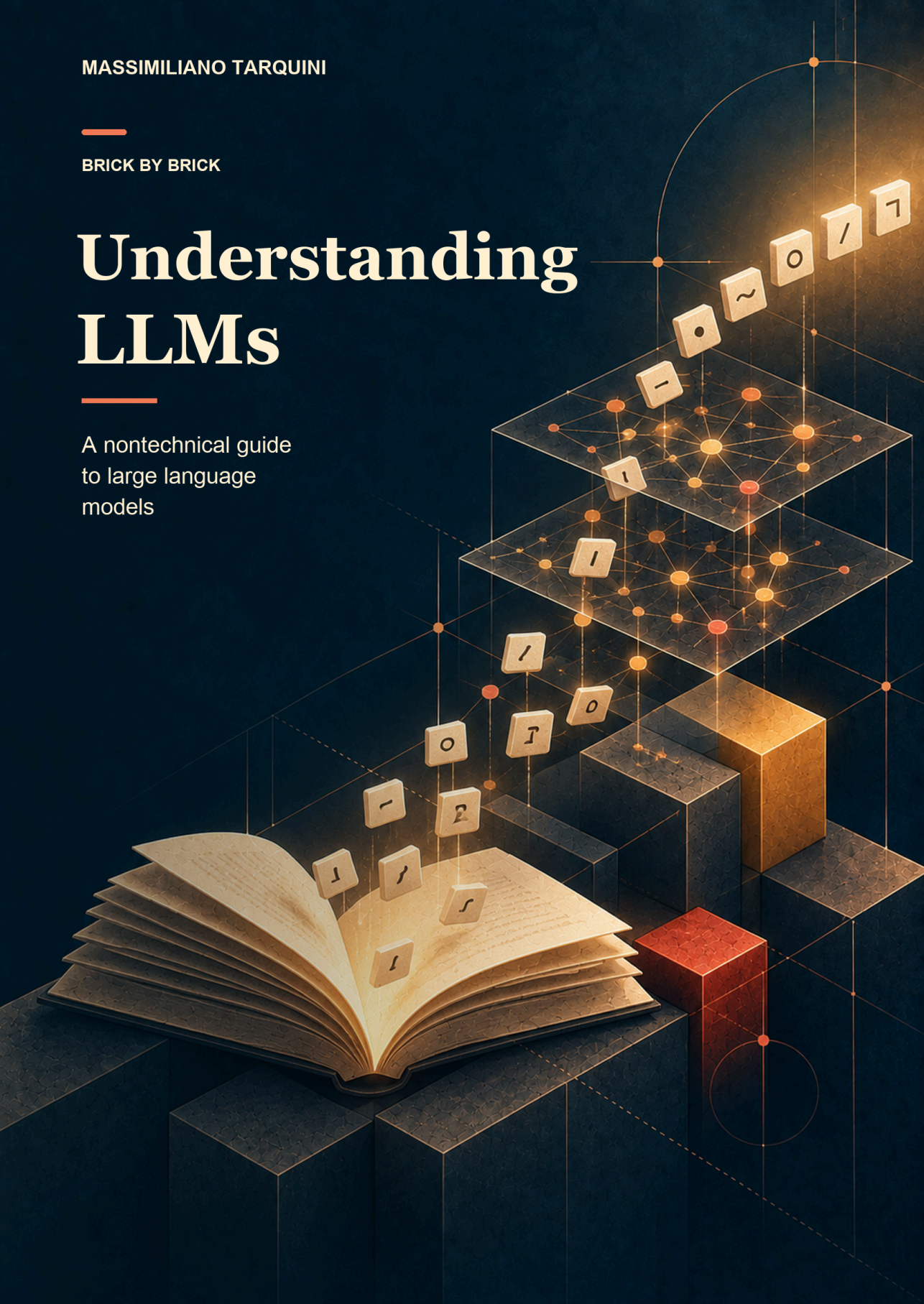


MASSIMILIANO TARQUINI

BRICK BY BRICK

Understanding LLMs

A nontechnical guide
to large language
models



BRICK BY BRICK

Understanding LLMs

A nontechnical guide

Massimiliano Tarquini

AiOne

Contents

Part I — Foundations	5
Introduction	5
What Are Large Language Models	5
How They Work Without Going into Too Much Detail	6
Part II — Capabilities and Interaction	9
What LLMs Can Do	9
Talking to LLMs — Why Prompts Matter and How Interfaces Are Designed	10
Emergent Abilities — What Models Learn on Their Own	11
Beyond Questions — How to Use Prompting to Harness Emergent Abilities	15
From Words to Abilities — Practical Examples of Emergent Prompting	17
Part III — Ethics, Limitations, and Society	19
Important Questions — What We Should Be Asking About LLMs	19
The Future of LLMs — Risks and Limitations of What They Cannot and Should Not Do	20
Using LLMs Wisely	20
Human-Centered AI and Explainability — Trusting What You Can Understand	23
Asking Models to Explain Themselves	24
Practical Examples of Explainability in Action	24
Myths and Reality — What LLMs Cannot Do Even When They Seem Able To	27
When Overconfidence Meets Artificial Intelligence	29
A Final Reflection	33
Part IV — Applications and Examples	34
Real-World Examples	34
Everyday Tools Quietly Powered by AI	35
Opportunities and Benefits — Why LLMs Can Be Truly Useful	35
Unexpected Stories from the World of Language Models	36

Contents

Part V — From General to Specialized Intelligence	38
From General to Specialized — How LLMs Learn New Skills	38
LLMs as Memory Machines — Retrieving Knowledge and Human Experience	39
A Whole World of Language Models — From Giants to Little Helpers	40
Not Just Words — When Language Models See and Understand Images	42
Part VI — Infrastructure and Sustainability	44
Power and Price — The Hidden Cost of Intelligence	44
Comparing AI Models — Cloud Giants and Local Experts	45
Part VII — Appendices	47
Before You Ask — A Checklist for Using LLMs Wisely	47
Conclusion	47
Glossary	48
Want to Know More — Resources to Explore	49

Part I — Foundations

Note on the revised Italian edition. This volume preserves the progression and topics of the original guide, supplementing them with technical corrections, verifiable references, and guided activities. Undocumented stories are explicitly presented as illustrative scenarios. The previous complete translation remains available separately. To follow the examples, you only need to be able to read a short table and compare two texts; no programming knowledge is required. The aim is to learn how to choose a use for the model, check its result, and recognize when another solution is needed.

Introduction

Artificial intelligence is no longer confined to science fiction: it is part of our everyday lives. You may already have interacted with it without realizing it, when your phone suggests words as you type, when a chatbot answers your questions, or when a tool helps you rewrite an email.

At the heart of many of these intelligent tools are Large Language Models, or LLMs for short. These artificial intelligence systems are changing the way we work, learn, and communicate. But what exactly are they? How do they work? And why are they attracting so much attention?

This guide helps you understand what LLMs are without resorting to technical jargon. Whether you are curious, concerned, or eager to keep up with the latest trends, you will find simple, clear answers.

The document will guide you through these topics:

- What LLMs are, how they are trained, and what they can do.
- How we interact with them through natural language prompts.
- Why they sometimes surprise us with unexpected abilities.
- What risks and limitations they have: bias, hallucinations, and lack of understanding.
- How to use them wisely, with a practical checklist.
- Where they are used today, from healthcare to education and creativity.
- New frontiers, such as multimodal models that see and understand images.
- How much they cost in energy, money, and computing power.

You do not need to be a developer to understand all this. All it takes is a little curiosity and a desire to engage with one of the most powerful technologies of our time, responsibly, creatively, and with a critical mindset.

What Are Large Language Models

Imagine wanting to teach a computer to understand and use language. Not just with a dictionary or a grammar book, but by letting it read almost everything: novels, newspaper articles, Wikipedia pages, social media posts, technical manuals, poems, and even movie scripts.

This is how large language models, or LLMs, are trained. They are exposed to enormous amounts of text so that they can learn how people actually speak and write in all kinds of situations.

But here is the surprising part:

Many generative LLMs initially learn to predict the next token. Further stages of training can adapt them to instructions and tasks.

A token is a unit of text, often a word or part of one. Prediction uses learned representations of relationships between elements: the mechanism is simple to describe, but the resulting abilities need to be tested. From the training objective alone, we infer neither human understanding nor the absence of reasoning abilities.

For example:

- If you write “The sky is full of...”, it might suggest stars, clouds, or planes.
- If you ask “Can you explain photosynthesis?”, it draws on everything it has read to construct an answer that appears useful and coherent.
- Ask it to write a poem, summarize an article, translate a sentence, or even help you with code: it is often able to do so.

Why does it work? Because language follows recurring patterns. The more examples the model sees, the better it becomes at guessing what comes next; these predictions begin to look a lot like reasoning, memory, or creativity.

This is why LLMs are often described as statistical engines for language. They do not know facts the way humans do, but they are extraordinarily good at imitating human communication on the basis of what they have read.

Here are some examples of what they can do:

- Write letters, blog posts, bedtime stories, and speeches.
- Converse: answer questions, hold conversations, and even play a role.
- Translate texts between different languages, including informal language and slang.
- Summarize long documents, meeting minutes, and academic articles.
- Help with code: explain programming concepts or generate code snippets.

They are used in virtual assistants, customer service bots, writing tools, educational applications, and creative projects. So, even though an LLM is just a very advanced word predictor, exposure to human language allows it to perform a surprising variety of tasks, often in ways that seem entirely natural.

How They Work Without Going into Too Much Detail

Suppose you are writing to a friend and type:

“I feel really...”

Before you even finish the sentence, your phone might suggest words such as tired, happy, or excited. This is a small example of what a language model does: it tries to guess the next word on the basis of the ones you have already written.

Now imagine doing this not just with a few sentences, but with billions of sentences, in different styles and on different topics. This is what LLMs do. They become extremely good at predicting what comes next in language.

At the heart of this process is a type of artificial intelligence called deep learning, which uses neural networks. These are digital systems loosely inspired by how the human brain works, made up of layers that process information step by step. Each layer looks for recurring patterns:

- One might deal with spelling or word structure.
- Another might focus on grammar.
- Yet another might look for deeper contextual elements, such as tone or intent.

The division between spelling, grammar, and context is an analogy, not a literal map of the layers: representations are distributed, and we do not assign just one task to each layer. Parameters are adjusted during training. Over time, the model learns to make increasingly accurate predictions that are increasingly similar to human ones.

A Simple Analogy — LEGO Bricks

Imagine building a sentence with LEGO bricks. The model learns which bricks tend to fit well together: “peanut butter” often appears with “jam”, “doctor” might appear with “hospital”, and “climate” with “change”.

