The Strange Nature of Interpretability

Lachlan Kermode

"What's in a name? That which we call a rose
By any other name would smell as sweet;
So Romeo would, were he not Romeo call'd,
Retain that dear perfection which he owes
Without that title. Romeo, doff thy name;
And for that name, which is no part of thee,
Take all myself."

— William Shakespeare, **Romeo and Juliet**, II.2.43–49 [1]

As neural nets now suffuse everything touched or threatened by computation, the field of machine learning interpretability has enjoyed a renaissance in relevance. Large neural nets, as we all know, are strangely effective when it comes to comprehending and generating textually transcribed language, and they have yielded impressive results in what has come to be called the 'multimodal' domain, i.e. generating an image on the basis of a textual prompt, or captioning an image with text by processing its pixels as an input query. The effectiveness of LLMs in these domains is strange, in part, because we don't have much in the way of an empirically robust theory about *how* neural nets work to generate the outcomes they do. Interpretability research seeks to give body to this open question, detailing the mechanisms at work in neural nets; such as the particular clusters of neurons in a model that fire when a particular kind of output is produced.

Salient interpretability researchers in industry have rhetorically aligned their work with the hard sciences: as in, for example, the 2022 paper from researchers at Anthropic and Harvard titled "Toy Models of Superposition" [2], where the term 'superposition' is transposed from quantum physics in an attempt to explain signification's eerie materiality in weights and connections. Another more recent example also comes from Anthropic, the 2025 paper titled "On the Biology of a Large Language Model" [3]. Much like experimentation in the 'hard' sciences such as physics or biology, these and other titles suggest, interpretability research aims to theorize mechanisms that appear to us through empirical experimentation, expressing in as precise a language as possible what neural nets really *do*. This rhetoric transposes a pretension to scientificity that is even more deeply ingrained in machine learning research; the idea that the study of artificial neural nets in computers approximates the study of biophysical neural nets in the brain. The study of artificial neural nets wants to be both a quantum physics and a neuroscience, a pancake-stack of the hardest kinds of science. If we take the semantics of Anthropic's interpretability research seriously, working out how large language models work is analogous to mathematically formalizing the microinteractions of the brain, or setting logically straight how miniscule particles that escape straightforward sensibility interact.

Neural nets are scientifically strange in the way that many objects of science such as black holes and brains are: we don't have satisfactory semantics to grasp the empirical mechanisms that work in them. Yet I want to suggest that the science of large language models is further strange, strange in a second way, as the phenomena that it seeks to explain are elusive in a categorically different register than the objects under the microscope in the hard sciences. Rather than activity in the real world—a tennis ball's bounce, a reaction between Coca Cola and Mentos, or a rat navigating a maze—the data at work in the science of machine learning interpretability are collected from the cultural domain of language: paragraphs of text, an audio recording, a photographic snapshot, or a moving image.

The data objects of machine learning interpretability, in other words, have a patently subjective character that the objective artifacts of the hard sciences do not. Words and images do not pretend to stand as sensible separate from our cognition of them, as a tree falling in a forest or a supernova might. A neural net 'works' when it paints a picture that *makes sense*, renders an audio file that is *recognizable*, or writes an essay that is *legible*. Machine learning models are only *effective* if they produce a paragraph rather than scrambled sentences, an image rather than a potpourri of pixels, or a video rather than fragmentary frames.

The mechanics of the strange objects that machine learning interpretability seeks to make sensible aren't obviously situated 'out there' in the world, as we might imagine of forces in physics such as gravity and electromagnetism, or substances such as carbon dioxide or haemoglobin. They are at least partially 'in there', in the mind, sensible only as concepts that allow us to structure shared meaning. In a word, sentences and images are strange objects of science because we cannot claim that they exist independent of our apprehending them.

If there is such a thing as a science of neural nets, then, it is perhaps something softer than the hard sciences such as chemistry, biology, or physics. Language and images are the objects of a decidedly social materiality, for their meaning is distinctly anchored in intersubjective space and not in strictly, supposedly objective phenomena. In taking this soft, social materiality as its object, machine learning interpretability seems more similar to the cognitive sciences, social sciences, or perhaps even history (when it is raised to the order of being systematic, as in *Geschichtswissenschaft*) than to chemistry. The subjects of its scrutiny—language and the image—are *social* artifacts, not natural ones; even as there may be mechanisms to discover in the way that neural nets produce and consume these artifacts.

