RADAR SPECIAL

N. 12A · 4 NOVEMBER 2026

LLM
The boundary between scale and intelligence

What limits today's models and which paths remain open



Gruppo Radar · Comprendere oggi. Costruire domani

by effettoantropico.life

Opening — Before calling it a wall

When an artificial intelligence system makes a mistake, it is natural to ask whether that result reveals something it cannot do. Before answering, however, we must understand where the error comes from. The system may lack important information, or the difficulty may concern how it represents and processes it. The criterion by which we assess the answer matters too: we can judge a result insufficient without having identified what produced it.

In the debate these differences are easily lost. Some present a mathematical proof as though it had already established how far any AI can go, while others take it for granted that ever larger models will eventually overcome every obstacle. Both readings promise an apparently simple answer, but to understand what really happens we must distinguish the different aspects of the problem: *not everything depends on the same cause and not everything is explained in the same way*.

To find our bearings, then, we start with a more precise question: **what limit are we observing, in which system and under what conditions?**

An **LLM** (Large Language Model) *is an artificial intelligence model trained to process and generate language*. It can underpin a service that also allows users to consult documents, use tools or work with images. The model and the complete service, however, are not the same. To understand the capabilities and limits of what we use, we must therefore look both at the model and at the functions and tools with which it is integrated.

We will start with what happens between a question and an answer. Then we will enter the numerical space of models, the laws describing their improvement and attempts to understand how they work. The thread holding the journey together is concrete: which obstacle are we measuring, in which system and under what conditions?

1. Before the answer, a transformation

We write: *"Can you help me understand this document?"* For us the sentence contains an intention, a situation and a need. For the model the work begins with a sequence of **tokens** (units that can be words, fragments of words, marks or other elements). **Tokenization** is the division of text into these units.

The system then associates each unit with a **vector** (an ordered list of numbers). This step is called **embedding** (a numerical representation used to process learned relationships). It does not give the model meaning as a ready-made label: *it makes the text usable by the network's operations.*

In a *Transformer* (an architecture, that is, a way of organizing the model), attention helps relate different parts of the available text. Consider the sentence: *"Marta took the book and opened it."* Linking *"it"* to *"book"* is one of the relationships the system must be able to process. The mechanism works through **numerical weights**, *values that regulate how much information in different parts of the text contributes to the current step*. Not all contributions therefore have equal weight, and this combination changes with the context. Other steps in the model further transform the result. *"Attention"* is thus the name of a mathematical operation: *it does not, by itself, denote the human experience of concentrating on something*. [E02]

During **training**, parameters are adjusted (numerical values that regulate the model's operations). In **inference** (the processing performed when we use the already trained model), those values are applied to available information. In models that generate text **auto-regressively** (building the sequence one unit at a time), tokens just produced also contribute to what follows.

We see this in everyday use: *by adding a page or clarifying a question, we can obtain a better answer*. We have changed the **context**, the information available during that processing, but that does not mean the model's parameters have changed too. The original GPT-3 work already showed how instructions and examples included in text could guide different tasks without updating the model each time. It is a historical example that clarifies this distinction, but it is not a description valid for all current services. [E03]

So when we say *"now it has understood"*, we may have observed a real improvement in the current answer, but to speak of stable learning we need to know which component was changed and how.

> 📌 **Did you know? — A token is not a concept**
>
> A word may require more than one token and the same token may appear in different contexts. An internal feature of the model does not necessarily correspond to a vocabulary entry either. Tokens, features and numerical coordinates describe different things: counting one does not allow us automatically to derive how many "ideas" the system possesses.