It does not know the meaning of these words the way humans do, but it knows how they tend to appear together.

Why Size Matters

Increasing the amount of data and the model's capacity can improve performance, but the result also depends on data quality, architecture, training, and the task. A larger model is not automatically better for every use. This is why we call them large language models: they consist of billions of parameters, similar to little knobs that the model adjusts to improve its predictions.

The result? A system that can:

- Write a press release in a few seconds.
- Answer complex questions.
- Simulate a conversation.
- Even explain jokes or wordplay, at least sometimes.

But remember:

Key concept — Prediction does not just mean completing sentences. A generative model produces tokens, units of text that can be words or parts of words, using learned representations. This mechanism can support complex tasks; by itself, it does not demonstrate consciousness, personal experience, or factual reliability.

Guided activity — Completing is not the same as verifying. You need to know an opening time, but the available text only says “The store opens at”. A plausible completion does not solve the problem. Instead, provide “The store opens at 9” and ask for the time to be extracted without adding any information: the expected result is 9. Check it against the text, then remove the time and check that its absence is reported. The comparison separates generation from verification; reuse it to extract explicit information, without inferring external facts from it.

Part II — Capabilities and Interaction

What LLMs Can Do

Large language models are like the Swiss Army knives of the digital world: versatile, powerful, and full of unexpected tools. Training in token prediction and subsequent stages can support many activities, which need to be evaluated in each specific case.

Here are some of the things they can do, with concrete examples.

Writing and Editing

Need help drafting an email? The model can suggest a polite, professional tone. Struggling with writer's block? It can suggest a few opening sentences. Want to rewrite a document in simpler language? It takes just a few seconds.

Example: “Dear Professor, I am writing to request an extension...” → the LLM completes the text in formal, respectful language.

Summarizing Long Texts

No time to read a ten-page report? Ask the model for a brief summary. Reviewing meeting notes or an organization's policy documents? The model can extract their main points.

Example: “Summarize this article about climate change in three bullet points.”

Translation and Language Help

Translate texts between different languages with surprising fluency. Rephrase a sentence to make it more formal, informal, or clear. Explain idioms, expressions, and even slang.

Example: “What does ‘break the ice’ mean?” → the model explains it like a friendly teacher.

Conversation and Assistance

Simulate a helpful assistant for customer service or technical support. Answer common questions about a product, policy, or procedure.

Example: an LLM-based chatbot can guide users through setting up Wi-Fi or booking an appointment.

Tutoring and Explanations

Help students understand complex topics by rephrasing them in simpler terms. Generate quiz questions or practice examples. Act as a “study buddy” that explains things step by step.

Example: “Explain gravity to me as if I were ten years old” → the model uses an accessible, friendly analogy.

Programming and Technical Help

Suggest solutions to programming problems. Generate small code snippets. Help identify errors or explain them.

Example: “Write a Python function that checks whether a number is prime.”

Understanding Data with the Right Setup

When connected to data sources or dashboards, LLMs can describe trends in natural language. They can help turn numbers into a narrative.

Example: “Explain what this chart means in simple words.”

As you can see, LLMs are not just for technology experts: they are already helping writers, students, professionals, healthcare workers, journalists, and artists. They do not replace human expertise, but they can amplify it, acting as digital copilots.

Talking to LLMs — Why Prompts Matter and How Interfaces Are Designed

Large language models do not use buttons, menus, or predefined commands. Instead, they respond to natural language inputs, also called prompts. This is more than a feature that makes them easier to use: it is a fundamental design principle.

Unlike traditional software, which follows rigid instructions, LLMs operate in a flexible, probabilistic space. A prompt acts as both input and context: it defines the task, sets the tone, and often suggests the expected response format.

Why Use Prompt-Based Interfaces

Prompts allow open-ended interaction. This is essential because LLMs are general-purpose systems. Rather than building a new application for every function, we can reuse the same model for different tasks simply by changing the prompt.

For example:

“Write a formal email requesting a meeting.”

“Explain the laws of thermodynamics in simple terms.”

“Summarize this company policy document in five bullet points.”

In each case, the user defines in natural language what the model should do.

Prompting as a Control Layer

In technical terms, prompting is a flexible interface layer. It does not change the underlying model, but it guides its behavior. A well-constructed prompt can:

- Assign a role, for example “You are a legal adviser...”.
- Set a task, for example “Draft a contractual clause on confidentiality.”
- Define the style or format, for example “Write a bulleted list of no more than one hundred words.”

This is why prompting is often described as a form of programming with words: through language that guides the model's outputs, rather than through code.

From Casual Use to Prompt Engineering

Simple prompts work for everyday tasks. But when users begin to go further, combining multiple constraints, using system messages, or chaining instructions, prompting becomes a skill in its own right.

This is where prompt engineering comes from: the practice of constructing inputs that optimize the relevance, precision, or creativity of the output.

Good prompting is not based on random attempts. It requires understanding how LLMs interpret and weigh textual cues, managing ambiguity to avoid hallucinations, and structuring inputs for multistep reasoning or domain-specific tasks.

Interfaces Built Around Language

Whether in chatbots, embedded assistants, development environments, or dashboards, the trend is clear: language itself is becoming the interface. Prompt-based design reduces barriers, increases accessibility, and adapts to the way humans naturally communicate.

This does not mean that graphical interface elements will disappear. Many tools already combine form-based inputs with prompt fields, offering users both structure and freedom.

Guided activity — Transforming a vague request. You want a confirmation email, but “Write an email” leaves the purpose and recipient undefined. Try “Write an email to a colleague confirming the meeting on Tuesday at 10; friendly tone; maximum 80 words; do not invent a location or names.” Check every constraint. As a variation, change the recipient and keep the content: the tone should adapt without changing the date or time. Reuse the method for textual tasks with constraints; do not expect a detailed prompt to guarantee true facts.

Emergent Abilities — What Models Learn on Their Own

This is where things become truly fascinating. As large language models grow and become more complex, they start doing things that nobody explicitly taught them.

Many generative models are initially trained to predict the next token; further training stages may follow, using examples of instructions, preferences, and tasks. During this process, abilities develop that are measured through specific tests. It is as if the model starts “picking up” more aspects of the world than expected.

What Emergent Abilities Are

We speak of emergent abilities when an evaluation detects behavior in larger models that was not detectable in smaller ones. This is not a law by which every ability appears beyond a universal threshold: data, training, prompts, and the metric, the rule used to assign a score, also matter.

Here are some examples:

- Solving riddles or logic puzzles: give the model a puzzle and it might find the answer.
- Writing working code: describe a task in natural language and the model generates code that works.
- Solving mathematical word problems: it can follow the logic of a question and find the right formula to solve it.
- Recognizing sarcasm or hidden meanings: it picks up subtle cues and the tone of language.
- Playing roles or imagining perspectives: it can simulate different viewpoints or “pretend” to be a teacher, doctor, or historical figure.

Why It Happens

When you train a model on enormous amounts of human text, millions of conversations, stories, explanations, and debates, the model begins to generalize. This means it can apply what it has seen in one context to a new, unfamiliar situation.

We can describe this behavior as “artificial intuition” only metaphorically: the model formulates hypotheses from learned representations. To evaluate them, we need to check whether they solve the problem, without inferring an inner experience.

It is similar to what happens to someone who reads a lot and can often answer questions or explain things that nobody formally taught them, because they have absorbed so many examples.

A Kind of Artificial Intuition

“Theory of mind” refers to the ability to attribute beliefs, knowledge, and intentions to others. Some language tests ask us to predict the response of a character who has different information from ours. Success in these tests must be distinguished from proof that the model has the same human ability in every situation.

How We Discovered Emergent Abilities

Many of the surprising abilities of LLMs were not explicitly designed. Nobody wrote code to teach them to solve riddles, write working programs, or translate less widely spoken languages. Yet during testing and exploration, models such as GPT-3 and GPT-4 began to display behavior that went far beyond simple word prediction.

How did we notice? Partly by chance, partly out of curiosity. Researchers and users began experimenting with different prompts, asking the model to “explain like a teacher,” “reason step by step,” or solve logic puzzles. Responses could be coherent and accurate on some tests. The generalization of learned representations contributes to this performance; the fluency of a response alone does not allow us to reconstruct the process or establish its reliability on another test.

Wei and colleagues discussed these abilities in 2022. Schaeffer, Miranda, and Koyejo later showed that some apparent jumps depend on the metric: a gradual improvement may seem sudden if the score distinguishes only success from failure. The debate invites us to examine how we measure performance, as well as model size.

Essential reference — Emergence and measurement. Wei and colleagues, Emergent Abilities of Large Language Models (2022), describe results in which scale is associated with new performance. Schaeffer, Miranda, and Koyejo, Are Emergent Abilities of Large Language Models a Mirage? (2023), discuss the effect of metrics. The two works help separate the measured ability from the explanation of its appearance; they do not establish a threshold valid for all models. Full references are provided in the final resources.

A Glimpse of Theory of Mind

In some studies, large models have demonstrated the ability to predict what another person might think: a faculty called theory of mind. For example, if a character in a story does not know something the reader knows, the model might still keep that character's knowledge limited.

Does this mean the model is conscious? No.

But it shows how far statistical pattern recognition can go when scaled to enormous sizes.

Emergent abilities are part of what makes LLMs both exciting and mysterious. They raise big questions about intelligence, creativity, and what machines might be able to do in the future.

These functions are not explicitly coded: they emerge from learning patterns in enormous volumes of human text.

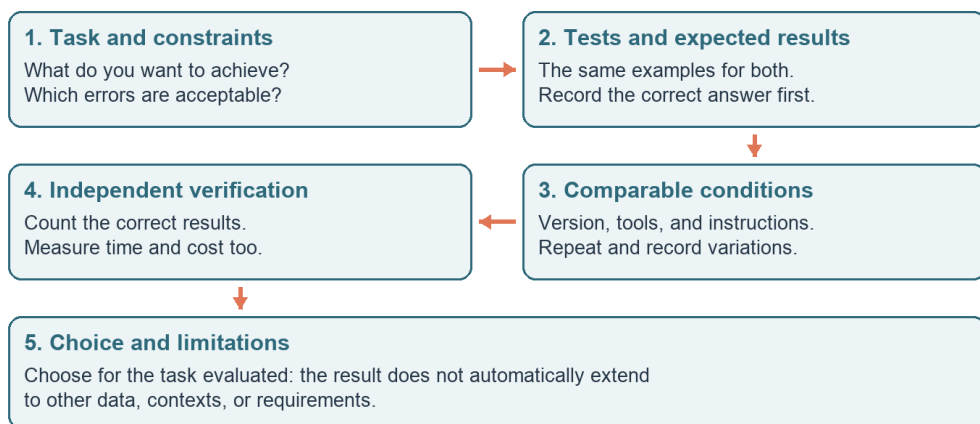
Ability	What the model can do	
Solve logical and mathematical problems	Uses step-by-step reasoning to solve puzzles, equations, or word problems, with increasing accuracy.	Shows early forms of symbolic reasoning and the ability to generalize logical patterns.
Write code from natural language	Generates working code in Python, JavaScript, and other languages, explains errors, and improves existing scripts.	Transforms language into executable logic, linking human reasoning to machine reasoning.