I want to try to show very briefly how the metaphor of machine learning as a hard science falls flat via a case study in the art of reasoning, a talent of which certain language models are now apparently capable. The idea of chain-of-thought reasoning in the domain of LLMs is a training technique through which a text-to-text generative model is given more time to arrive at an answer in an effort to produce a better result. Chain-of-thought reasoning builds on an idea put forward by DeepMind researchers in a 2017 paper, which proposed that language models could "learn to solve algebraic word problems by inducing and modeling programs that generate not only the answer, but an **answer rationale**, a natural language explanation interspersed with algebraic expressions justifying the overall solution" [4]. The problem here, according to the paper, is that natural language is not necessarily the most natural medium in which to solve problems that involve mathematical reasoning. As the paper's introduction states:

> Solving these tasks is challenging as it involves recognizing how goals, entities, and quantities in the real-world [sic] map onto a mathematical formalization, computing the solution, and mapping the solution back onto the world [4].

Mathematical reasoning was hard to achieve in natural language, according to these researchers, because it involves commuting between computation, determinate functions that take as input numerical quantities, and the "real-world", an indeterminate phenomenological surround that needs to be represented in qualitative 'entities' before it can be computed or quantified. This is provided in way of an explanation in 2017 for why, empirically speaking, language models struggled to produce the correct answers to question prompts that required mathematical reasoning, even as they could produce other kinds of text that was impressively 'natural'. It was not clear that model training methods such as attention and RLHF (Reinforcement Learing from Human Feedback), methods which had reinvigorated interest in language models and enabled products such as ChatGPT, could also carry the day when it came to the forms of *mathematical* reasoning we readily associate with intelligence in the Western world today.

Both chain-of-thought reasoning and its precursor in DeepMind's 'answer rationales' were inspired by a pedagogical insight that we might remember from our high school mathematics class. If one tries to arrive immediately at the answer to a complicated question, most minds will skip over intermediate steps in the necessary mathematical reasoning, resulting in an incorrect answer. The remedy for this psychological insufficiency in schoolchildren, of course, is to ask students to show their working steps in full, so that the solution can built up piecemeal, and the mathematical reasoning at each step verified.

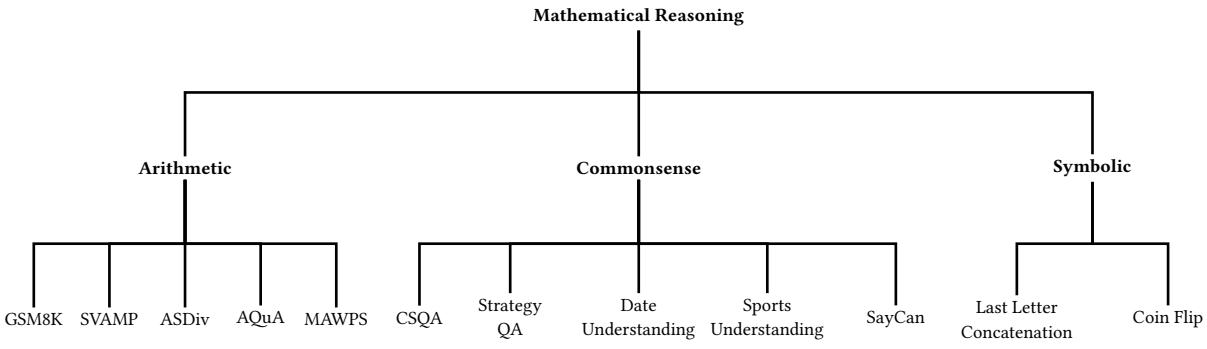
This is what chain-of-thought reasoning means in the context of LLMs. In order to make LLMs more effective, researchers introduce a special delimiting token during SFT (Supervised Fine Tuning), a late stage of the LLM training pipeline. The LLM then learns to produce this special 'think' delimiting token just as it does any other token. It learns to produce a weighted probability distribution from which the next token is sampled through a process known as reinforcement learning, where the model is incentivized to produce reasoning traces that will result in more correct answers.