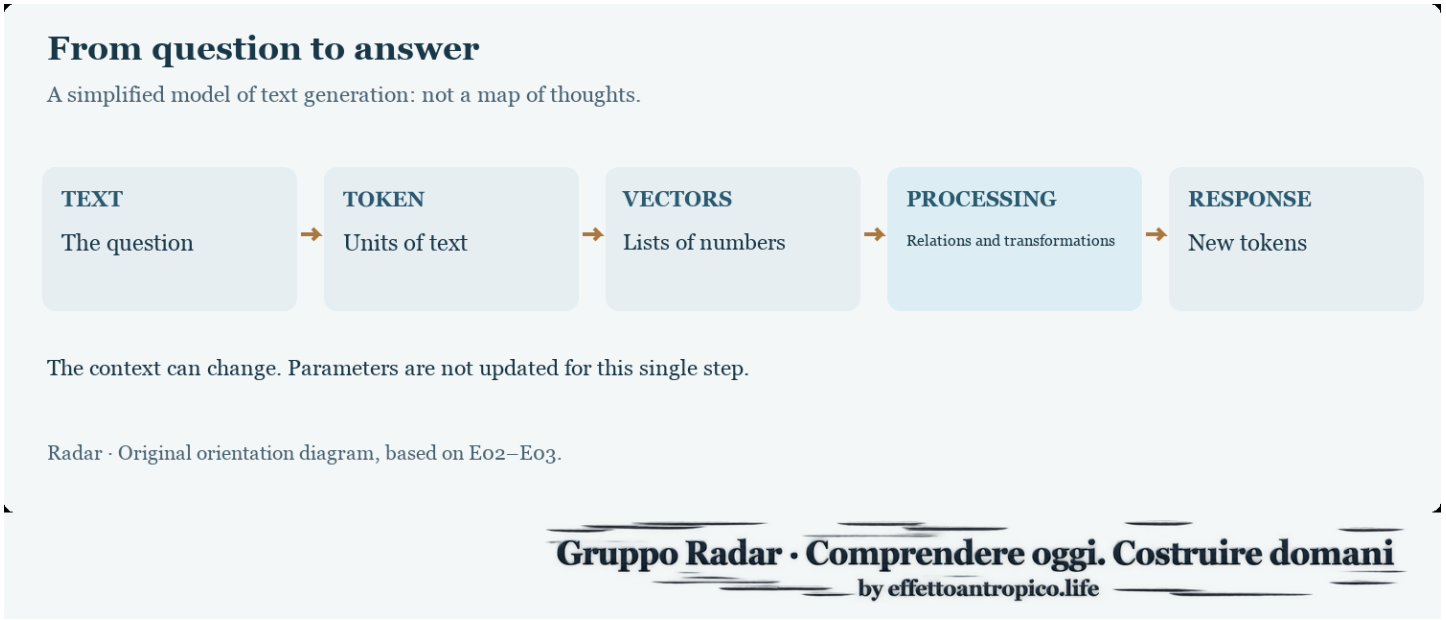


Figure 1. From question to answer: context and training act at different times..

2. A shared space can preserve many differences

The Italian word *"rete"* can mean the Internet, a football goal or a set of relationships. We distinguish the meaning through the sentence and the situation. In the model too, the representation is transformed along the way: *recognizing a unit of text is only the beginning of the work.*

A **dimension** is a coordinate in numerical space. It is not a shelf dedicated to an idea. With two coordinates we can describe many directions on a plane; this does not make all those directions independent and does not guarantee that every combination can be separated without ambiguity.

Research speaks of features (properties and patterns that contribute to processing). In the simplified models studied by Elhage and colleagues, more features can share the same dimensions. This organization is called superposition. It becomes particularly interesting when features are **sparse** (only some of the possible ones are active in the same example). Shared space can be used because not everything must be represented simultaneously. [E04]

Sharing, however, introduces the **possibility of interference** (the unwanted contribution of one representation to the reading of another). The system must be able to manage this interference, not merely contain more information. **Nonlinear** transformations (operations that change the signal differently from a simple sum or proportion) can help; their effect depends on the model's construction and conditions. [E04]

These overlaps can suggest that the model inevitably confuses information. To understand whether that really happens, we must test what it can distinguish, through which operations and in which tasks. The fact that one representation can interfere with another does not mean that separating them is impossible.

> 🧠 **Radar Insight — Containing is not enough**
>
> An efficient organization can represent many features in the same space and encounter difficulty separating some of them. That is why the useful question is twofold: *which information is represented, and how well is it used when needed?* The comparison must be made on the task, not on the allure of a number of dimensions