Ability	What the model can do	
Understand multimodal content	Interprets images, charts, or screenshots and explains them in words, as happens, for example, with GPT-4V and Gemini.	Connects text to vision, a crucial aspect for education, medicine, and research.
Reason across multiple exchanges	Maintains logical consistency and context during a conversation or multistep analysis.	Enables extended dialogues, planning, and task decomposition.
Recognize tone, emotions, and sarcasm	Recognizes humor, irony, and emotional tone in conversation.	Makes human-AI interaction more natural and sensitive to context.
Translate less widely spoken or mixed languages	Handles low-resource languages, dialects, and code-switching, the alternation between two languages.	Expands access and inclusion across different language communities.
Generate structured data	Produces tables, JSON, or schema-based outputs directly from natural text.	Is essential for automation, data pipelines, and integration with business tools.
Infer hidden context in theory-of-mind tests	Predicts what a character in a story knows or believes, distinguishing it from what another knows or believes.	Is an early sign of modeling context, perspective, and intention.
Simulate roles or experts	Takes on personas such as teacher, doctor, or analyst and responds accordingly.	Makes AI adaptable to users' goals and communication styles.
Explain an answer through intermediate steps	Can propose a sequence of steps to check.	Improves the readability of the solution, without certifying the internal process that produced it.
Generate hypotheses and analogies	Suggests possible explanations or creative comparisons.	Supports ideation, research, and education.
Self-correct and reflect	Revises or critiques its own previous output when asked.	Introduces iterative refinement and reasoning about reasoning.
Follow complex multistep prompts	Executes sequences such as “analyze → reason → summarize → advise.”	Provides the foundation for tool use and agentic behavior.

Surprising things LLMs have learned to do

Model performance changes over time, but a comparison requires defined tasks and conditions. The radar chart in the original guide assigned scores to GPT-4 and GPT-5 without specifying tests, scale, or method. In this edition, we replace it with the following diagram: it serves to construct a verifiable comparison without introducing new measurements.

A comparison you can verify



Method diagram • No scores assigned to models

Figure 1 — How to compare two models on the same task. The diagram describes a method and does not assign scores to products.

The figure connects five steps: choose the task, prepare examples with an expected result, keep conditions comparable, check outputs, and decide on the basis of quality, cost, and time. A benchmark is a set of tests with an evaluation rule; it measures performance on those tests, not general intelligence on a single scale.

Worked example — A useful comparison. You need to extract the date and amount from ten fictional receipts. First, you manually record the correct answers; then you send the same texts to both models, with the same tools and the same requested format. You count the correct fields, record the time, and repeat the test, because outputs can vary. If one model is more accurate but too slow for your service, the choice remains open. As a variation, add a receipt without a date: a reliable model for this task must flag the missing information instead of inventing it. Reuse the comparison only for similar inputs and conditions; an extraction test does not demonstrate an ability to provide medical advice.

Beyond Questions — How to Use Prompting to Harness Emergent Abilities

Prompting is not about giving orders, but about tuning a resonance. When we write to a language model, we are not merely asking it to produce words: we shape how it connects the many layers of knowledge and reasoning it already possesses. A prompt can encourage reasoning, analogy, creativity, and self-correction on the task; it does not guarantee that an ability is present or exercised correctly.

A prompt can provide new instructions, facts, and examples in the context of the response. This adaptation does not imply that parameters are updated: we distinguish what the system uses in the conversation from what it learned during training.

How Prompts Can Activate Emergent Abilities

Emergent behaviors often require explicit contextual cues. The same model can respond superficially or brilliantly, depending on how the task is formulated. By constructing prompts that reflect how humans approach reasoning, we can help the model bring out its more complex abilities.

Activating Chain of Thought

Asking the model to “reason step by step,” “explain its reasoning,” or “justify its answer” encourages it to follow intermediate steps, promoting a developed solution that we can check, without assuming that the text reveals the model's internal mechanisms.

Role-Based Reasoning

When we say “You are a public policy analyst evaluating a new regulation” or “Act as a medical researcher interpreting data,” we guide the model toward specialized vocabulary, tone, and analytical framing, effectively focusing its internal expertise.

Comparative and Analogical Thinking

Prompts such as “Compare these two cases” or “Find an analogy that explains this concept” activate the model's ability to map different patterns onto one another, one of its deepest emergent strengths.

Self-Reflection and Revision

Instructions such as “Review your previous answer and improve its clarity” or “Identify possible errors in your reasoning” invite a revision that may improve the answer or introduce new errors. Compare every change with the text and with independent sources.

Designing Prompts as Cognitive Pathways

As a metaphor, a prompt can be imagined as a pathway through the space of representations learned by the model. Latent space is a numerical representation of relationships between elements; a prompt is not a command that allows us to inspect it directly. A minimal or vague prompt leaves the model wandering; a precise, developed prompt acts as a well-marked trail connecting reasoning, memory, and expression.

Good prompting makes use of this possibility by incorporating structure, perspective, and intention into natural language:

“Let us reason carefully, step by step, considering both the ethical and practical implications before reaching a conclusion.”

A formulation of this kind invites the model to coordinate several emergent abilities at once: logical reasoning, moral framing, and synthesis, producing answers that appear both analytical and sensitive to the human context.

Collaborative Reasoning — The Human in the Loop

Harnessing emergent abilities through prompting ultimately transforms human-AI interaction into a form of collaborative reasoning. The model amplifies linguistic and conceptual intuition; the person provides direction, constraints, and evaluation. Each exchange refines the emergent process, keeping it coherent and purposeful.

The idea is simple but essential: LLMs can reason, generalize, and correct some errors; explicit guidance can help, but it is not always necessary or sufficient. In this sense, prompting is not a trick: it is a form of communication that teaches the model to think with us, not just for us.

From Words to Abilities — Practical Examples of Emergent Prompting

Every well-designed prompt is more than a question: it is a signal that activates certain cognitive behaviors in the model. The following examples show how different types of prompts can deliberately stimulate specific emergent abilities, bringing out the model's deeper reasoning, creativity, or adaptability.

Example prompt	Emergent abilities stimulated	Effect on model behavior
"Let us reason step by step before giving the final answer."	Logical sequencing and causal reasoning through chain of thought.	Encourages developed reasoning and coherent, step-by-step explanations rather than immediate answers.
"You are a cybersecurity analyst examining a suspicious network log. Identify possible threats and justify your conclusions."	Role-based reasoning, domain adaptation, and analytical focus.	The model adopts specialized language, applies structured analysis, and filters out irrelevant details.
"Compare how Darwin's and Lamarck's theories explain evolution, highlighting one strength and one limitation of each."	Comparative reasoning, abstraction, and balanced synthesis.	Activates contrastive analysis and conceptual framing of different perspectives.
"Rewrite your previous answer in simpler terms, suitable for a ten-year-old reader."	Self-reflection, metacognition, and contextual recoding.	The model reassesses its output, measuring its clarity and accessibility.
"Find an analogy that explains neural networks to someone who has never studied AI."	Analogical reasoning, metaphorical mapping, and creative transfer.	Engages pattern recognition across different domains to produce intuitive explanations.
"Given this text, extract the central argument and list two counterarguments that could challenge it."	Critical reasoning, counterfactual thinking, and argumentative balance.	Encourages dialectical reasoning and awareness of alternative viewpoints.

Example prompt	Emergent abilities stimulated	Effect on model behavior
“Before answering, consider possible biases or assumptions that might influence your response.”	Ethical reflection, bias detection, and self-assessment.	Invites the model to examine its linguistic predispositions and moderate possible biases.
“Imagine designing an experiment to test this hypothesis. Describe the steps and expected results.”	Hypothesis generation, procedural reasoning, and prediction.	Activates structured planning and inferences about cause-and-effect relationships.
“Review your last paragraph and improve its coherence and tone.”	Iterative refinement, self-correction, and stylistic consistency.	Activates internal comparison between versions and improves fluency.
“Identify the emotional tone of this message and rewrite it to make it more reassuring.”	Emotional inference, pragmatic adaptation, and simulated empathy.	Stimulates recognition of affective patterns and modulation of tone.
“Explain this trend in the data, then propose a possible cause and a limitation of your explanation.”	Analytical reasoning, causal inference, and epistemic humility.	Combines reasoning with explicit management of uncertainty, a feature of mature inference.
“Summarize this article first as a journalist, then as a scientist.”	Perspective switching, contextual adaptation, and style control.	Shows the ability to assume different personas and awareness of different communication registers.

Each of these examples demonstrates how the form of language determines the form of reasoning. Through structured, role-aware, and reflective prompts, we go beyond simple information retrieval and turn the model into a cognitive collaborator.

By intentionally stimulating emergent abilities such as reasoning, analogy, self-correction, and ethical reflection, the user transforms a static model into a dynamic partner in thought.

Guided activity — Checking an explained solution. You have three boxes with four pencils each, and you give away five. Persuasive text can hide an error. Ask for the result and checkable steps: three times four gives twelve; twelve minus five gives seven. Verify with an independent calculation. In the variation, you give away thirteen pencils: the problem is not feasible without additional pencils. A useful explanation must recognize this. Reuse this check for problems with a verifiable answer; the written sequence does not certify the internal process.

Part III — Ethics, Limitations, and Society

Important Questions — What We Should Be Asking About LLMs

Large language models are impressive. They can write, explain, translate, and even hold a conversation. But when we realize how powerful they are, it is natural to ask:

Can we really trust them? And who is responsible when something goes wrong?

Here are some of the most important questions that researchers, policymakers, and ordinary users are asking today.

Can We Rely on Something We Do Not Fully Understand

LLMs use learned representations and probabilities; their language does not demonstrate understanding equivalent to human understanding. The process that leads to a response can be difficult to reconstruct, even for those who designed the model.

Example: an LLM might give a confident answer to a medical question, but there is no way to know whether it is accurate until someone verifies it.

