2210.15237v2.

Lester, B. et al. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. https://arxiv.org/abs/2104.08691. arXiv:2104.08691v2.

Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. https://arxiv.org/abs/2005.11401

Li, H. et al. (2024). Digger: Detecting Copyright Content Mis-usage in Large Language Model Training. https://arxiv.org/abs/2401.00676. arXiv:2401.00676v1.

Li, S. et al (2024). Evaluating Quantized Large Language Models. https://arxiv.org/abs/2402.18158. arXiv:2402.18158v1.

Li, Y. et al (2023). A Survey on Fairness in Large Language Models. https://arxiv.org/abs/2308.10149. arXiv:2308.10149.

Liang, P. et al. (2023). Holistic Evaluation of Language Models. https://arxiv.org/abs/2211.09110. arXiv:2211.09110v2.

Liu, T. et al (2022). Autoregressive Structured Prediction with Language Models. https://arxiv.org/abs/2210.14698. arXiv:2210.14698v2.

Liu, Y. et al (2024). Datasets for Large Language Models: A Comprehensive Survey. https://arxiv.org/abs/2402.18041. arXiv:2402.18041v1.

Liu, Y. et al (2023). Robustness Over Time: Understanding Adversarial Examples' Effectiveness on Longitudinal Versions of Large Language Models. https://arxiv.org/pdf/2308.07847.pdf. arXiv:2308.07847v1.

Luo, Y. et al. (2023). An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning. https://arxiv.org/pdf/2308.08747.pdf. arXiv:2308.08747v3.

Management Solutions (2023). Explainable Artificial Intelligence (XAI): challenges in model interpretability. https://www.managementsolutions.com/en/microsites/whitepapers/explainable-artificial-intelligence

Management Solutions (2022). AutoML, towards the automation of models. https://www.managementsolutions.com/en/publications-and-events/industry-reports/white-papers/auto-machine-learning-towards-model-automation

Management Solutions (2014). Model Risk Management: Quantitative and Qualitative Aspects. https://www.managementsolutions.com/en/publications-and-events/industry-reports/white-papers/model-risk-management

Meeus, M. et al. (2024). Copyright Traps for Large Language Models. https://arxiv.org/abs/2402.09363. arXiv:2402.09363v1.

Mehta, S.V. et al. (2023). An Empirical Investigation of the Role of Pre-training in Lifelong Learning. https://arxiv.org/abs/2112.09153. arXiv:2112.09153v2.

Mikolov, T. et al. (2013). Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/abs/1301.3781. arXiv:1301.3781v3.

Minaee, S. et al. (2024). Large Language Models: A Survey. https://arxiv.org/abs/2402.06196. arXiv:2402.06196v2.

MindsDB (2024). A Comparative Analysis of Leading Large Language Models. https://mindsdb.com/blog/navigating-the-llm-landscape-a-comparative-analysis-of-leading-large-language-models

Mökander, J. et al. (2023). Auditing large language models: a three-layered approach. arXiv:2302.08500v2.

Nasr, M., et al. (2023). https://arxiv.org/pdf/2311.17035.pdf. arXiv:2311.17035v1.

Neelakantan, A. et al. (2022). Text and Code Embeddings by Contrastive Pre-Training. https://arxiv.org/abs/2201.10005. arXiv:2201.10005v1.

NIST (2023). AI Risk Management Framework : NIST. https://www.nist.gov/itl/ai-risk-management-framework

Oneto, L., Chiappa, S. (2020). Fairness in Machine Learning. 2012.15816.pdf (arxiv.org) arXiv:2012.15816v1.

OpenAI (2024). Prompt engineering. https://platform.openai.com/docs/guides/prompt-engineering

Ovadia, O. et al (2024). Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. https://arxiv.org/pdf/2312.05934.pdf. arXiv:2312.05934v3.

Pankajakshan, R. et al (2024). Mapping LLM Security Landscapes: A Comprehensive Stakeholder Risk Assessment Proposal. https://arxiv.org/html/2403.13309v1. arXiv:2403.13309v1.

Parikh, A. P., et al. (2016). A Decomposable Attention Model for Natural Language Inference. https://arxiv.org/abs/1606.01933. arXiv:1606.01933v2.

Penedo, G. et al (2023). The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. https://arxiv.org/abs/2306.01116. arXiv:2306.01116v1.

Pew Research Center (2023). Experts Predict the Best and Worst Changes in Digital Life by 2035.

Project Gutenberg (2024). https://www.gutenberg.org/.

Rae, JW, et al (2021). Scaling Language Models: Methods, Analysis & Insights from Training Gopher. https://arxiv.org/abs/2112.11446. arXiv:2112.11446.

Rafailov, R. et al (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. https://arxiv.org/abs/2305.18290. arXiv:2305.18290v2.

Rejeleene, R.; Xu, X.; Talburt, J.; (2024). Towards Trustable Language Models: Investigating Information Quality of Large Language Models. https://arxiv.org/abs/2401.13086. arXiv:2401.13086v1.