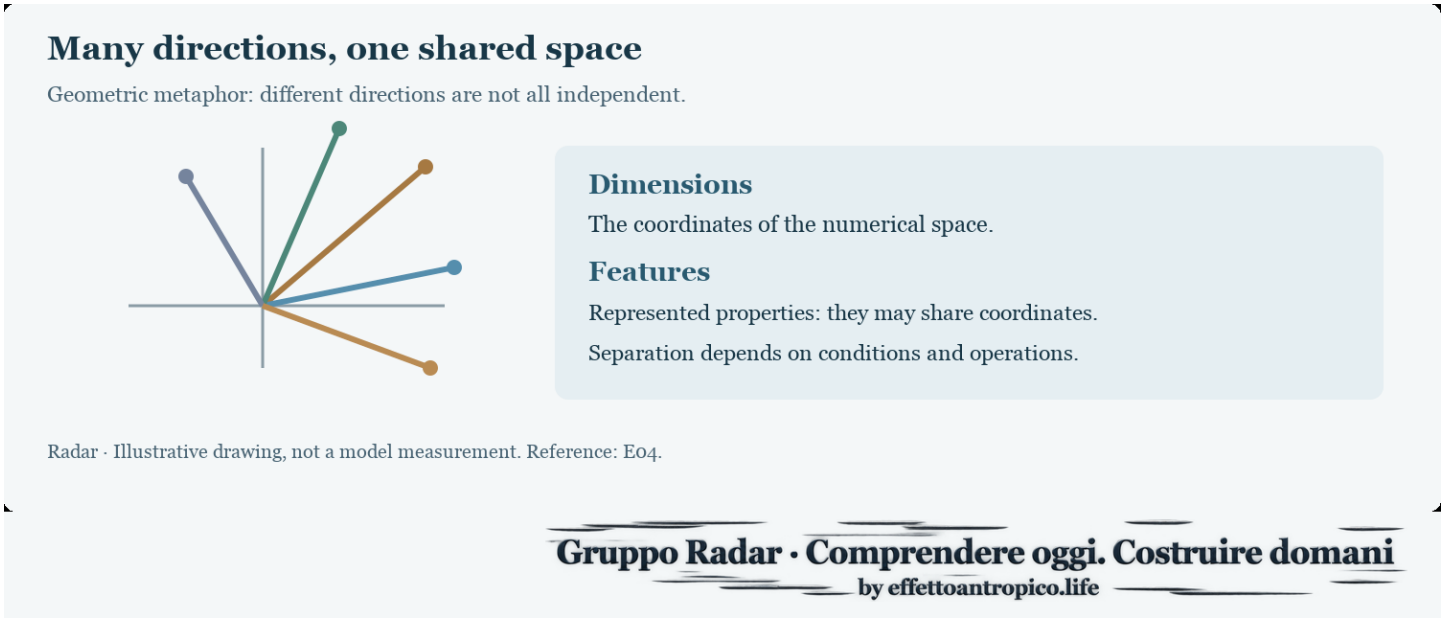


Figure 2. Illustrative drawing of shared space. The directions are not measured features of a real model..

3. What research on scale measures

To study a mechanism, researchers can build a simplified model, often called a toy model (a reduced study system that isolates some properties and leaves others out). It is useful precisely because it permits controls that are difficult in a huge network. Its strength and its limit arise from the same choice.

In *Superposition Yields Robust Neural Scaling*, presented at the **NeurIPS** scientific conference in 2025, Yizhou Liu, Ziming Liu and Jeff Gore study representations with different frequencies. In the strong-superposition regime analyzed, representation error decreases approximately in inverse proportion to width (the number of internal dimensions). The authors also compare the framework with families of open language models. [E01]

The study's discussion explicitly delimits the result: *the simplified model is not rigorously solved in all its behaviors and the study concentrates on representation, leaving other contributions to error outside its main scope. This is not a flaw to hide: it is what allows us to understand which conclusion the work supports*. [E01]

At the 2026 **ICML** conference, devoted to methods by which systems learn from data, two studies in the same line address other steps. The first examines **depth** (the number of successive processing layers). In the systems analyzed, many layers make similar updates; an averaging effect among contributions can reduce part of the error. The authors indicate this as a likely explanation and acknowledge that their statistical examination does not reveal the function of each individual layer. **Groups of layers with different roles remain compatible with the findings**. [E05]

The second study examines learning peaked **probability distributions** (situations in which a few alternatives receive much of the probability). With **softmax** (a function that turns scores into probabilities) and **cross-entropy** (a measure that penalizes a low probability assigned to observed outcomes), it finds a decrease in error according to a power law with a one-third exponent, in the regime studied. *"Universal"* in the title concerns the mechanism's robustness to different details, not every phase of every training process. [E06]

These works distinguish representing, transforming and learning. A synthesis by Yizhou Liu combines the contributions in a **semi-empirical** formula (built by joining theoretical indications with fitting to experimental data). The proposed sum is not a definitive law of all AI: *interactions between contributions remain a research topic*. [E07

4. Average error is not a complete measure of intelligence

Loss (the error measure used in training or evaluation) *tells us how the model behaves with respect to a precise objective*. In next-token prediction, it assesses the probabilities assigned to observed outcomes. It is a valuable measure, but does not necessarily contain everything we ask of an assistant.

