ARTIFICIAL INTELLIGENCE FRAMEWORK: TOWARD THE NOTION OF AGI

Anonymous authors
Paper under double-blind review

ABSTRACT

Within the limited scope of this paper[1], we argue that artificial general intelligence cannot emerge from current neural network paradigms regardless of scale, nor such approach is healthy for the field at present. Drawing on various notions, discussion, present-day development and observations, current debates and critiques, experiments and so on in between philosophy, including Chinese Room Argument and Godelian argument, neuroscientific idea, computer science, the theoretical consideration of artificial intelligence, and learning theory, we address conceptually that neural networks are architecturally insufficient for genuine understanding. They operate as static function approximators of a limited encoding framework - a 'sophisticated sponge' exhibiting complex behaviours without structural richness that constitute intelligence. We critique the theoretical foundations the field relies on and created of recent time, for example, an interesting heuristic as *neural scaling law* (for example, Kaplan et al. (2020b)) made prominent in a wrong way of interpretation, he Universal Approximation Theorem addresses the wrong level of abstraction, and in parts, partially, the question of current architectures lacking dynamic restructuring capabilities. We propose a framework distinguishing existential facilities (computational substrate) from architectural organization (interpretive structures), and outline principles for what genuine machine intelligence would require, and furthermore, an conceptual method of structuralize the richer framework on which the principle of neural network system takes hold. Such is then of conclusion, that the field's repeated AGI predictions fail not from insufficient compute, but from fundamental misunderstanding of what intelligence demands structurally.

Large Language Model, or LLM (Vaswani et al. (2017); Devlin et al. (2019); Brown et al. (2020); Zhao et al. (2023; 2024); Radford et al. (2019; 2018); Raffel et al. (2020); Touvron et al. (2023b;c); Chowdhery et al. (2022a); Ouyang et al. (2022); Wei et al. (2022); Kaplan et al. (2020a); Hoffmann et al. (2022); Bai et al. (2022)) is one of the most successful, most advanced, and most developed type of model in the current modern machine learning landscape, and of AI (Artificial Intelligence) research at large. Its success has not been lacking, and its reputation and widespread uses have been proved over time. The effect of LLM has been realized, and indeed has been changing the landscape of society in a very much difficult way. Latest model of such architecture, like OpenAI (2025a); Wang et al. (2025); OpenAI (2025b) GPT-5, Guo et al. (2025); DeepSeek AI / Hugging Face (2025)'s DeepSeek AI, Anthropic (2024); Anthropic Research (2025)'s Claude AI, Touvron et al. (2023a)'s LLaMA, Chowdhery et al. (2022b); Anil et al. (2023)'s PaLM / PaLM-2, and Jiang et al. (2023); Mistral AI (2023)'s Mistral (Mistral-7B), pushed this boundary further and further, and levelling up many tasks and purposes with AI system in practice, and further onward with techniques like Wei et al. (2022)'s *Chain-of-Thought prompting*, Kaplan et al. (2020a)'s scaling law, and more (Hoffmann et al. (2022); Bai et al. (2022)).

However, with such development, come great expectation, great speculation, and also great hallucination. New development of the field of AI even earlier than Vaswani et al. (2017) paper on the Transformer neural network which fuelled the revolution of AI exposure, has gathered a group of people speculated about the further exponential growth of AI, almost to a degree of religious, about the topic of a *Singularity*, where AI will become **Artificial General Intelligence** (AGI). This is re-

[1]The paper is not finished, this is the latest draft as of Monday 3rd November, 2025 at 02:11, will be uploaded to at least arxiv later.