Risk.net. (2024). The bank quant who wants to stop gen AI hallucinating. https://www.risk.net/risk-management/7959062/the-bank-quant-who-wants-to-stop-gen-ai-hallucinating.

Sachdeva, N., et al (2024). How to Train Data-Efficient LLMs. https://arxiv.org/html/2402.09668v1. arXiv:2402.09668v1.

Samsi, S., et al (2023). From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. https://arxiv.org/pdf/2310.03003.pdf. arXiv:2310.03003v1.

Sarti, G. et al (2023). Inseq: An Interpretability Toolkit for Sequence Generation Models. [2302.13942] Inseq: An Interpretability Toolkit for Sequence Generation Models (arxiv.org). arXiv:2302.13942v3.

Searle, J. (1980). Minds, Brains, and Programs. The Behavioral and Brain Sciences, vol. 3. Cambridge University Press. https://web.archive.org/web/20010221025515/http://www.bbsonline.org/Preprints/OldArchive/bbs.searle2.html

Shaikh, O. et al. (2022). On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. https://arxiv.org/abs/2212.08061. arXiv:2212.08061v2.

SHAP documentation. https://shap.readthedocs.io/

Shaw, P. et al (2018). Self-Attention with Relative Position Representations. https://arxiv.org/abs/1803.02155. arXiv:1803.02155v2.

Sherstinsky, A. (2018). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. https://arxiv.org/abs/1808.03314. arXiv:1808.03314v10.

Shi, W. et al (2024). Detecting pretraining data from large language models. https://arxiv.org/abs/2310.16789. arXiv:2310.16789v3.

Singh, C. et al (2024). Rethinking Interpretability in the Era of Large Language Models. https://arxiv.org/abs/2402.01761. arXiv:2402.01761v1.

Sinha, K. et al (2021). Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. https://arxiv.org/abs/2104.06644. arXiv:2104.06644v2.

Soskek (2019). BookCorpus. https://github.com/soskek/bookcorpus.

Su, J., et al (2021). Roformer: Enhanced transformer with rotary position embedding. https://arxiv.org/abs/2104.09864. arXiv:2104.09864.

Sutskever, I. et al (2014). Sequence to Sequence Learning with Neural Networks. https://arxiv.org/abs/1409.3215. arXiv:1409.3215v3.

The Next Web (2023). When will AGI arrive? Here's what our tech lords predict. https://thenextweb.com/news/when-will-agi-arrive-tech-experts-predict-artificial-general-intelligence

Tian, Y. et al (2024). TinyLLLM: Learning a Small Student from Multiple Large Language Models. https://arxiv.org/abs/2402.04616. arXiv:2402.04616.

Tirumala, K. et al. (2023). D4: Improving LLM Pretraining via Document De-Duplication and Diversification. https://arxiv.org/abs/2308.12284. arXiv:2308.12284v1.

UK Government (2023). The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Vartziotis, T. et al (2024). Learn to Code Sustainably: An Empirical Study on LLM-based Green Code Generation. https://arxiv.org/html/2403.03344v1. arXiv:2403.03344v1.

Vaswani, A. et al. (2017). Attention Is All You Need. https://arxiv.org/abs/1706.03762

Wan, Z. et al (2024). Efficient Large Language Models: A Survey. https://arxiv.org/pdf/2312.03863.pdf. arXiv:2312.03863v3.

Wang, Q. et al (2024). LLMCheckup: Conversational Examination of Large Language Models via Interpretability Tools. [2401.12576] LLMCheckup: Conversational Examination of Large Language Models via Interpretability Tools (arxiv.org). arXiv:2401.12576v1.

Wang, Y. et al (2024). Two-stage LLM Fine-tuning with Less Specialization and More Generalization. https://arxiv.org/html/2211.00635v3. arXiv:2211.00635v3.

Wei, J. et al (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. https://arxiv.org/abs/2201.11903. arXiv:2201.11903v6.

Wenzek, G., et al (2019). CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. https://arxiv.org/abs/1911.00359. arXiv:1911.00359v2.

Wettig, A. et al. (2024). QuRating: Selecting High-Quality Data for Training Language Models. https://arxiv.org/abs/2402.09739. arXiv:2402.09739v1.

Weights & Biases: The AI Developer Platform (wandb.ai). https://wandb.ai/site

Wikipedia (2024). Dumps. https://dumps.wikimedia.org/zhwiki/latest/.

Wired (2023). OpenAI's CEO Says the Age of Giant AI Models Is Already Over. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/

Weizenbaum, J. (1966). ELIZA-a computer program for the study of natural language communication between man and machine. https://dl.acm.org/doi/10.1145/365153.365168

White House (2022). Blueprint for an AI Bill Of Rights. Making Automated Systems Work for the American People. https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

White House (2023). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Wu, X. et al. (2023). Depn: Detecting and editing privacy neurons in pretrained language models. https://arxiv.org/abs/2310.20138. arXiv:2310.20138.

Xin Zhao, W., et al. (2023). A Survey of Large Language Models. https://arxiv.org/abs/2303.18223. arXiv:2303.18223v13.