Consider a system that summarizes contracts. We may care about the correctness of the summary, its ability to find an unusual clause, its willingness to acknowledge uncertainty and how it documents what it asserts. These are different questions, and an improvement in average prediction error does not verify all these capabilities.

Even a minimum threshold for loss must be interpreted in context. If a sequence admits more than one possible continuation, uncertainty associated with the data

remains. **Entropy** (the measure of this uncertainty in the distribution of outcomes) contributes to the minimum of the statistical measure. From this we cannot directly derive a maximum for creativity, usefulness or the ability to tackle a problem. [E07, E09]

A **benchmark** (a set of tests used to compare systems) adds another level. What matters is which questions it contains, how it awards points and how closely it resembles the use that interests us. A high result can be important and remain insufficient for another situation. To understand its value we must know the test, not only the score.

> ⚠️ **To watch — Before the number, the question**
>
> When we read that a model is *"better"*, let us look for the task, the measure, the system compared and the test conditions. Let us also ask which errors the average may conceal. This information changes the meaning of the result: an assistant reliable for a summary is not automatically reliable for a sensitive decision

5. Growing helps, but improvement has a shape

Scaling laws (relationships describing how performance changes as resources grow) have helped design language-model training. Many take the form of power laws (relationships in which a quantity changes according to an exponent) and can describe regular improvement that becomes progressively more costly. [E08]

To understand the shape of the curve, imagine gradually increasing the computing capacity available to the model. If error falls only slightly, we can continue to improve, but to obtain a benefit similar to the previous one ever greater resources will be needed (for example, a first doubling of compute might reduce error substantially, while another doubling might reduce it only slightly). We then speak of **diminishing returns**: *progress has not stopped, but each step forward becomes more expensive*.

How resources are distributed matters too. Under its experimental conditions, the study showed the importance of balancing model size and the amount of training data under a compute budget. *The most economical choice for training does not necessarily coincide with the most economical choice for a service that must then answer millions of requests*. [E09]

The picture remains open also because what is learned changes. Research by **Bordelon, Atanasov and Pehlevan** studies how **feature learning** (the change in representations during training) can modify scaling laws in the systems analyzed. The

result does not authorize a promise for every model, but reminds us that the way of learning is part of the problem. [E10]

Increasing size, then, is one lever among several. To judge it we must know which contribution to error we are trying to reduce and what cost we are willing to bear..