flected in popular culture phenomena, speculation, researches, interpretation of reasoning behaviours and so on, for example, in Barrat (2013); Birch (2024); Yudkowsky & Soares (2025); Hao (2025); Bostrom (2014), and more broadly on LLM in specific, Mumuni & Mumuni (2025); Shang et al. (2024); David Ilić (2024); Goertzel (2023); Feng et al. (2024); Ryunosuke Ishizaki (2025) and Cui et al. (2024). The claim is clear - we are pushing toward the age of AGI, and perhaps sooner or later, reach the state of Artificial Superintelligence (ASI) of which cited in popular culture as the cultivation point of the Singularity - the shift of society toward a society of abundance, post-scarcity state. Most proponents point to LLM for such advancement, as it is one of the most widespread success and accessible form of interaction with AI systems on large scale. Pushbacks against such movement, including such as Friedman (2024); Quentin FEUILLADE-MONTIXI (2024); Bender et al. (2021); Baan (2021); Uddin (2023) on the "Stochastic Parrot Hypothesis", Villani (2024) questioned of LLM path to AGI, the reverend Forum (2025) post itself, Abdur Rahman Bin Md Faizullah (2024); Zhijian Xu et al. (2025) on generating suggestive limitation of research paper, Marcus (2025; 2023) critique on generative AI on world models and failure of LLM, Sandra Johnson (2024) on limitation of LLM, similarly Zhang et al. (2025), mathematics critique in Mirzadeh et al. (2024), and more. However, the generally public, and more so of the positivity of the inner market on AI focus on the development and increment of larger models toward such goal. It is not too offset to hear the phrase "AGI will be in X days/month/year", as much as it is a social phenomena even in small or large circle. Objectively, such positivity is not without basis. Furthermore, it is rather with certain amount of irony that the research made use of AI itself, for reference taking purposes.

Nevertheless, a critical task can be given out of such argument and thorough development of the current debate. What can then be extrapolated from the ongoing dilemma? What has to do with the architecture, the consideration about AGI that is now turned into the debate of will LLM be AGI? How is our understanding of the concept of AI, AGI, and ASI in general? And of a sense, what will provide us a pathway toward such goal?

1 Understanding Artificial Intelligence

The topic of artificial intelligence has been debated since the long form of intellectual pursuit, started with, not as comprehensive starting point as possible, Turing (1950); McCarthy et al. (1955). The conversation following such cultivated into different voices of definition and opposition via problems, such as McCarthy (1987) on the definition of strong and weak AI, Searle (1980b) definition of Chinese Room Argument, Russell & Norvig (2010) definition of intelligence on basis of agents action, Penrose (1989b) argument against the philosophy of computational intelligence, Floridi (2004)'s ethics. More can be found on Stanford Encyclopedia of Philosophy (2018) and further sources, but it is sufficed to say that the topic has received no less of contributions thereof to itself. However, it is reasonable to see that such argument and debate is often very shallow and hypothetical.

Inside the term *artificial intelligence*, there is the word *intelligence*. A normal person will tell you that they are intelligent. But it just so happens that this notion of qualification is harder to define when one participates in the active action of finding it. So, what is it? This is the question we should take in.

That said, this question is very much arduous in its public view, and distastes of the technical crowd. just as Øygarden (2019) mentioned, for philosophers whom are assumed to be interested in such endeavour, the topic of intelligence has traditionally appeared of a less interesting concept than consciousness. Pardon their minds and view, the philosophy of mind and body seems to attract more on the front of abstract thought, than something considered to be mechanical by thinking, and attribute no more and no less to the notion of being human as it is, with the soul and body being the main question. Developments of AI have caused new philosophical interest in the concept of intelligence, though seldom appropriately decoupled from its closely related phenomenon consciousness. Even in the field of AI itself, there are the avoidances of defining intelligence, though there have been no shortage of finding one in the middle of the forest. It is said that, for someone to work in the field of artificial intelligence, it would be wise to couple oneself with a definition of AI on him/herself, rather than not. Such is to say as to fix a philosophical standpoint before working in the field, which both contributes to the enormous amount of opinionated definition, but also the rigid framework on which artificial intelligence is considered.

As such, there exists no satisfying definition of artificial intelligence beyond the notion of artificial, of which is still dubiously believed upon. However, to fully capture the notion of AGI, we need the notion of AI on it.