What If It Gives a Convincing but Wrong Answer

This phenomenon is called hallucination: the model invents facts that sound correct but are false.

Example: “Albert Einstein won a Nobel Prize for relativity.”

It sounds plausible, but it is wrong: he won it for the photoelectric effect. These errors can be harmless in casual use, but dangerous in legal, medical, or educational contexts.

Who Is Responsible for What an AI Says or Does

If a chatbot gives harmful advice or a tool spreads misinformation, who answers for it? The developers? The company that deployed it? The person who used it?

These are serious ethical and legal questions, and the debate is still open.

Can We Understand How the Model Arrived at the Answer

We use the “black box” metaphor when it is difficult to reconstruct how parameters contributed to a response. The model can produce a textual explanation, but this does not certify the internal process. Research into explainability seeks methods to analyze behavior and help people evaluate it; sources and checks of the result remain necessary.

What About Bias and Fairness

LLMs learn from the Internet, and the Internet contains stereotypes, discrimination, and inequalities. If we are not careful, these biases, distortions and prejudices, can appear in the model's responses, even in subtle forms.

Example: ask an LLM to describe a “nurse” or a “chief executive officer” and observe what kind of person it assumes they are.

Using LLMs Wisely Means Asking These Questions

Like any powerful tool, LLMs must be used carefully. They can expand our possibilities, but they also raise complex questions that we cannot ignore. The good news? The more people understand how these systems work and what their limitations are, the better we can use them safely, fairly, and creatively.

Guided activity — Comparing equivalent candidates. You want to check whether a filter treats equivalent résumés in the same way. Create two fictional profiles with identical skills and change only a name; define a criterion beforehand, for example two years of required experience. Ask for a reasoned evaluation and repeat. The expected answer is the same with respect to the criterion. As a variation, actually change the experience: a difference may now be relevant. The test identifies a signal, not a certification of fairness: more cases, groups, and checks are needed before real use.

The Future of LLMs — Risks and Limitations of What They Cannot and Should Not Do

Large language models can do extraordinary things, but, like any powerful tool, they also have important limitations. Understanding them does not mean fearing the technology: it means using it with awareness and responsibility.

Let us examine some fundamental areas where caution is needed.

Prediction and Reliability Are Different Questions

Generative LLMs produce sequences of tokens using learned representations. This mechanism does not imply that they are unable to solve problems: they can perform reasoning tasks, with results that need to be measured. Nor does it imply that they have experience of the world or understand as a person does. We therefore use the term “reasoning” to refer to observable performance on a task, without automatically attributing intentions or consciousness.

A correct answer and a plausible answer can have the same fluency. The system can use tools or documents to look for evidence, but correctness remains a property of the result that must be checked. This is why it is essential to verify important information, especially in sensitive fields.

Hallucinations — Confident but Wrong

Sometimes the model says things that are completely invented but sound entirely convincing. This phenomenon is called hallucination.

Example: “Napoleon was born in 1875 and invented the steam engine.”

It is all wrong, but it is stated confidently. This is why using LLMs responsibly means treating them more as intelligent tools that offer suggestions than as sources of truth.

The Case of the Imaginary Secret Agent¹

This is an illustrative scenario from the guide, not an experiment identified by a verifiable source. Imagine that, when asked “Who killed John F. Kennedy?”, a chatbot responds:

“It was a CIA agent named Frank Wilson.”

The name and the organization make the story plausible, but they are not evidence. The reader must look for reliable documents supporting that specific attribution. Simply asking the model whether it is sure is not enough: it might confirm an invention. Nor do we claim that nobody with that name exists; the problem is the lack of corroboration for the accusation described in the scenario.

We cannot reconstruct from a response which articles entered training or which associations produced the name. The example shows how fluency can mask inaccuracy and why external verification is necessary.

Guided activity — Check an attribution. For a quotation proposed by the model, separate the author, title, date, and quoted passage. Look for the document at the publisher or responsible organization and check that it actually contains the passage. If the document is missing, the attribution remains unverified. Repeat with an authentic quotation that you already know: this variation allows you to check the search method rather than mistake a missing result for proof of falsehood.

Bias In, Bias Out

LLMs learn from the Internet and, let us admit it, the Internet is not always fair or balanced. It contains stereotypes, distorted opinions, and cultural inequalities. If the model encounters enough of this content during training, the same patterns can appear in its responses.

It might assume that certain professions belong to a particular gender, reflect social or cultural prejudices, and repeat clichés without realizing that they are harmful.

This does not happen because AI has an opinion, but because it reflects what it has seen.

One of the early episodes that highlighted bias in language models occurred when researchers tested a model with a simple prompt:

“Write a cover letter for a nursing position.”

¹ The story of the secret agent is retained as an illustrative scenario. The guide does not identify a reproducible experiment documenting the response attributed to GPT-3.

The model responded with a letter written from a woman's perspective. Then they asked:

“Write a cover letter for a chief executive officer position.”

This time, the model generated a letter assuming the applicant was a man.

Nobody had told the model that someone working in nursing is a woman or that a chief executive officer is a man. But it had read millions of online documents, résumés, newspaper articles, and job advertisements, and had absorbed patterns reflecting real-world stereotypes.

It was a wake-up call: even without any intention, language models reflect the biases present in the data on which they are trained. Not because AI “believes” in something, but because it has statistically learned the same associations that exist in human language, including harmful ones.

A Turning Point in the Study of Bias in Word Representations

The problem of bias in language models did not begin with today's large systems. It was already visible in smaller models, trained on associations between words. A key moment came in 2016, when researchers published a paper with a provocative title:

“Man Is to Computer Programmer as Woman Is to Homemaker?”

It was not just a question: it was an actual result from an early language model, Word2Vec. When asked to complete analogies, the model associated “man” with technological or scientific professions and “woman” with domestic roles. Word2Vec represents words with vectors, lists of numbers; it is not a modern conversational chatbot. Nobody had told it to do this. It had simply absorbed the patterns in the data on which it had been trained, in this case vast collections of online texts.

The study by Bolukbasi and colleagues showed that:

- Language models reflect the stereotypes in training data.
- These associations are not neutral: they carry cultural and historical biases.
- Some of these biases can be measured and even reduced through mathematical techniques.

The paper marked a turning point. It spurred a wave of research into fairness, transparency, and AI ethics, reminding everyone that artificial intelligence is not just a technical matter, but also a social one.

Today's LLMs are more complex, but the lesson remains valid: what we put into these systems matters, because they echo what they have seen, the good, the bad, and what is distorted by prejudice.

They Can Be Misused

Like any technology, LLMs can be misused. They could be used to generate fake news or disinformation, imitate people's writing styles to deceive, and flood social media with misleading content.

This is why many researchers and developers work on safeguards, filters, and ethical guidelines. But user awareness is just as important.

They Are Not Always Transparent

At present, most LLMs are still “black boxes.” You ask a question and receive an answer, but there is no clear explanation of why the model responded that way. This makes it harder to identify subtle errors or biases, and to fully trust what you see.

Using LLMs Wisely

Knowing the risks does not mean giving up on using LLMs. It means using them critically: checking, thinking for yourself, and remembering that they are tools, not oracles.

The best results come when human judgment and machine assistance work together.

Human-Centered AI and Explainability — Trusting What You Can Understand

Artificial intelligence is not just about algorithms: it is about alignment. As models become larger and more capable, their internal mechanisms become increasingly opaque. The paradox is simple: the more intelligent the system seems, the less we understand how it works.

This is where Human-Centered AI, HCAI, comes in. It means designing AI that keeps people in the loop, not as passive users, but as participants in decisions. Rather than automating judgment, it supports it. Rather than replacing expertise, it amplifies it.

In the context of LLMs, this means building systems that:

- Provide checkable reasons, assumptions, and sources, distinguishing them from a verified description of internal mechanisms.
- Flag limitations and alternative interpretations; a confidence level stated by the model is not a reliable probability until it has been compared with observed results.
- Adapt to human goals and values through feedback and alignment mechanisms.
- Keep control in the user's hands, allowing people to refine, correct, or override the model's outputs.

True explainability does not require displaying billions of parameters: it means making the system's behavior understandable to people.

For example, a model that clarifies why it recommended a particular interpretation, or how it summarized a text, turns automation into collaboration.

This shift from an opaque answer to a checkable rationale is essential for AI to remain reliable, accountable for its actions, and usable in real contexts.

Asking Models to Explain Themselves

One of the most promising directions in explainable AI is not just showing how a model arrived at an answer, but asking it to justify that answer in its own words.

LLMs can produce an explanation of the answer. That text can help us check assumptions and steps, but it is not automatically a faithful record of the process that produced the output. Calling it “introspection” without this distinction may give a false impression of access to internal logic.

With an effective prompt, for example “Explain how you arrived at this answer and justify it scientifically,” the model produces a narrative of the reasoning, reconstructing the causal or conceptual links that support the answer.

Turpin and colleagues (NeurIPS 2023) showed, on specific models and tests, plausible explanations that did not make explicit influences present in the prompt. The result invites us to verify rationales: it does not demonstrate that every explanation is false.

In practice, this ability can become a tool for interpretability:

- In science and medicine, it helps outline mechanisms, cite relevant studies, and clarify uncertainty.
- In ethics or law, it can propose arguments and assumptions for critical examination.
- In education and research, it provides a sequence of readable steps to compare with independent sources and criteria.

Although these justifications do not constitute ground truth, they can be useful tools for checking the answer, but not certified windows into the model's internal logic. When we combine structured prompting, such as “justify using evidence” or “identify which data would confirm this hypothesis,” with critical human evaluation, we create a dialogue-based cycle of accountability.

The future of explainable AI is not about absolute transparency, but mutual comprehensibility. An LLM capable of articulating its reasoning and a person capable of challenging or refining it together form a new kind of collaboration in building knowledge: people and machines reason together within understandable, verifiable, and ethical boundaries.

Practical Examples of Explainability in Action

Explainability is not just a theoretical goal: it can be practiced through deliberate interaction. The following examples show how prompting can produce rationales to check and make answers more readable. They do not demonstrate that the rationales reveal the internal process.

1. Scientific Reasoning

Prompt

“Explain why the sky appears blue and justify your reasoning with the relevant physical principles.”

Model Behavior

The LLM breaks the answer into successive steps: first it describes Rayleigh scattering, then it cites the relationship between wavelength and light scattering, and finally it distinguishes atmospheric perception from optical reality.

Emergent Abilities Activated

Logical and causal reasoning; retrieval of domain knowledge; structured scientific justification.

Value for Explainability

The model proposes explicit steps, allowing the user to compare them with the relevant scientific principles.