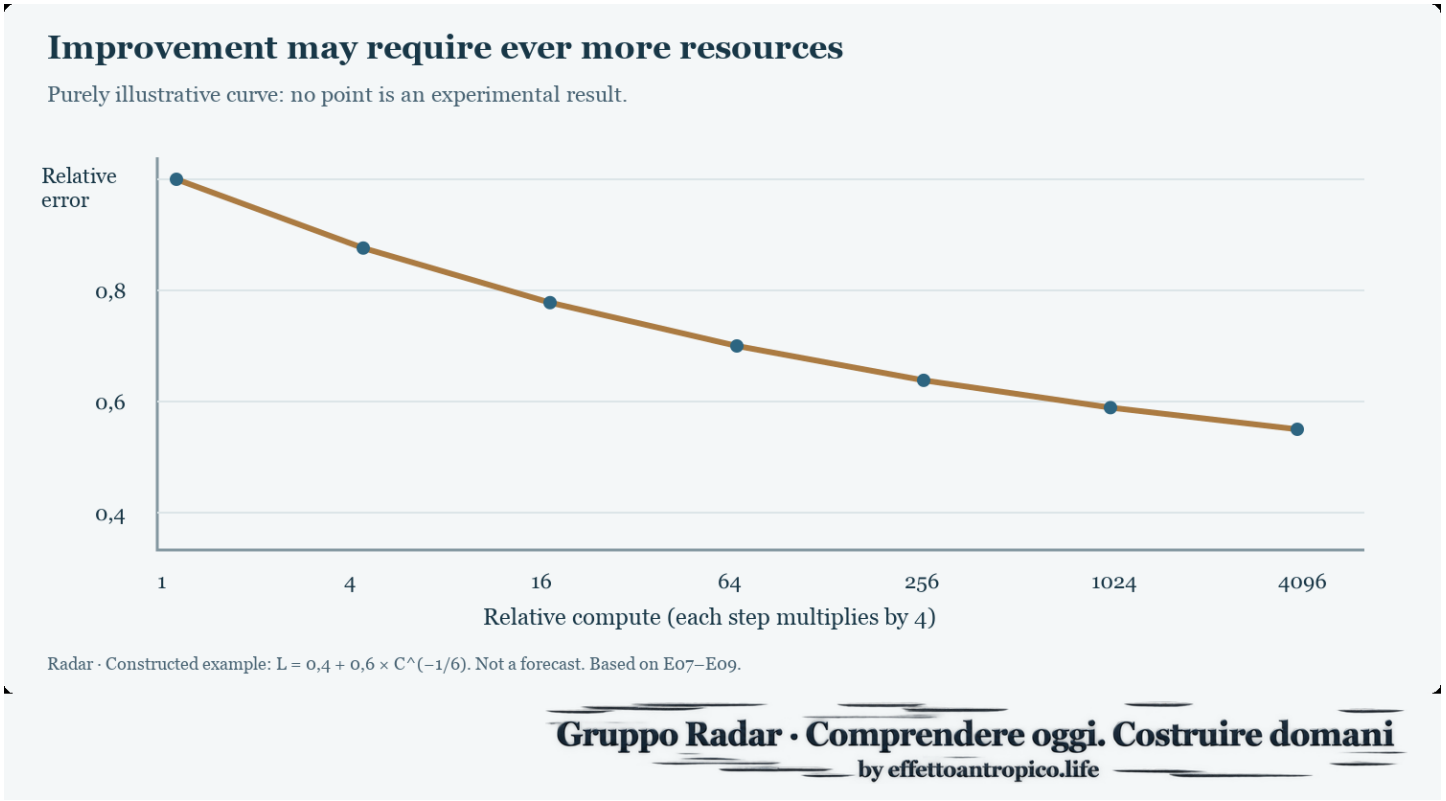


Figure 3. A mathematical example constructed to show diminishing returns. It contains neither experimental data nor a forecast..

6. When a theorem meets a product

A **theorem** is a conclusion demonstrated from assumptions. If the assumptions change, we must check whether that conclusion still holds. This sounds obvious, but in the story told about AI it is one of the steps most easily lost.

Michael Hahn demonstrated limitations for attention-based systems in certain formal tasks, such as recognising structures with parentheses. In the results on **hard attention** (which directly selects particular positions) and **soft attention** (which combines weighted contributions), the conditions and conclusions are not identical. Decisive elements include depth fixed relative to input length (the incoming information) and, in the softer case, constraints on how the model's transformations may vary. The analysis considers sequences that can grow without limit. [E11]

Another work, by P**érez, Barceló and Marinkovic**, demonstrates the **Turing completeness** (the theoretical possibility of simulating the computation of a general machine) of a construction with attention. It uses hard attention and a**rbitrary numerical precision** (the possibility of retaining numbers with as much detail as needed), conditions different from ordinary hardware operation. It does not demonstrate that a commercial assistant solves any problem, learns it from data or does so with reasonable resources. [E12]

How, then, can the results of the two works be reconciled? They do not contradict each other, because they study different constructions and conditions. They do, however, invite us to distinguish what we often put together: is a computation possible in theory? Can a system learn to perform it from data? And can it succeed reliably, with sustainable time and costs? Demonstrating the first possibility is not enough to guarantee the others

> 📚 **To explore further — Conditions are part of the result**
>
> To read a proof, let us ask again: which architecture does it use? How much memory does it allow? What precision does it assume? And how does processing time grow? A result concerning unbounded sequences can illuminate a structural limit without describing every chat error. A possibility with idealized resources can guide research without being a capability available today.

7. A false answer can have more than one cause

We call a **hallucination** a generated answer that presents false information or information without the necessary support as reliable. The name says a great deal, but groups different situations: *an invented citation, a contradictory step, a detail added to a document, a confident answer to a question that does not permit a conclusion.*

The causes should not all be assigned to the same internal property. Information may be missing, the wrong source may be retrieved, a detail of context may be lost, or the system may be rewarded for attempting an answer even when it should stop. This last mechanism is central to the manuscript *Why Language Models Hallucinate*: *the authors study how some training and evaluation criteria favor guessing over acknowledging uncertainty*. It is a bounded explanation, not the sole origin of every error. [E13]

Where information is located matters too. *Lost in the Middle* observed, in the models and tests studied, difficulties using information placed in the middle of long contexts. It is a historical result from 2024 and leaves us a practical question to test: *are having a document in context and correctly using a passage from it really the same thing?* [E14]

When checking an error, we can conduct simple, bounded experiments: *provide the correct source, clarify the task, move the relevant detail, explicitly allow an uncertain answer. If behavior changes, we have a useful indication for diagnosis*. Establishing an internal mechanism, however, requires further investigation.