1.1 UNDERSTANDING ARTIFICIAL

What separated the artificial and the 'natural product' by definition? Below is the table on such terms of being artificial taken from the most basic of the knowledge, etymology and definition available in large, of Merriam-Webster (2025); Cambridge Dictionary (2025); Oxford Learner's Dictionaries (2025); Dictionary.com (2025); Justia Legal Dictionary (2025); Merriam-Webster Etymology (2025); Bianchini (2021); NASA (2023); IBM (2024); Legg & Hutter (2007a); Goertzel (2014) and so on so forth. Indeed, for as long as the field artificial intelligence is formed formally of 1956, in the *Dartmouth workshop*. In the process of making something artificial, one must then have to reproduce what they considered natural, in certain perspective — not man-made — into a man-made form, with certain criteria. This is discussed in Simon (1969); Haugeland (1985); Boden (1987; 1996); Boden & Edmonds (2019); Onyeukaziri (2022); Boden (1990), in particular, about natural and artificial concept.

Table 1: Comparative definitions of *artificial* across different domains

Domain	**Core Definition**	**Notes / Nuances**
General dictionaries	Made by humans; imitation of nature [a,b,c,d]	Often connotation of "fake" or "not sincere"; opposed to natural.
Technical (AI, computing)	Human-designed systems that simulate or replicate functions of intelligence [e,f,g,h]	Must handle unpredictable environments, learn, and adapt. Debate on narrow vs broad definitions.
Biology / Synthetic biology	Engineered biological systems (synthetic cells, genetic circuits) [i]	Blurs the line between "natural" and "artificial"; challenges classical dichotomy.
Classification (taxonomy)	Groupings based on superficial traits rather than evolutionary lineage [d]	Used in contrast to "natural classification."
Legal / Social	Constructs created by law, rules, or human institutions [j]	Examples: "artificial person" (corporate law), borders, price manipulation.
Philosophy / Etymology	Derived from Latin *artificialis*, meaning "produced by human skill" [k]	Distinction between natural and artificial debated since antiquity; less clear in modern science.

In all, the definition of artificial can be considered diluting, since essences and the term is usually wholly considered in different voices and perspectives. For example, one might mistakenly classify biological to being natural — certainly a farm can be just as biological as a plain field without intervention, but not natural as it is[2]. On another matter, every object that are man-made still follows the law of physics, they obey the law of which is itself an expression of how the universe works, but such artifacts are called artificial, not natural, even though a rock still obeys such rulesets. We can then see that to equate something with the property of being artificial, requires more than just putting on its label, for the term natural itself is very hard to grasp. The science of artificial itself is now being advocated, much to Simon (1969) idea that a lot of the current world is now artificial in perspective. However, one can propose such dilemma by equating artificial with something entirely: *process*. Simply put, we define natural as without intervention. By our scientific fact, we know that the world changes and evolves itself, develops and mutate by time, of the second law of thermodynamics. Then, we can define something being artificial as a thing created, from a process outside such evolution, of which then subjectively, means created by human. To be created by human also means to conform to the notion of **interpretation**, **encoding**, of which is then represented in certain ways or form. While philosophical, it is trivial to see that without interpretation, an object itself will remain as plain of the object itself, without purposes, without triviality. Just as a neural network without the interpretation

[2]Under the same line, we also then have the discrete notion of *biologically artificial*. A child that is born with intention of its father and mother is a natural-born child, while artificial childbirth would be a non-biological process with intervention and non-biological form to facilitate such process.

to be a neural network is useless, and just as mathematics is useless without the interpretation of the subject and objects of consideration. Thereby, we then accept such provisional overview as the definition of artificial, from the subjective view of being created by human, hence constructible, and is created of purposes, its process and operations can be given meaning and representation.

The definition has a drawback, though, that is, if we later on discover an alien species with the capability to construct such artificial intelligence, would we not call it artificial intelligence, because it is not made by human? We need a definition that generalize to certain notion of intended creation, for in such event if we ever discover alien lifeform, then their "AI" would not be argued against such to be not artificial because of the word's basic meaning related to human, and hence can be considered a stipulative definition in its stead. Doing such, requires the notion of **intention**, and the notion of reason.