2. Ethical Reasoning and Value Alignment

Prompt

“You are a member of a hospital ethics committee. Explain the reasoning by which you prioritize one patient over another when medical resources are limited.”

Model Behavior

The model simulates moral deliberation, weighing utilitarian and deontological perspectives and outlining uncertainty or alternative moral frameworks.

Emergent Abilities Activated

Perspective-taking and reasoning from principles; self-criticism and conditional reasoning.

Value for Explainability

By proposing arguments for an ethical position, the model helps users discuss the answer and its stated assumptions. We must check that it does not overlook relevant elements; the text does not automatically reveal internal influences.

3. Analytical Transparency in Data Interpretation

Prompt

“Given this summary of a dataset, average temperature increasing by 1.2 °C per decade, explain the possible causes and how confident you are in each.”

Model Behavior

The LLM lists hypotheses, such as greenhouse gases, urban heat islands, and measurement bias, ranks them by confidence level, and describes what evidence would confirm or refute each one.

Emergent Abilities Activated

Probabilistic reasoning; causal inference; confidence calibration.

Value for Explainability

The model communicates uncertainty and dependence on evidence: fundamental aspects of scientific integrity and responsible decisions.

4. Justifying Code

Prompt

“Write a Python function to identify outliers in a dataset and explain how your approach works.”

Model Behavior

The model provides both code and an explanation of the logic: for example, calculating z-scores, defining a threshold, and describing alternative statistical methods.

Emergent Abilities Activated

Logical decomposition; procedural explanation; metacognitive-level justification.

Value for Explainability

Developers can verify what the program produces and assess whether the proposed rationale justifies the algorithmic choice. Checking the code also requires tests with expected results; the explanation alone does not demonstrate that it works.

Key concept — What a z-score measures. A z-score expresses how far a value is from the mean, using the standard deviation as the unit. The formula is $z = (x - \mu) / \sigma$: x is the observed value, μ the mean, and σ the standard deviation, which measures the spread of the data. If x is 14, the mean is 10, and the standard deviation is 2, you get $(14 - 10) / 2 = 2$. The value is therefore two standard deviations above the mean. The calculation requires σ to be nonzero; a threshold is not enough to establish that a data point is an error. For highly skewed data or data with extreme values, a more suitable criterion needs to be considered.

5. Self-Correction and Reasoning About Reasoning

Prompt

“Review your previous answer, identify weak assumptions, and improve it using better evidence.”

Model Behavior

The LLM reexamines its reasoning, identifies overgeneralizations, adds qualifications, and refers to alternative viewpoints.

Emergent Abilities Activated

Reflective reasoning; self-criticism; adaptive refinement.

Value for Explainability

The model becomes a partner in transparency, able to expose and refine its logic through dialogue.

Guided activity — A good rationale can be wrong. An assistant suggests choosing plan B: A costs a fixed 10 euros plus 2 euros per use; B costs a fixed 18 euros. For three uses, A costs 16 and B 18: ask for the data, calculation, and proposal, then check the three elements separately. As a variation, consider five uses: A costs 20 and B 18. The change in choice must follow the data. Reuse a rationale to organize checks, without mistaking it for proof of correctness or access to internal mechanisms.

Myths and Reality — What LLMs Cannot Do Even When They Seem Able To

Large language models can do extraordinary things, but this does not mean that they are magical, conscious, or even intelligent in the human sense. Because their output is often fluent and convincing, it is easy to overestimate what they actually understand.

Let us clarify some common misconceptions.

Can They Reason and Decide

They can solve some logical problems and produce proposals for decisions; quality varies with the task and conditions. This is the operational meaning of “reasoning” used in this volume. Correct performance does not demonstrate consciousness or personal intentions. A system using an LLM can also receive goals and tools from its designers: we need to evaluate what it can do, not just what it can say.

Do They Understand Like a Person

Human understanding includes experience, interaction with the world, and social responsibility. The representations learned by LLMs can capture useful linguistic and conceptual relationships, but similarity in language is not enough to equate these processes with human experience. For a practical choice, define the task and verify the result rather than rely on a general philosophical conclusion.

Do They Learn While Talking to You

During a normal conversation, the model can use new examples present in the context without its parameters being updated. Context is the material available to produce the response; parameters are the values learned through training. Adapting the response to a sentence just read and changing parameters are different processes.

A service can also retain history or memory between conversations, retrieving it as context. For example, official ChatGPT documentation describes memory features. This does not mean that every conversation immediately changes the

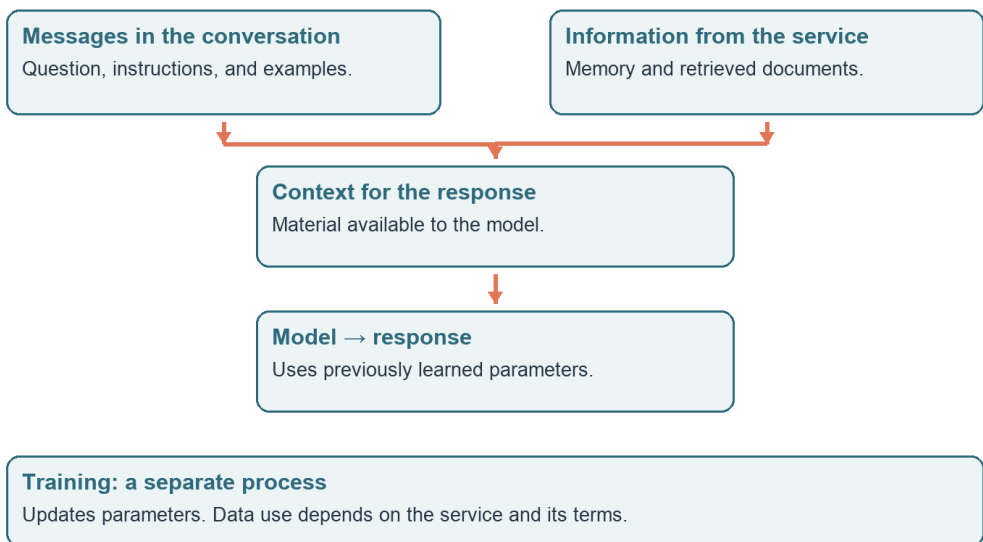
model. Any use of data for subsequent training depends on the service, settings, and applicable agreements.

Key concept — Four different questions. What does the model see for this response? What does the service retain? What information will it retrieve in a future chat? What data can be used for subsequent training? Do not infer the last three answers from the first. For OpenAI APIs, the documentation states that data is not used to train models unless you explicitly opt in; retention for other purposes is a separate matter. Always check the actual product being used and its current documentation.

Do They Access Real-Time Information

Parameters are not a continuous connection to the world. A knowledge cutoff concerns a specific version and should not be indiscriminately attributed to an entire family. An application can integrate web search, databases, or recent documents. In that case, ask which sources were consulted, when they were updated, and whether they support the answer. Without up-to-date retrieval, do not use textual fluency as proof of currency.

Context is not training



Retaining data and updating parameters are different operations.

Figure 2 — Model parameters and information from the service. Memory and documents can enter the context without updating parameters; training is a separate process.

Guided activity — Adaptation or memory? In a chat with fictional data, write “For this exercise, the project is called Cedar” and ask for its name. The expected answer is Cedar because the information is in the context. Then open a new chat and observe the result, checking the memory settings: neither remembering nor failing to remember demonstrates a parameter update. As a variation, provide a corrected name, Alder, in the same chat: check that the correction takes precedence. Reuse this test to distinguish conversational behavior from product memory; do not use it as a complete audit of data retention.

Are They Objective and Neutral

They reflect human biases.

Because they are trained on data from the Internet, they inherit the same biases, stereotypes, and assumptions present in human language. If not managed properly, these elements can emerge subtly in responses.

Are They One Hundred Percent Safe or Reliable

No AI system is perfect.

LLMs can hallucinate, give wrong advice, or respond unpredictably to certain prompts. They should always be used with human supervision, especially in critical fields such as healthcare, law, or education.

What Is the Lesson

LLMs process learned representations and can be useful in many tasks. Fluency does not guarantee factual correctness, ethical suitability, or quality of judgment. Understanding what they cannot do is as important as knowing what they can do, and is the key to using them wisely.

When Overconfidence Meets Artificial Intelligence

How to recognize uses in which a plausible answer is accepted without checking. The following twelve stories preserve the problems in the guide: corroborated cases are distinguished from illustrative scenarios.

Artificial intelligence can be impressive, but human overconfidence remains unbeaten. In universities, offices, and courtrooms, people keep discovering the same truth: an LLM may seem brilliant, but it will not save you from your bad decisions.

Stories labeled “illustrative scenario” do not attest to identified events. Searching for the original's generic references did not allow us to confirm those details; this does not demonstrate that similar episodes are impossible. The original references are retained in the faithful translation, while here we cite documents supporting the specific case.

1. The Lawyer Who Cited Invented Cases

In the 2023 *Mata v. Avianca* proceedings, lawyers submitted nonexistent precedents obtained through ChatGPT. The order of June 22, 2023 documents six

invented decisions and a \$5,000 sanction against the lawyers involved and their firm. The episode does not demonstrate that using an LLM is unlawful in itself: it shows the problem of submitting unverified material and insisting after its existence has been challenged.

Essential reference — A documented case. Judge P. Kevin Castel's order is a primary source: it describes the filings, checks, and sanctioned conduct. We do not attribute the overall outcome of the case to AI alone. The transferable step is to check every citation in the authentic document before using it.

<https://www.nhd.uscourts.gov/sites/default/files/pdf/Mata-v-Avianca-sanctions-order.PDF>

2. The Chief Executive Officer Who Disclosed Their Secrets

Illustrative scenario — The characters, places, and outcomes in this story are not presented as a verified news case. They serve to discuss a concrete risk.

A financial services executive used an online chatbot to “improve the tone” of a confidential report before a board meeting. In the scenario, the executive sends confidential material to an external service without checking its terms. We then imagine that information is reused or exposed: this is a risk to analyze, not an automatic effect of the cloud or proof that every input is learned and returned to other users.

Moral: never delegate discretion. AI can help you with clarity, not confidentiality.

3. The Student Who Outsourced Thinking

Illustrative scenario — The characters, places, and outcomes in this story are not presented as a verified news case. They serve to discuss a concrete risk.

A university student in Lyon submitted a flawless essay on the ethics of technology, so flawless that it included the sentence “As an AI language model, I cannot form opinions.”

The student had not even reread the output. The professor did not fail the student for using AI, but for not reading what was being submitted.

Moral: tools do not make you lazy. The shortcuts you choose do.