8. Data, compute and money are not interchangeable

The amount of text available does not tell us how much a model can learn. Quality, variety, provenance and relevance matter. The study by Villalobos and colleagues examines the availability of public human-generated data and proposes scenarios under certain assumptions. It is not an expiry date for human knowledge, nor proof that research will stop in a particular year. [E15]

Then there are **synthetic data** (content produced artificially, including by other models). The work by Shumailov and colleagues shows risks of **model collapse** (loss of fidelity to the original distribution) in recursive training on generated content. *One question concerns the loss of less frequent parts of the data. The result depends on the process studied: it does not mean that any use of synthetic data necessarily causes collapse.* [E16]

Compute indicates the operations performed; money indicates what it costs to make them available, but between the two lie **hardware** (the physical components of devices), energy, facility utilization, memory, networks and service organization. A formula approximating training compute is not a formula for a company's budget. Likewise, a system cheaper to train may prove more expensive to use if each answer requires substantial additional work.

Therefore, when faced with an efficiency claim, it is wise to specify the denominator: *fewer resources for what result, at what quality and in which phase?* It is a simple way to prevent a technical advantage from becoming a claimed economic saving without having measured it**.**

9. The open paths concern more than size

An assistant can improve because the model changes or because the system around it changes. Distinguishing the two levels helps us understand what we have actually obtained.

With **RAG** (Retrieval-Augmented Generation, generation assisted by document retrieval), the system searches for relevant material and brings it into the model's context. Information can thus come from an external archive instead of having to be

derived only from parameters. The method does not guarantee a true answer: *the search, source updates and fidelity of the synthesis must be checked*. [E17]

Mixture of Experts architectures (networks that select some components to process an input) make it possible to activate only part of the parameters in a step. *Switch Transformers* studies this path. "*Experts*" is the name of the components: *it does not mean that each possesses a professional competence understandable like a person's*. [E18]

Another lever is to devote more compute to the answer. **Test-time compute** (resources used while tackling a task) can serve to search for alternatives or verify solutions. **Snell** and colleagues study its allocation in specific tasks and budgets. The benefit depends on the difficulty of the problem and the method: allowing more time is not an automatic guarantee of correctness. [E19]

Reinforcement learning (training guided by reward signals) can steer behavior towards desired outcomes. Work on **DeepSeek-R1** explores this for reasoning tasks and also discusses difficulties such as reward hacking (obtaining a score without truly satisfying the purpose that score should represent). The quality of the objective matters, not only the strength of the reward. [E20]

Finally, there are organizations different from attention alone. **Mamba**, for example, studies selective **state-space models** (systems that update an internal state, choosing how to retain sequence information in it). It is an architectural direction and does not prove that one alternative has already resolved all the limits of other systems. [E21]

These paths act on different problems: *access to information, use of components, search for the solution, learning criterion and sequence management*. Their value must be tested against results and resources, without confusing a new possibility with a general solution..