Definition 1.1 (Artificial)**.** *A subject A is artificial, if it is created by an object B, with intention and tasks p, constructed in a way of which the logic system of B can interpret such object, the representation of B can simulate and construct A, and A's operation lies within such domain of interpretation and logic of B.*

The notion of artificial will always be troublesome, as it is based on the ground that there exists certain subject of the current universe, B, that can facilitate and create another object A of various means, with interventions and interferences from one's natural evolution. While not satisfactory, it will be our provisional definition on the topic of artificial. In a sense, however, it is wise to note that the definition and treatment of the term artificial is purely subjective, per subject in question. The fallacy of such provisional definition comes in the very simple thought experiment. If ants are to create its own fungus farm, would we call that fungus farm artificial, or natural? Indeed, one can attribute that to the fact that the being, observer is attributing such terms and meaning to the structure that is interpreted as fungus farm, while in fact in natural it is just a natural behaviour of the ant colony. However, how far will such interpretation stretch, and edge cases of which it can handle is unknown. Where does nature end, and where artifice begin?

Of such basis, of the artificial platform in which we as subject B create, artificial intelligence theory then posit that the notion of intelligence can be represented, interpreted, and constructed using a system called **computer**. Though variations differ of which what is considered computation, how is symbolic manipulation considered, this theory is the philosophy of Computational Intelligence, or Computationalism, of which formed the basis of artificial intelligence research of date. Supporters of the theory include as earliest being Putnam (1988); Fodor (1975); Churchland (1986); Dennett (1978; 1991), and of the 21st century, Scheutz (2002); Dodig-Crnkovic (2012); Gauvrit et al. (2015) and Copeland & Proudfoot (2018). This theory then posits that we can represent intelligence, of the arbitrary definition that it is understood upon, using the language of representation of computer, using numbers and computational processes as for simulation, using algorithms to simulate thinking processes, and else. Of course, there exists pushback against such notion, of which creation of artificial intelligence is not totally configurable using computers, notably Searle (1980a; 1992); Penrose (1989a; 1994); Nagel (2012); Müller (2025); LaForte (1998). To answer if the theory is false or not, and what is the implication of such theory, we have to shift to understanding the second term of the word — intelligence. Furthermore, we also have to realize what is said as a computational structure, and as the artificial construct that we created of the computational framework, how and what will constitute the creation of artificial intelligence in such specific representation.

1.2 Understanding intelligence

Inside the term *artificial intelligence*, there is the word *intelligence*. A normal person will tell you that they are intelligent. But it just so happens that this notion of qualification is harder to define when one participates in the active action of finding it. So, what is it? This is the question we should take in.

We start this section with a series of historical accounts. Shane, Marcus (2007)'s paper *A collection of definitions of intelligence*,Legg & Hutter (2007b), and Masahiro's (2023) *Descartes and Artificial Intelligence*[3] might be a great place to start this, since they provide a non-trivial amount of definitions and attempts already there, which serve us more as exhibition for observant in this section and beginning.

[3]See more in Journal of Philosophy of Life Vol.13, No.1 (January 2023):1-4

R. J. Sternberg ...I prefer to refer to it as 'successful intelligence.' And the reason is that the emphasis is on the use of your intelligence to achieve success in your life. So I define it as your skill in achieving whatever it is you want to attain in your life within your sociocultural context — meaning that people have different goals for themselves, and for some it's to get very good grades in school and to do well on tests, and for others it might be to become a very good basketball player or actress or musician.

D. K. Simonton ...certain set of cognitive capacities that enable an individual to adapt and thrive in any given environment they find themselves in, and those cognitive capacities include things like memory and retrieval, and problem-solving and so forth. There's a cluster of cognitive abilities that lead to successful adaptation to a wide range of environments.