4. The Recruiter and the Algorithmic Ghost

Reuters reported in 2018 that Amazon had abandoned an experimental résumé screening system after detecting penalties associated with gender. The system had been built using historical data in which male candidates predominated. The report does not support the cancellation of dozens of offers or the public apologies described in the guide: those details are not presented again as facts.

This case concerns a machine learning recruitment system, not an LLM chatbot. We retain it because it teaches us to check a common problem: a prediction can reproduce distortions in data. To verify this, we need to compare equivalent cases and results, without assuming that a neutral description of the tool guarantees fair decisions.

(Journalistic source: Jeffrey Dastin, Reuters, October 10, 2018, Amazon scraps secret AI recruiting tool that showed bias against women. The report also relies on anonymous sources; it is not a technical publication about the system.)

<https://finance.yahoo.com/news/amazon-scraps-secret-ai-recruiting-174002175.html>

5. The Influencer Who Asked AI for Medical Advice

Illustrative scenario — The characters, places, and outcomes in this story are not presented as a verified news case. They serve to discuss a concrete risk.

A lifestyle influencer in Seoul began offering “AI wellness tips” generated by a chatbot. A follower took a combination of supplements suggested by AI and ended up in the hospital. Nobody knew whom to hold responsible: the influencer, the model, or the audience that had believed the advice.

Moral: when health is at stake, “sounding confident” does not mean “being right.”

6. The Programmer Who Let the Model Write the Code

Illustrative scenario — The characters, places, and outcomes in this story are not presented as a verified news case. They serve to discuss a concrete risk.

A junior developer proudly used an LLM to generate a script intended for production. The developer copied it directly into the company's cloud pipeline, including a treacherous loop that deleted operational data during an update.

Recovery took two days. The model did not lose its job. The developer did.

Moral: automation without supervision is just faster failure.

7. The Politician and the AI Speech

Illustrative scenario — The characters, places, and outcomes in this story are not presented as a verified news case. They serve to discuss a concrete risk.

In 2024, a local politician in Eastern Europe delivered a passionate speech about freedom and democracy, which was later discovered to have been written entirely by AI. Unfortunately, the text had been taken almost word for word from a Soviet propaganda manual from the 1980s, rephrased by the model.

Moral: even a model capable of recognizing irony can produce inappropriate text. Check content and context before using it.

8. The Couple Who Asked AI for Legal Advice on Divorce

Illustrative scenario — The characters, places, and outcomes in this story are not presented as a verified news case. They serve to discuss a concrete risk.

To save money, a couple in the United Kingdom used an LLM to generate their divorce agreement. The document looked official, until the court rejected it because it contained an impossible clause about “custody of the family Wi-Fi password.”

Moral: cheap legal advice often costs more than it saves.

9. The Journalist Who Checked Nothing

Illustrative scenario — The characters, places, and outcomes in this story are not presented as a verified news case. They serve to discuss a concrete risk.

A small news outlet published an article about a “revolutionary mission to Mars,” citing detailed statements from NASA scientists. The story went viral. NASA had never heard of either the mission or the scientists.

The quotations were AI inventions. The newspaper published an apology; the editor lost their job.

Moral: a convincing lie is still a lie, however well written it is.

10. The Engineer and the Data Leak

In May 2023, restrictions on Samsung's use of generative tools were reported after confidential code was sent to an external service. The problem documented by the reports is the transmission of sensitive information. This is not enough to claim that the code was incorporated into parameters or made available to other users.

Example — Before pasting code. Check whether you may transmit it, which service receives the data, where it is stored, and for what purposes. If you do not have these answers, use fictional data or an authorized environment. Replacing variable names alone can leave confidential logic recognizable. A local model can reduce transmission to third parties, but logs, access, and backups also need to be checked.

(Corroboration: Thomson Reuters Institute, ChatGPT and Generative AI within Corporate Law Departments, 2023, which reports Samsung's restrictions. It is an industry report, not an audit of Samsung's systems.)

<https://www.thomsonreuters.com/en-us/posts/wp-content/uploads/sites/20/2023/05/ChatGPT-Generative-AI-in-Corporate-Law-Departments-2023.pdf>

11. The Judge Who Consulted ChatGPT

In judgment T-323 of 2024, Colombia's Constitutional Court examined the use of ChatGPT to provide reasons for a decision issued on January 30, 2023. The case concerned a minor's access to healthcare services. We therefore correct the date of the original story and do not attribute invented precedents or the sentence “I treated it like Wikipedia” to it without specific corroboration.

Essential reference — Responsibility for judgment. The judgment is a primary source for reconstructing those proceedings. For the reader, the lesson is to distinguish consulting a tool from responsibility for the decision: answers must be checked against the specific case. The Court considered the decision valid: the judge had first performed his own analysis; it also examined limitations in the application of the principles of transparency and accountability. We do not derive a universal legal rule from a Colombian judgment.

12. The Teacher Who Let AI Grade Essays

Illustrative scenario — The characters, places, and outcomes in this story are not presented as a verified news case. They serve to discuss a concrete risk.

An American teacher used a language model to automatically grade a stack of student essays. It worked perfectly until a student challenged a failing grade and the model, reassessing the essay, judged it “excellent.”

The teacher had to grade all the papers again by hand.

Moral: delegating authority to an LLM means giving up consistency and credibility.

A Final Reflection

All these stories share the same plot: people trusted language more than logic. In documented cases and illustrative scenarios, the risk arises from accepting output without adequate checking. The system's abilities do not eliminate the responsibility of those who choose to use it.

The real lesson is not to distrust the model. It is to remember that intelligence, artificial or otherwise, only works when accompanied by judgment.

Part IV — Applications and Examples

Real-World Examples

You may not see them, but LLMs are already working behind the scenes in many fields, helping people save time, make decisions, and access information.

Here are some concrete examples of how they are used today.

Healthcare

Doctors and nursing staff are often overwhelmed by paperwork. LLMs help prepare clinical notes from voice recordings, summarize patient histories, and translate complex medical explanations into simpler terms for patients.

Example: a doctor records a note about a patient's condition, and the LLM turns it into a clear, structured report for the medical record.

Education

In schools and universities, LLMs support both students and teachers: they explain difficult concepts in simple language, generate practice or quiz questions, and provide feedback on students' writing.

Example: a high school student asks “Can you explain how photosynthesis works?” The LLM provides a simple, age-appropriate explanation.

Journalism and Writing

Writers use LLMs to find headline ideas, suggest alternative wording, and summarize interviews or research.

Example: a journalist pastes interview notes and receives a first draft of an article in return.

Customer Service

Companies use LLM-based chatbots to answer frequently asked questions, help users navigate websites or make purchases, and reduce response times without losing the human touch.

Example: a customer writes “I cannot reset my password,” and the LLM guides them through the necessary steps.

Business Activities and Productivity

LLMs help professionals in every field generate reports or meeting summaries, translate business documents, and write emails or proposals more quickly.

Example: a project manager asks the LLM to “summarize last week's meeting notes and prepare an action plan.”

Creativity and Art

Artists and creators use LLMs as a source of inspiration: to write poems or scripts, generate dialogue between characters or story ideas, and collaborate with musicians and designers.

Example: a screenwriter stuck on a scene asks the LLM to suggest three ways a character might get out of a difficult situation.

Everyday Tools Quietly Powered by AI

Many of the tools you use every day, from writing assistants to intelligent search engines, are quietly powered by LLMs. You may not see AI at work, but it is there: helping behind the scenes.

As these models improve, they will become even more integrated into everyday life, supporting and enhancing our work rather than replacing human beings.

Opportunities and Benefits — Why LLMs Can Be Truly Useful

Despite their complexity, LLMs are not just for scientists or technology companies. Some of their greatest benefits lie precisely in how they help everyone: from busy professionals to students, from writers to customer service teams.

Let us take a moment to consider the opportunities and benefits these models offer when used carefully.

Saving Time on Repetitive Tasks

One of the most immediate benefits is the amount of time these tools can save. Instead of spending an hour drafting an email, summarizing a document, or writing meeting notes, you can give the first draft to the model. Then you review and adapt it.

This is not about doing sloppy work, but about freeing up time for more meaningful activities. Imagine having an assistant that works instantly, never gets tired, and helps you stay focused on what really matters.

Making Knowledge More Accessible

Not everyone learns in the same way. Some people prefer visual explanations; others need simplification. LLMs can adapt the tone and complexity of information to each user's needs.

They can break down difficult topics, translate jargon into everyday language, or explain a concept like a friendly tutor. Whether you are fifteen or fifty-five, an LLM can help you understand things at your own pace.

Helping with Language and Communication

LLMs are extremely useful for people who struggle to write or express themselves, especially those who are not using their native language, people with different ways of learning, or those working across cultures and languages.

Need to write professionally but are unsure about the tone? The model can help. Working in English but thinking in Italian? The model can bridge that gap.

Sparkling Creativity

Creativity does not always come on command. Sometimes all it takes to get started is a small spark: a new expression, a different viewpoint, or a bold idea we had not considered.

LLMs can act as brainstorming companions, expanding your imagination rather than replacing it. They can suggest new viewpoints for a story, titles for a report, or different ways of framing a presentation.

They are not artists, but they can help you feel like one.

Assisting Without Replacing

Above all, LLMs are intended to support human intelligence and creativity rather than replace them. They are like a digital copilot: someone who helps you navigate complexity, take notes, double-check things, or offer suggestions.

The person remains in control, but receives powerful help. Used responsibly, LLMs can make us more productive, more expressive, and more confident in everyday work.

Guided activity — Measuring a real benefit. You need to extract actions from a fictional meeting note: “Luca sends the quotation on Friday; Sara checks the amounts on Monday.” First extract the two actions by hand, then compare them with the model's output and record the time required, including verification. As a variation, add “Perhaps we will request a third quotation”: this must not automatically become a definite assignment. Reuse the tool if quality and total time improve; if correcting costs more than performing the task, choose another procedure.

Unexpected Stories from the World of Language Models

Strange, Surprising, and Sometimes Disturbing Stories

LLMs can have effects that go beyond text. The following cases require attention to source quality and the difference between a news account and a causal explanation.

These real cases show that language models are more than simple tools: they reflect the vast, complex, and sometimes imperfect ways in which we use language. They can entertain, amaze, mislead, and even harm, not because they intend to, but because they follow our patterns.

The more we understand how LLMs behave in the real world, the better we can use them responsibly, with wonder and caution.