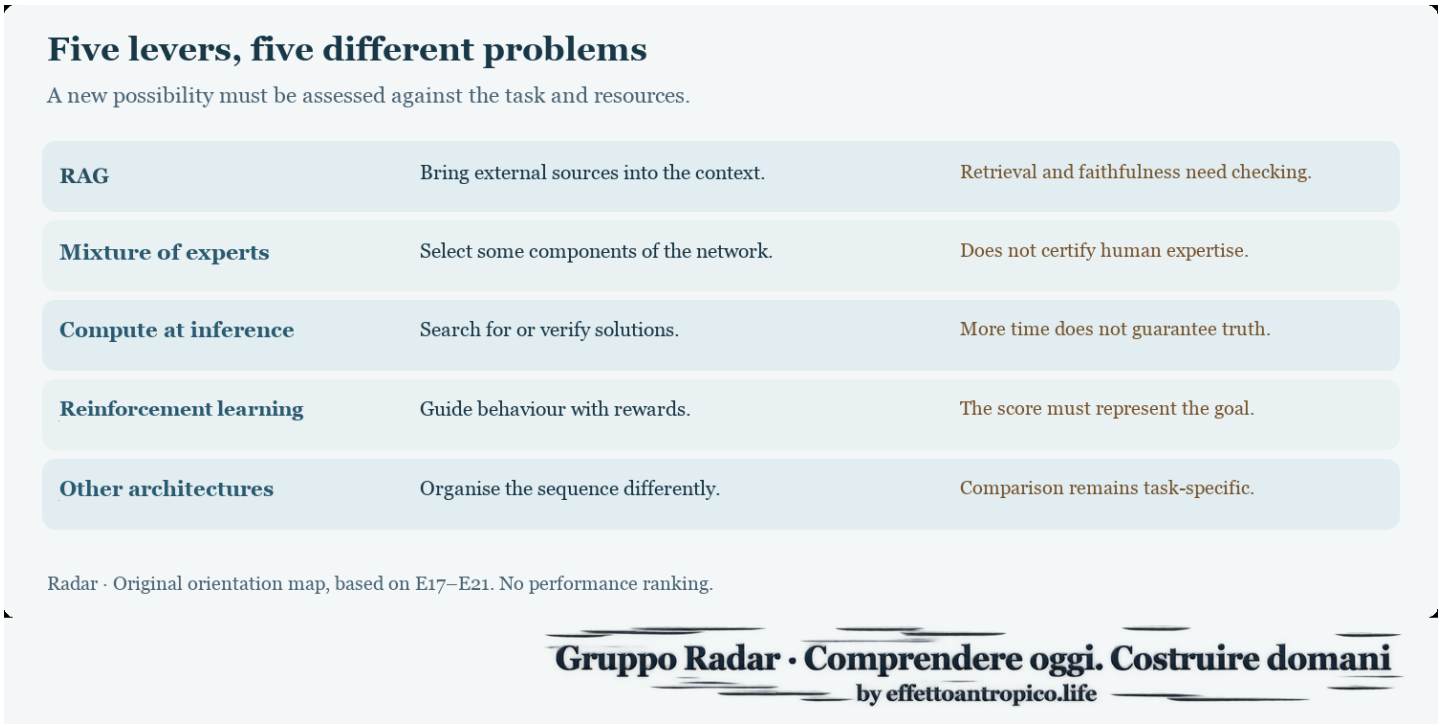


Figure 4. Different levers address different problems: every benefit requires verification on the task.

10. Understanding how it works requires more than a convincing explanation

Asking *"why did you answer that way?"* can produce an explanation useful to read. But that explanation too is generated text. Its coherence is not enough to demonstrate that it faithfully describes the internal operations that produced the first answer.

Mechanistic interpretability (research that attempts to reconstruct which internal components contribute to a behavior) addresses precisely this problem. One method studied by Bricken and colleagues attempts to break down model activity into more distinguishable features through dictionary learning (the construction of components that help reconstruct observed signals). What becomes more legible depends on the method, the data and how well the original activity is reconstructed. [E22]

In *On the Biology of a Large Language Model*, Anthropic's team uses attribution graphs (maps of estimated contributions among components) and interventions to examine some behaviors of **Claude 3.5 Haiku**. The maps are partial and the substitute system used for the analysis is imperfect. The authors also state that the cases studied were selected: *illuminating results do not thereby become the complete anatomy of the model.* [E23]

A **causal intervention** (changing an element and observing the effect on the result) provides different information from **correlation** alone (two phenomena varying together). Even a bounded intervention does not answer every question, but it allows a

hypothesis to be tested. It is this move from description to verification that makes the research valuable.

> ⚠️ **To watch — The account is not the trace**
>
> An explanation produced by the model can help us examine a line of reasoning or check a result. It should not automatically be presented as a record of its internal operations. Reconstructing a mechanism requires observations, tools and targeted tests; where the map remains partial, the limit must be stated.

11. The useful question after reading

We can now return to the starting point with a more precise question. *Faced with a limit, do we want to know whether it depends on representation, processing, data, training, context, resources or the test we have chosen?* **The more precisely we can locate it, the better we can understand which interventions make sense**.

For someone using an assistant in everyday work, this means checking the concrete task: *which documents can it consult? Which steps must be checked? When can it state that it does not know? And who answers for the final decision?* For someone developing a system, it also means stating the conditions under which the result was obtained and those under which it has not been demonstrated.

The research we have encountered offers tools to distinguish the opportunities available, showing mechanisms of improvement, diminishing returns, theoretical constructions, risks and alternatives. For the future, the result will come from the value we are able to assign to the field of observed results, which are currently more precise than the field of conclusions being attributed to them.

> 🧠 **Radar Insight — A boundary that guides the work**
>
> A documented limit deserves attention. Its meaning, however, depends on what was studied. Clarifying it allows us to use available tools better and recognize which questions remain open. We do not have to choose between blind trust and rejection: we can choose verifiable use, proportionate to the task.