Xu, L. et al. (2023). Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. https://arxiv.org/pdf/2312.12148.pdf. arXiv:2312.12148v1.

Xu, Y. et al. (2021). Non-Autoregressive Text Generation with Pre-trained Language Models. https://aclanthology.org/2021.eacl-main.18/

Xu, Z. et al. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. https://arxiv.org/abs/2401.11817. arXiv:2401.11817v1.

Yang, J. et al. (2024). Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. https://arxiv.org/abs/2304.13712. arXiv:2304.13712v2.

Yidiz, C. et al (2024). Investigating Continual Pretraining in Large Language Models: Insights and Implications. https://arxiv.org/html/2402.17400v1. arXiv:2402.17400v1.

Yu, C. et al. (2023). Unlearning bias in language models by partitioning gradients. https://aclanthology.org/2023.findings-acl.375.pdf.

Yogarajan, V., et al (2023). Tackling Bias in Pre-trained Language Models: Current Trends and Under-represented Societies. https://arxiv.org/pdf/2312.01509.pdf. arXiv:2312.01509v1.

Zaharia, M. et al (2018). Accelerating the Machine Learning Lifecycle with MLflow. https://people.eecs.berkeley.edu/~matei/papers/2018/ieee_mlflow.pdf.

Zeng, Y., et al (2023). CLIP2: Contrastive Language-Image-Point Pretraining from Real-World Point Cloud Data. https://arxiv.org/abs/2303.12417. arXiv:2303.12417v2.

Zhang, B. et al (2024). When Scaling Meets LLM Finetuning: The Effect of Data, Model and Finetuning Method. https://arxiv.org/abs/2402.17193. arXiv:2402.17193v1.

Zhang, L. et al (2024). Enhancing Large Language Model Performance To Answer Questions and Extract Information More Accurately. https://arxiv.org/html/2402.01722v1. arXiv:2402.01722v1.

Zhang, S. et al (2023). Instruction Tuning for Large Language Models: A Survey. https://www.researchgate.net/publication/373263398_Instruction_Tuning_for_Large_Language_Models_A_Survey.

Zhang, Y. et al (2024). Bias Mitigation in Fine-tuning Pre-trained Models for Enhanced Fairness and Efficiency. https://arxiv.org/html/2403.00625v1. arXiv:2403.00625v1.

Zhao, B., et al (2023). Tuning LayerNorm in Attention: Towards Efficient Multi-Modal LLM Finetuning. https://arxiv.org/abs/2312.11420. arXiv:2312.11420v1.

Zhou, C. et al (2023). LIMA: Less Is More for Alignment. https://arxiv.org/abs/2305.11206. arXiv:2305.11206v1.

Zhou, N., et al (2021). Bias, Fairness, and Accountability with AI and ML Algorithms. https://arxiv.org/abs/2105.06558. arXiv:2105.06558v1.

Our aim is to exceed our clients' expectations, and become their trusted partners

Management Solutions is an international consulting services company focused on consulting for business, risks, organization and processes, in both their functional components and in the implementation of their related technologies.

With its multi-disciplinary team (functional, mathematicians, technicians, etc.) of more than 4,000 professionals, Management Solutions operates through its 46 offices (20 in Europe, 22 in the Americas, 2 in Asia, 1 in Africa and 1 Oceania).

To cover its clients' needs, Management Solutions has structured its practices by sectors (Financial Institutions, Energy, Telecommunications and other industries) and by lines of activity, covering a broad range of skills -Strategy, Sales and Marketing Management, Risk Management and Control, Management and Financial Information, Transformation: Organization and Processes, and New Technologies.

The R&D department provides advisory services to Management Solutions's professionals and their clients in quantitative aspects that are necessary to undertake projects with rigor and excellence through the implementation of best practices and the continuous monitoring of the latest trends in artificial intelligence, data science, machine learning, modeling and big data.

Javier Calvo Martín
Partner at Management Solutions
javier.calvo.martin@managementsolutions.com

Manuel Ángel Guzmán Caba
Partner at Management Solutions
manuel.guzman@managementsolutions.com

Luz Ferrero Peña
Manager at Management Solutions
luz.ferrero@msgermany.com.de

Segismundo Jiménez Láinez
Manager at Management Solutions
segismundo.jimenez@msspain.com

Management Solutions, Professional Consulting Services

Management Solutions is an international consulting firm whose core mission is to deliver business, risk, financial, organization, technology and process-related advisory services.

For further information please visit **www.managementsolutions.com**

Follow us at:

www.managementsolutions.com

Madrid Barcelona Bilbao Coruña Málaga London Frankfurt Düsseldorf Paris Amsterdam Copenhagen Oslo Stockholm Warszawa Wroclaw Zürich Milano Roma Bologna Lisboa Beijing Istanbul Johannesburgo Sydney Toronto New York New Jersey Boston Pittsburgh Atlanta Birmingham Houston Miami SJ de Puerto Rico San José Ciudad de México Monterrey Querétaro Medellín Bogotá Quito São Paulo Rio de Janeiro Lima Santiago de Chile Buenos Aires