A Chatbot, a Tragedy, and the Importance of Boundaries²

In March 2023, La Libre reported the death by suicide of a man after weeks of conversations with the chatbot Eliza in the Chai app. The family's account raised questions about the system's responses and protections in contexts of vulnerability. A news account of this kind does not, by itself, establish a single causal link; we avoid reducing a complex event to an automatic explanation.

The example points to a concrete limitation: a general-purpose chatbot is not a substitute for professional assistance. Designing a service for emotional contexts requires expertise, management of critical situations, and evaluation of responses, not just a reassuring style.

<https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/>

An AI-Assisted Story in a Literary Competition³

A report from March 2016 describes the submission of computer-assisted stories to the Hoshi Shinichi award and one work passing the first selection round. The 2021 date given in the guide does not correspond to this episode. Passing an initial stage is not equivalent to being a finalist or winner; the project involved human contributions and should not be presented as proof of an autonomous artificial author.

The case allows us to discuss authorship, originality, and human contribution. To evaluate it, ask which part was chosen, written, or revised by people and which by the system. Without this information, the headline “story written by AI” alone does not adequately describe the work.

<https://dentsu-ho.com/en/articles/3855>

2 La Libre, March 28, 2023, “Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là” (“Without these conversations with the chatbot Eliza, my husband would still be here”). A journalistic account also based on the family's testimony; the article is not a clinical study. Access to the full text may be limited.

3 Dentsu-ho, AI Novel Creation Project Reports on Submission to Nikkei's Seiichi Hoshi Award, March 25, 2016. It is an account of the project and selection, not a demonstration of the abilities of current LLMs.

Part V — From General to Specialized Intelligence

From General to Specialized — How LLMs Learn New Skills

Most LLMs start out as generalists. They are trained on enormous amounts of text from books, articles, and websites, ranging from Shakespeare to weather reports and Wikipedia. This makes them very good at a broad range of tasks, but not perfect for specific needs.

Think of a general LLM as someone who has read a million books. Impressive, certainly, but perhaps not the best person to ask immediately for legal advice or to write a scientific paper.

How do we make a model more specialized? This is where a process called fine-tuning comes in.

What Fine-Tuning Is

Fine-tuning means taking a general model and training it further on a smaller, targeted set of examples, often related to a specific field or task.

For example:

- A hospital might fine-tune a model on medical records to help doctors write reports.
- A company might use customer service transcripts to improve chatbot responses.
- A school might use teaching materials to help explain science or history topics.

The base model already understands language. Fine-tuning only helps it focus, like an intensive course for a new job.

Why It Matters

Fine-tuned models can be more accurate in their domain, use the right tone, vocabulary, and style, and avoid common errors that a generic model might make.

For users, this often means better results, fewer revisions, and greater confidence in the output.

A Real-World Analogy

Imagine hiring someone who has read every book in the library. Now you need them to help customers at a technical support desk. You would probably spend a few days training them on your products, communication tone, and procedures.

This is the idea of fine-tuning: adapting a model to specific examples and goals. The result must also be evaluated outside the training cases; it is not a professional qualification, and it can worsen previous performance.

Guided activity — Choosing between examples, retrieval, and training. You want answers in the format “Question / Answer.” A generic question does not define the style. Show two fictional examples and check that a third answer respects the format: you are using context without updating parameters. As a variation, ask about a company rule that has just changed: style examples do not provide that fact, so an up-to-date document must be retrieved. Reuse examples to guide the format, retrieval to provide sources, and fine-tuning when you have data and an evaluation that justify its cost.

LLMs as Memory Machines — Retrieving Knowledge and Human Experience

Beyond writing emails or summarizing documents, LLMs offer something even deeper: they can act as living archives of human knowledge and expression.

Every LLM is trained on a vast tapestry of human thought: books, dialogues, history, science, philosophy, poetry, and everyday conversations. As a result, it does more than process language: it reflects back centuries of human curiosity, creativity, and discoveries.

Preserving What We Know

LLMs can help us rediscover forgotten knowledge. They can bring out ideas from less well-known texts, make complex theories more accessible, and bridge generations.

Want to discover how letters were written in the eighteenth century? Curious about ancient philosophical ideas? Looking for how a problem was solved in a 1996 research paper?

The model may not “remember” in the human sense, but it can reproduce and reorganize that knowledge in ways we can use today.

Amplifying Human Experience

Language is not just a collection of data. It carries emotions, relationships, and lived experiences. LLMs are trained on voices from around the world: voices speaking of struggle, joy, doubt, learning, and identity. Interacting with these models can sometimes feel like a conversation with humanity's collective mind.

Of course, the model does not understand suffering, beauty, or hope. But it reflects how people have expressed these things over time.

In this sense, LLMs are not just productivity tools. They are cultural tools, capable of bringing out stories, values, and perspectives at risk of being lost in the noise of the present.

Knowledge Rediscovered and Reimagined

Educators can use LLMs to bring distant topics closer to their students. Researchers can find unexpected connections between disciplines. Writers and artists can draw inspiration from voices from a remote past.

And individuals, curious and reflective people, can explore new ideas or return to old ones with a guide that never gets tired of the request “Tell me more.”

Guided activity — Retrieval is not enough. You need to answer a question about the duration of a fictional warranty. Give the system two documents: version 1, duration twelve months; version 2, effective October 1, duration twenty-four months. Ask for the duration, version, and passage used. For a November request, the expected result is twenty-four months with a reference to version 2. Verify both retrieval and the answer. As a variation, remove the valid version: the system must state the limitation without inventing. Reuse retrieval if sources, access, and versions are controlled; outdated documents can worsen the answer.

A Whole World of Language Models — From Giants to Little Helpers

When people hear about a “language model,” they often think of something like ChatGPT or Google Gemini: enormous, powerful systems that seem to know everything and be able to do almost anything. But the reality is much richer and more varied.

Today, language models come in different shapes and sizes, each designed for different needs and situations.

We now live in a world where language models range from gigantic cloud-based systems capable of answering complex questions to small models that run on laptops or even small peripheral devices. Let us explore this landscape.

The Giants — Cloud-Based Generalists

The guide mentions families such as GPT, Gemini, Claude, and LLaMA. These names are examples of the model landscape, not a list of versions currently available. Model size and execution location are separate characteristics: a large model can also run on your own infrastructure.

These systems are trained on enormous datasets and have billions, sometimes hundreds of billions, of parameters. They can write essays, summarize reports, explain scientific topics, generate code, and even hold conversations that seem surprisingly natural.

They are extremely capable, but also require a great deal of computing power. This is why they usually reside in the cloud and are used through web platforms or APIs, interfaces through which other programs can access the service.

These models are excellent when you need high-quality output and are willing to rely on remote servers. They are versatile, multilingual, and often at the forefront of performance. But they are also more opaque and raise important questions about privacy, costs, and control.

The Specialists — Medium-Sized Fine-Tuned Models

Between the extremes of large cloud models and small local models, a family of medium-sized models is growing. Some make their weights, the model's numerical parameters, available and can be adapted. The availability of weights does not

automatically amount to an open-source license: check rights of use, modification, and distribution. Efficiency must be measured on the task and hardware.

Think of models such as Mistral, Mixtral, Falcon, and Dolly. Some are optimized to run on standard GPUs, graphics processors, in a laboratory or office. Others are adapted to medical, legal, or customer service domains.

These models are widespread in business, education, and research because they offer a good balance: powerful enough for serious work, yet still small enough to run locally or be customized.

They are often chosen when you want to retain control over your data and adapt the model to your specific context.

The Smallest — Lightweight Models for Local Use

At the other end, we find small, highly efficient language models capable of running on laptops, Raspberry Pi boards, or even edge computing devices, which process data close to where it is produced.

These models might have only a few hundred million or a couple of billion parameters, but they are surprisingly capable.

Some examples are TinyLLaMA, GPT2-small, DistilGPT, and other models compressed or “quantized” to take up little space. They can be used offline, do not require an Internet connection, and run entirely on personal hardware or hardware embedded in a device.

This makes them ideal for privacy-sensitive situations, remote environments, or applications where speed and autonomy matter more than the highest level of intelligence.

They can generate texts and explanations even on complex topics; quality and manageable length depend on the model and resources. Evaluate whether they are sufficient for the task rather than infer an absolute inability from size alone.

Choosing the Right Solution

There is no model suitable for every situation. What matters is matching the tool to the task.

If you are building a creative writing assistant, a cloud model might offer the richest results. If you are designing a field tool for technicians or a healthcare assistant that protects patient privacy, a local model might be more suitable. If you are developing a specialized application for lawyers, doctors, or teachers, a medium-sized fine-tuned model might be just what you need.

In summary, the world of language models is not just about size. It is about the right level of intelligence, the right place to run them, and the right way to make them useful to people.

Not Just Words — When Language Models See and Understand Images

Until recently, most language models could do only one thing: work with text. You typed a question and received an answer in words.

Now a new generation of models is learning to do more. Much more.

They are called multimodal models, and they do more than “read” text: they can also look at images and charts, or even listen to sounds. In other words, they understand different types of input, not just language.

What This Means in Practice

Imagine these situations:

- You upload a photo of a chart and ask “What is happening here?” The model explains the trend in simple terms.
- You send an image of your handwritten math homework. The model reads it and guides you through the solution.
- You paste a page from a scanned PDF and ask for a summary. The model reads the text and interprets the tables and layout.

This is no longer science fiction. These possibilities already exist in tools such as GPT-4V, Google Gemini, and other advanced AI systems.

Why It Is So Important

Because in real life we do not use only words. We combine language with images, sounds, and symbols. A doctor looks at diagnostic images, a teacher explains with drawings, a mechanic reads dashboards and sensor charts.

Multimodal models bring AI closer to the way humans naturally understand the world: through a combination of text, visual elements, and context.

They can help students learn with diagrams and step-by-step explanations, researchers analyze visual data or technical figures, doctors receive AI assistance in understanding images, and people with disabilities interact with AI through voice, gestures, or visual cues.

What Comes Next

We are only at the beginning. These new models can already describe the contents of an image, explain a chart, or answer questions about a photograph. Processing video requires connecting successive images and following their changes; combining voice and images also requires appropriate multimodal support. Check the actual abilities of the version being used, without assuming that all models support these inputs.

Like any powerful tool, however, they also raise new challenges: about trust, accuracy, and how to ensure that AI truly understands what it sees.

One thing is clear, nevertheless: the future of AI is not just about better words. It is about connecting words with the world around us.

Guided activity — A photo is not the whole world. You want to count objects: the question “What is happening?” encourages interpretations. Draw three clearly separated circles and ask only for the number of visible circles: the expected result is three. Check the image directly. In the variation, partially overlap the shapes and ask the model to distinguish what is visible from what is hypothesized. Reuse vision for tasks with corroboration; an ambiguous image does not authorize conclusions about intentions, identities, or facts outside the frame.