Closing — Distinguish in order to build

A **model's size** tells us something about the resources it contains. A score tells us something about the test it faced. A **proof** tells us something about the construction

and assumptions it examines, but none of this information, taken alone, tells the whole story.

The truly interesting work begins when we bring all these options into relation. What really matters is the value we give to our ethics during study and analysis: we can recognize a result without any need to inflate it, and an error without turning it into a condemnation.

Understanding a limit is not meant to stop thought, but to tell us where to begin again, with which tools and with what responsibility.

Primary sources and reading path

E01. Y. Liu, Z. Liu, J. Gore, *Superposition Yields Robust Neural Scaling*, NeurIPS 2025. Paper originale.

E02. A. Vaswani et al., *Attention Is All You Need*, 2017, manoscritto degli autori. Testo originale.

E03. T. B. Brown et al., *Language Models are Few-Shot Learners*, 2020, manoscritto degli autori. Testo originale.

E04. N. Elhage et al., *Toy Models of Superposition*, 2022, pubblicazione del gruppo di ricerca. Originale su arXiv.

E05. Y. Liu, S. Kangaslahti, Z. Liu, J. Gore, *Inverse Depth Scaling From Most Layers Being Similar*, ICML 2026, PMLR 306:77093–77118. Editore e paper.

E06. Y. Liu, Z. Liu, C. Pehlevan, J. Gore, *Universal One-third Time Scaling in Learning Peaked Distributions*, ICML 2026, PMLR 306:77529–77558. Editore e paper.

E07. Y. Liu, *Neural Scaling Laws Trilogy: Representation, Transformation, and Training*, 28 febbraio 2026, sintesi firmata dall'autore. Sintesi primaria.

E08. J. Kaplan et al., *Scaling Laws for Neural Language Models*, 2020, manoscritto degli autori. Studio originale.

E09. J. Hoffmann et al., *Training Compute-Optimal Large Language Models*, 2022, manoscritto degli autori. Studio Chinchilla.

E10. B. Bordelon, A. Atanasov, C. Pehlevan, *How Feature Learning Can Improve Neural Scaling Laws*, ICLR 2025. Pubblicazione scientifica.

E11. M. Hahn, *Theoretical Limitations of Self-Attention in Neural Sequence Models*, TACL 8:156–171, 2020. Editore e paper.

E12. J. Pérez, P. Barceló, J. Marinkovic, *Attention is Turing-Complete*, JMLR 22(75):1–35, 2021. Editore e paper.

E13. A. T. Kalai, O. Nachum, S. S. Vempala, E. Zhang, *Why Language Models Hallucinate*, settembre 2025, manoscritto. Versione degli autori.

E14. N. F. Liu et al., *Lost in the Middle: How Language Models Use Long Contexts*, TACL 12:157–173, 2024. Editore e paper.

E15. P. Villalobos et al., *Position: Will we run out of data? Limits of LLM scaling based on human-generated data*, ICML 2024, PMLR 235:49523–49544. Studio originale.

E16. I. Shumailov et al., *AI models collapse when trained on recursively generated data*, Nature 631:755–759, 2024. Articolo originale.

E17. P. Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, 2020, manoscritto degli autori. Studio originale.

E18. W. Fedus, B. Zoph, N. Shazeer, *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*, JMLR 23(120):1–39, 2022. Editore e paper.

E19. C. Snell, J. Lee, K. Xu, A. Kumar, *Scaling LLM Test-Time Compute Optimally can be More Effective than Scaling Model Parameters*, 2024, manoscritto degli autori. Studio originale.

E20. D. Guo et al., *DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning*, Nature, 17 settembre 2025. Articolo originale.

E21. A. Gu, T. Dao, *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*, 2023, manoscritto degli autori. Studio originale.

E22. T. Bricken et al., *Towards Monosemanticity: Decomposing Language Models With Dictionary Learning*, 5 ottobre 2023, ricerca originale del gruppo Anthropic. Presentazione scientifica.

E23. J. Lindsey et al., *On the Biology of a Large Language Model*, 27 marzo 2025, ricerca originale del gruppo Anthropic. Studio e limiti del metodo.

© 2026 Barbara Gattobigio · Gruppo Radar

Comprendere oggi. Costruire domani · by effettoantropico.life

All rights reserved in the original content. Free access does not authorize reproduction or republication without consent, except for uses permitted by law. Third-party materials remain the property of their respective owners. To share this issue, use the link to the article.