H. Nakashima Intelligence is the ability to process information properly in a complex environment. The criteria of properness are not predefined and hence not available beforehand. They are acquired as a result of the information processing.

P. Voss ...the essential, domain-independent skills necessary for acquiring a wide range of domain-specific knowledge - the ability to learn anything. Achieving this with 'artificial general intelligence' (AGI) requires a highly adaptive, general-purpose system that can autonomously acquire an extremely wide range of specific knowledge and skills and can improve its own cognitive ability through self-directed learning.

Jensen, Huarte, Dearborn ...the ability to learn, the ability to understand, either principles, truths, facts, or common sense, to profit from experiences; the ability to comprehend, or the capacity to reason.

A. Anastasi Intelligent is functionally of multiple components combined.

J. Peterson ...a bunch of stimuli.

Humphreys ...the resultant of the process of acquiring, storing in memory, retrieving, combining, comparing, and using in new contexts information and conceptual skills.

Out of those definitions, there are two kinds of defining the notion of intelligence, we call it the **top-down** and the **ground-up** approach. The top-down line of thought demonstrate, most of the time conjectures, the existence of intelligence as a whole, without finding the actual shell that contains it. If intelligence is *general*, then their implementation follows, but to a sufficient degree, it can be achieved everywhere. It guarantees, partially, of certain school of thoughts the generalizability of intelligence as the ground base to re-create such, which is characterized, often, by current machine learning discipline. This approach beside from guarantees such existence, also has the capability to 'test' a subject of being, 'intelligent'.

This is done by setting up agenda and criteria, of which the current theory serves as more of a black box for the actual 'machine' that contain it, but enough exhibitions fitting those criteria for intelligent. Fortunately, this also sets certain criteria for artificial intelligence to be specified so in the name. The Turing test, which posit different observable properties to be examined, and the Gödel's argument is one of such example in this line of thoughts, theorized by J. R. Lucas (1961), Penrose (1994, 1989), and Benacerraf (1967), similar to the Chinese Room Argument (Searl, 1980). Coincidentally, the notion of **computationalism** is also formed out of this approach. In a sense, it works as the following definition.

Definition 1.2 (Intelligence, top-down approach)**.** *We say that we observe intelligence in any given circumstances, of any arbitrary object regardless of structure, if it exhibits observable behaviours to the environment, the surrounding, the interested space such that can be clarified, and identified, to the nearest high-intelligent specimen (human), to certain degree of operational arbitrariness, of its own activities, properties, and functions. In such case, intelligence is defined per speculative reference point (human) and of criteria that fits the such point (human) model of intelligent.*

The definition in the top-down sense is then entirely subjective from the human perspective.

The *ground-up* approach of defining intelligence is simply the polar opposite: Instead of defining intelligence by criteria, they create machines or models that have intelligence seems to be the emergence behaviour from those model. That is to say, they define intelligence by not defining it but constructing it. Though, this type of approach still requires the intuitive feeling of intelligence to figure out or identify such emerging signs of a growing construct, but it is more or less general, as it does not depend on certain opinion, or fixed high-level criteria to classify it. There are many ways to achieve such insight, either by examining the source of intelligence in high forms - neuroscience on human brains, or by analysing them in a representation form - as modellings, and anecdotal analogue that can be found, and so on.

As of date, no such consensus has been found about the definition of intelligence. As we have said earlier, philosopher refrains from talking of such topic, artificial intelligence practitioners rely on certain intuitive sense and reason to interpret such intelligence definition, and some argue about such notion with terms from the discipline of AI itself. In general, it is led to believe that the overall strategy is to pick on such intuition and work with it, rather than doing much about it. And it aligns well with the philosophy of defining the notion of intelligence.

1.3 Artificial Intelligence

Let us come up with an understanding of the term artificial intelligence. Due Stanford Encyclopedia of Philosophy (2018), it is the field devoted to building artificial animals (or at least artificial creatures that – in suitable contexts – appear to be animals) and, for many, artificial persons (or at least artificial creatures that – in suitable contexts – appear to be persons). However, such definition is fairly limited, and would not capture the essence of what practically can be artificial intelligence. Though, uncovering the current measure of which we make up artificial intelligence, we can come up with a provision definition that fits the current philosophical choice.