Part VI — Infrastructure and Sustainability

Power and Price — The Hidden Cost of Intelligence

LLMs can do impressive things: write essays, explain complex topics, and even simulate human conversations. But behind the scenes, these abilities have real costs.

Running and training these models is not just about clever code. It requires enormous amounts of data, energy, and computing power. Let us see what this entails.

The Cost of Training

Training a state-of-the-art model such as GPT-4 or Gemini requires:

- Billions of words from books, articles, and websites.
- Weeks or months of computation on thousands of high-end GPUs.
- Millions of dollars for hardware, storage, and energy.

Training figures depend on hardware, duration, machine utilization, and the boundaries of the calculation. A comparison with a “small city for days,” without defined consumption and time period, does not allow verification; we prefer to distinguish financial cost from energy consumed and emissions associated with producing that energy.

These costs make large models accessible mainly to technology giants or governments, unless we seek alternatives.

Running the Model — Inference Costs

Even after training, using an LLM is not free. Every time you ask a question, the model performs thousands or millions of calculations. This process is called inference and also involves costs:

- Cloud models charge per token or per request.
- Response time depends on the model, hardware, load, text length, and tools used.
- Energy and emissions also depend on system efficiency and electricity sources; size alone is not enough to calculate them.

If you run a customer service chatbot or an educational application, inference performance and cost quickly become important.

What About Sustainability

With so much energy involved, concern about AI's environmental impact is growing. Data centers require cooling systems, continuous electricity, and water; the carbon footprint of a single large model can be significant.

This is why many developers are exploring smaller, more efficient models, such as TinyLLaMA, DistilGPT, and Mistral; quantization and compression techniques to reduce size and energy consumption; and on-device AI, with models running locally without depending on data centers.

These approaches can reduce the resources required and increase control over the environment. The benefit must be measured: a local system still requires access management, updates, and data protection.

Choosing the Right Balance

Bigger does not always mean better. For many tasks, such as summarizing emails, preparing reports, or guiding users, a medium-sized or optimized model can work well without high costs.

It is about choosing the right model for the right task, paying attention to both performance and impact.

Comparing AI Models — Cloud Giants and Local Experts

How do we choose the right model for the right situation? It depends on the requirements: raw power, flexibility, speed, cost control, or privacy. Not every task requires a gigantic model running in the cloud: sometimes a smaller local model is more than sufficient. In some cases, it is even the better choice.

Here is a simple comparison to understand benefits and trade-offs.

Feature	Large cloud-based LLM, for example GPT-4 or Gemini	Optimized local model, for example Mistral or TinyLLaMA
Performance	Must be measured on the task: depends on version, tools, and input.	Must be measured on the task; adaptation can help, without guaranteeing superiority.
Access	Requires Internet, an API, or a platform.	Can work offline on a laptop, server, or edge device.
Speed	Depends on servers, load, connection, and text length.	Depends on hardware, model, load, and text length.
Cost	Check prices, subscriptions, and request volume.	Hardware, energy, maintenance, and management; compare total cost.
Energy consumption	Depends on infrastructure, utilization, and load; requires comparable measurements.	Depends on hardware and utilization; local deployment does not guarantee savings.
Privacy	Transmission to a provider: check contract, access, and retention.	Can avoid external transmission; still check access, logs, and backups.
Customization	Depends on the options and constraints offered by the service.	Depends on the license, access to weights, tools, and available resources.
Most suitable use cases	Complex questions, creative writing, and research.	Customer support, educational tools, and assistants for the Internet of Things, IoT.

Guided activity — Evaluating cost and constraints together. For one hundred requests, a fictional service costs 0.02 euros per request: the variable cost is 2 euros. A local alternative costs 20 euros to manage over the period, plus any hardware and energy costs. The price per token alone does not settle the choice. Check the calculation and also measure errors and time. As a variation, increase requests to two thousand: the service costs 40 euros, but you need to check whether the alternative can handle that load. Reuse the comparison for the measured volume and the same requirements; do not automatically transfer prices or performance to another environment.

Part VII — Appendices

Before You Ask — A Checklist for Using LLMs Wisely

LLMs are powerful, but not magical. As with any tool, the quality of what you get depends on how you use it and how well you understand its limitations.

Here is a quick checklist for using LLMs responsibly, effectively, and safely.

Area	What to do	What to avoid
Know the tool	Understand the model you use: name, size, and purpose.	Using it blindly without knowing how it works.
Formulate clear prompts	Be specific, provide context, and define roles, for example “You are a teacher...”.	Writing vague or ambiguous prompts.
Think critically	Always double-check facts, especially in sensitive fields.	Trusting the output without verifying it.
Pay attention to bias	Test the model in different scenarios and from different perspectives.	Assuming the model is neutral or fair by default.
Protect privacy	Avoid pasting personal or confidential data into public models.	Assuming the system is private or secure by default.
Use responsibly	Consider the impact of the output and the people involved.	Using LLMs to automate tasks without human supervision.
Refine when necessary	Revise prompts and instructions iteratively to improve results.	Expecting perfect answers on the first attempt.

Conclusion

LLMs are changing how we interact with technology. They can write, translate, summarize, explain, and even surprise us with their creativity.

Their operation is based on learned representations and token generation. They can perform complex tasks, but a fluent answer does not demonstrate subjective experience, emotions, or reliability. The choice of use must depend on checked results and recognized limitations.

Yet, when used carefully, they can be powerful partners:

- They help us save time.
- They make complex ideas more accessible.
- They support learning, creativity, and communication.

The key is to remember that control remains with us. LLMs are here to amplify our work, spark new ideas, and remove some obstacles in everyday activities, rather than replace people.

Used with care, curiosity, and responsibility, these models can help us do more and think better.

Whether you are a teacher, a business executive, a student, or simply a curious person, you now have the basic knowledge to engage with this technology more confidently.

Glossary

Term	Definition
AI — Artificial intelligence	Technology that imitates human abilities to make decisions, solve problems, or perceive.
Language model	A system trained to predict and generate human language on the basis of patterns in text.
LLM — Large Language Model	A type of AI trained on enormous amounts of text to generate and understand language.
Prompt	A natural language instruction or question given to the model to guide its output.
Prompt engineering	The practice of designing clear, effective prompts to obtain better results from LLMs.
Token	A portion of text, a word or part of a word, that the model uses in prediction.
Inference	The process by which a trained model generates a response.
Fine-tuning	Further training of an LLM on specialized data to adapt it to specific tasks or fields.
Bias	Unintentional assumptions or stereotypes reflected in the model because of patterns in training data.
Hallucination	A situation in which an LLM generates false but convincing information.
Emergent ability	Performance detected in larger models and not detected in smaller ones in a specific evaluation; the result also depends on the metric.
Multimodal model	A model that can process and generate not only text, but also images, audio, or other types of data.
Cloud-based model	An LLM used through the Internet, running on remote servers.
Local model	A model running in an environment controlled by the user; possible sizes depend on available hardware.
RAG — Retrieval-Augmented Generation	A technique that allows LLMs to retrieve information from external sources during generation.
Quantization	A method for reducing a model's size and resource use, often to run it locally.

Want to Know More — Resources to Explore

Foundational Reading

Attention Is All You Need

Vaswani and colleagues, 2017. The work introduces the transformer and evaluates its performance in translation. This architecture underlies many modern LLMs; the original work does not demonstrate the abilities of every subsequent model.

<https://arxiv.org/abs/1706.03762>

Man Is to Computer Programmer as Woman Is to Homemaker

Bolukbasi and colleagues, 2016. A study of gender bias in word embeddings, numerical representations of words, which also proposes a method to reduce some of its manifestations. It is not proof that an intervention eliminates every form of bias.

<https://arxiv.org/abs/1607.06520>

On the Dangers of Stochastic Parrots

Bender and colleagues, 2021. A critical analysis of risks, biases, and ethics in large language models.

<https://dl.acm.org/doi/10.1145/3442188.3445922>

European Union — Artificial Intelligence Regulation

Regulation (EU) 2024/1689 is available on EUR-Lex, the official portal for European Union law. For a concrete use, consult the current text, applicable scope, and deadlines; this guide does not replace a legal assessment. The original's informational AI Act website is not the European Commission's website.

<https://eur-lex.europa.eu/eli/reg/2024/1689/oj>

Educational Guides

OpenAI — Prompt Engineering Guide

Practical examples and suggestions for constructing better prompts.

<https://platform.openai.com/docs/guides/prompt-engineering>

DeepLearning.AI — ChatGPT Prompt Engineering for Developers

A short course with examples for developers. Check the website for access arrangements and prerequisites, which may change.

<https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/>

OECD — AI and the Future of Skills

Reports on AI's impact on education and work.

<https://www.oecd.org/en/topics/artificial-intelligence-and-education-and-skills.html>

References for the Revision

These references support the added corrections and distinctions. The examples with receipts, pencils, prices, and company documents are original exercises with fictional data. The sources do not certify those data or the performance of a specific product.

Emergence and Evaluation

Wei and colleagues, Emergent Abilities of Large Language Models, 2022. On the relationship between scale and observed performance.

<https://arxiv.org/abs/2206.07682>

Schaeffer, Miranda, and Koyejo, Are Emergent Abilities of Large Language Models a Mirage?, 2023. On the role of the metric in apparent performance jumps.

<https://arxiv.org/abs/2304.15004>

Explanations and Their Limitations

Turpin and colleagues, Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting, NeurIPS 2023. On the distinction between a plausible explanation and a faithful account of influences on the result.

https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract.html

Information Retrieval

Lewis and colleagues, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2020. The work combines a generative model with document retrieval and evaluates tasks requiring knowledge. It does not guarantee that every retrieved source is current or correct.

<https://arxiv.org/abs/2005.11401>

Memory and Data Handling

OpenAI, documentation on ChatGPT memories. Describes product features: it should not be generalized to every LLM.

<https://learn.chatgpt.com/docs/customization/memories>

OpenAI, Data controls in the OpenAI platform. Distinguishes training use from data retention in APIs. Terms may change; recheck them for the applicable service and contract.

<https://developers.openai.com/api/docs/guides/your-data>

How to Read News Cases

Sources for the Mata order, the Colombian judgment, the Thomson Reuters Institute report, and the Amazon, Eliza, and Hoshi accounts are provided alongside the cases. News accounts have a different value from court documents and experimental studies. Full access to some articles is limited: the revision flags this, avoids unsupported attributions, and retains the original translation for comparison. References for the revision consulted on October 5, 2026.