`<think>` ——— **`[chain-of-thought]`** ——— `</think>` ——————— **`[answer]`**

Training models to produce chain-of-thought reasoning has been empirically validated as a way to improve a model's ability to correctly solve certain kinds of problems. The breakthrough paper that secured chain-of-thought reasoning as an essential part of the LLM toolkit was a 2022 paper from researchers at Google Brain [5]. (Google Brain and DeepMind merged as facets of the Alphabet company, best known for its Google search engine, in 2019 to form a unified research division that is now known as Google DeepMind.) This paper coins the chain-of-thought terminology, and also makes a more general claim about what the mechanism enables in language models:

> We explore how generating a *chain of thought*—a series of intermediate reasoning steps—significantly improves the ability of large language models to perform complex reasoning.... [C]hain-of-thought prompting improves performance on **a range of arithmetic, commonsense, and symbolic reasoning tasks**. [5]

Note that 'mathematical reasoning' in 2017 has been expanded to encompass "a range of arithmetic, commonsense, and symbolic [kinds of] reasoning" in 2022. True to its word, the Google Brain paper is broken up into three sections, each of which details the results of experiments using chain-of-thought models on a set of benchmarks.



Each of the benchmarks named in the bottom layer of this graph represent a set of problems analagous to a high school exam. The questions are put to a model, and the answers are held out. A

model's performance on a benchmark is how well its answers match up with the textbook answers to the test, which is how measures such as 90% are produced. Benchmarks are an essential apparatus in the laboratories of machine learning. After all training is done and dusted, they are the end-of-year exams in which we see how adequately models can respond to questions carte blanche. A machine learning model's capabilities are measured, almost exclusively, through their performance on benchmarks such as these. We say that a model can 'do arithmetic' because it can solve problems in a benchmark, or set of benchmarks, that are presumed to be representative of 'arithmetic' problems in general.

In this way, benchmarks are the empirical tasks through which machine learning researchers make the claim that 'model X can do, or solve, Y'. Something important happens when we gloss the empirical reality of a model's performance on certain benchmarks using abstract names such as 'arithmetic' and 'commonsense', or indeed when we gloss the empirical mechanism of training a model through reinforcement learning with specialized 'think' tokens as 'reasoning'.

(Side note: the claim that some of these models are now capable of 'PhD level reasoning' is also somewhat strange, in light of the fact that benchmarks are the empirical proof of a model's capabilities, as I think that many in the sciences and the humanities would agree that what one learns during a PhD is strictly irreducible to standardized testing.)

First and most evidently, there is a naturally-occurring semiotic leakiness in the top-level categories that group benchmarks together in this diagram. There are kinds of tasks one could reasonably call arithmetic, I have no doubt, that are not adequately represented by the bunches of benchmarks here. When Google DeepMind researches announce that reasoning models are better at arithmetic, we must critically scrutinize the experiments that were conducted in order to substantiate this claim, critically interrogating whether 'commonsense' is appropriately precise as a nominative abstraction for the tasks in the benchmarks beneath it.

The lack of this kind of properly scientific scrutiny of the terminology tossed around in machine learning research, I would argue, is an important component of what allows some to delude themselves that LLMs are intelligent, or even (more preposturously) that state-of-the-art models have achieved or are close to achieving AGI, Artificial General Intelligence. We could first put the empirical un-sensibility of these nominated states of achievement under a semiotic microsope. How impoverished must our notion of 'intelligence' be in order to be fooled into thinking that ChatGPT or Claude achieves it? We live in a dull and hopeless world if intelligence amounts to nothing more than being able to produce code that functionally satisfies a formal specification, or to read and write email in the assembly line of digital white-collar work.

What is beautiful and frightening about language as a domain is that it cannot be fully functionalized in terms of the tasks and objects that sometimes travel through it. The professor of electrical engineering and computer science Ben Recht provides a concrete definition of evaluation in machine learning that shows why language's functional slipperiness complicates machine learning's metaphorical self-conception as a hard science. Recht's definition is as follows:

> Evaluation is measuring the difference between articulated expectations of a system and its actual performance [6].

This definition is as expansive as it is general. In order to evaluate a model, we must articulate (name) an expection of the system at hand, and measure through empirical experminentation whether or not the model conforms to this expectation in practice. What evaluation puts under the microscope, in other words, is not only a model, but also *a name*. Evaluation is a process by which we can meaningfully claim that a name *works* to represent the empirical reality of what a model can do. To ensure that 'mathematical reasoning' is an appropriate name for the kind of tasks a reasoning

model can carry out, we evaluate the model's performance on a set of benchmarks that we presume to be indicative of 'mathematical reasoning'.

Yet names are strange objects with which to do science, as their true meaning can never actually be nailed down: they slip away from themselves. To treat a name as a transparent reflection of reality is analagous to grasping at a shadow. It misses the nature of the materiality at hand by eliding two fundamentally discontinuous domains: sense and vision in the case of a shadow, language and phenomena in the case of a name.

In a lucid piece that was recently published in *Critical Inquiry*, Mercedes Bunz ruminates on what the territorial distinction between language and phenomena, words and things, means for LLMs. Citing the philosopher Hilary Putnam, Bunz notes that "langauge extends into the world—or rather, the world and its things extend into language, though nothing seems to touch the other realm directly" [7, p.438]. Language is a domain that has an *extension* into worldly affairs, for Bunz, and the nature of language's extension is critical for thinking about its utility, its meaning, and its reality. Because LLM-produced language does not reach out and attempt to grasp the world in the same way that human language does—because it has a different extension—Bunz argues that computer-generated writing operates in a fundamentally distinct domain than natural language:

> Generated writing, however, is first and foremost different writing, one that comes with its own tendency as it operates in a different mode of production, doing words without things. [7, p.441]

Science, in contrast to computer-generated writing, is maybe best thought as a kind of doing words *with* things. It is the verification of names through controlled and consistent experimentation, an empirically evidenced claim that a concept is adequate to the world. If we agree that reasoning models are good at mathematical reasoning, we do so because we agree that the set of benchmarks shown above are adequately representative of what it means to mathematically reason. Science, in both its hard and softer varieties, is strictly a matter of terminological consensus regarding a realm of material affairs. Naming is a deficient instrument in the sense that it does not extend directly into empirical reality: names and things do not strictly overlap in language.

It is perhaps for reasons of wanting to cover up the semantic slipperiness of science's work with names that the field calls itself machine learning *interpretability*, rather than machine learning *interpretation*. Interpretation is a technique in the social sciences and the humanities, firmly situated on the side of the subject in the sense that it acknowledges reality must be filtered through and distorted by the viewer's situation within it. Interpretability, on the other hand, lends itself suggestively to objectivity, in that it frames the neural net as a natural given that we simply and pragmatically seek to explain without stepping into theoretical quicksand. We need to train models that have this quality of interpretability, the field suggests, as interpretable models will be able to produce language and images whose cause and reason can be explained.

What the 'interpretability' framing glosses over, however, is the strange natures of language and image as scientific domains. The phenomena that interpretability puts under the microscope are mechanisms that produce words and images, objects that are self-evidently cultural rather than natural in any obvious sense. In experiments such as those in which a reasoning model performs better on mathematical benchmarks than a non-reasoning model, the analogy to chemistry or biology can gloss over the fact that language, the substance being produced, has a fundamentally different nature than chemical compounds or proteins. Concepts are the actual objects that experiments refine, and language is the real domain in which science takes place.

If we want to think about the empirical study of neural nets as a science, we need to be more careful about how we employ metaphors to suggest it as such. Neural nets are not obstinate materialities

that simply do things, and which we therefore cannot change or control. They are a curiously compelling way to transform projections of language and visuality, digital text and images, into other kinds of visuality and language. A robust practice of interpretation in the field of machine learning must be deeply sensitive to the social nature of the domains in which models work; for what we consider 'good' or 'reasonable' language is not objectively fixed, but historically and socially contingent. A genuine science of neural nets, in other words, must recognize that the subjective nature of language is something to be taken seriously.

Bibliography

[1] W. Shakespeare, *Romeo and Juliet*. in The Arden Shakespeare Third Series. Bloomsbury Academic, 2012.

[2] N. Elhage *et al.*, "Toy Models of Superposition," no. arXiv:2209.10652. arXiv, Sep. 2022. doi: 10.48550/arXiv.2209.10652.

[3] J. Lindsey *et al.*, "On the Biology of a Large Language Model. Anthropic," *Preprint. https:// transformer-circuits. pub/2025/attribution-graphs/biology. html*, 2025.

[4] W. Ling, D. Yogatama, C. Dyer, and P. Blunsom, "Program Induction by Rationale Generation : Learning to Solve and Explain Algebraic Word Problems," no. arXiv:1705.04146. arXiv, Oct. 2017. doi: 10.48550/arXiv.1705.04146.

[5] J. Wei *et al.*, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[6] B. Recht, "Benchmarking Culture." Dec. 2023.

[7] M. Bunz, "On the Calculation of Meaning, or Doing Words without Things," *Critical Inquiry*, vol. 52, no. 3, pp. 423–446, Mar. 2026, doi: 10.1086/739755.