For the definition of artificial intelligence, or the construct that supports it, to make sense, we need to evaluate again, from what we have seen, what is even the term. As noted by definition on the notion of *artificial* in the preceding section, being *artificial* mostly comes of from the consideration of evolutionary processes - of which the interaction in the physical worlds, the biological worlds, and overall, anecdotally, of anything that is non-human of its (human) own capability to morph objects into an intended state - this is what normally resided to. Then, artificial intelligence refers to a set of observations, observable qualities deemed sufficiently of all intents and purposes intelligent, by any given constructs that is created artificially so.

This breaks down to the two conceptual ways to talk about artificial intelligence.

Conjecture 1.1 (Artificial intelligence)**.** *Artificial intelligence is the classification for any such object of constructs sufficiently reflects those qualities that fit the standard of intelligence, of which also created* ***artificially*** *of intent and purposes (as reflected in definition of artificial).*

The second conjecture then interprets that, toward the theory of computational intelligence. If such theory is indeed proved to be feasible, then we might have the following core argument:

Conjecture 1.2. *Artificial intelligence refers to (a)****construct(s)*** *- of which consists of the* ***machine*** *and its* ***process****, for such that the machine supports the process to reflects the observed results quantified in one way or another, to be interpreted as intelligence by the construct that is standard for those terms. Those constructs however, are absent, or not, by choice, of the existential facility - or of either a rigid static facility of such - and hence artificially made.*

Arguably, the second conjecture is far more interesting and familiar than the first one. However, the claims, of such, can be hypothesized as perhaps not so ideal generalization. The term artificial intelligence, generically, refers to the comparison between two actual constructs. If the current human - or us - are the ones evaluating certain constructs as intelligent, then it is equivalent to generalize human into a construct on its own, of sufficient analysis such that the comparison can be conducted. If so, then the definition ultimately is reinforced, as for now, to be relative and subjective. Would we be able to find generality in such structures, if of current time we rely solely on our own construct to evaluate the criteria, though the creation that we are making is inherently different?

1.4 THE LANGUAGE MODELS (LM)

One of the main, major example of artificial intelligence application is indeed the formulation of language, manifested in a model. Attempts has been made to trying to understand why language emerges as a proxy of information exchange, either by newer treatments, as seen of Galke et al. (2022); Worden (2025), or per historical developments, as Grimm (1819); Humboldt (1836); Schleicher (1874); Darwin (1871); Saussure (1916); Bloomfield (1933); Hockett (1960); Chomsky (1957; 1965); Goldberg (1995); Fillmore (1988;?); Hauser et al. (2002); Pinker & Bloom (1990); Deacon (1997); Bickerton (1990); Kirby (2001); Christiansen & Kirby (2003); Nowak et al. (2001); Evans & Levinson (2009); Christiansen & Chater (2008); Hauser et al. (2002) and more, that is both specific of the field of linguistic and broader. The ability of forming language is considered one of the many things, up to the capacity that human is capable of, hints of intelligence that human exhibits. There, it is just natural that one of the application since the early onset of AI theory, is to recreate this form of language. Some of the first major applications are, of the onset of the Cold War, the task of machine translation (MT). The first demonstration of MT, the Georgetown-IBM experiment, showed a great promise, with limited ability that was proposed to be increased even further in the future. Though, such development did not end well, and by the time the ALPAC report came out (et al. (2006)) the field of MT has already been hit hard. It, and with the addition of Lighthill (1973) report on AI, ultimately, then officially begun the first AI winter.

Looking back as some of the failures in the theory of natural language modelling, it is perhaps surprising when looking at advancements of **Natural Language Processing (NLP)** has as the successor of such research direction in the prelude of AI research. Using analytical view upon the language, pragmatic approach to 'dictionaririze' the copula of words and sentence structures (word encoding, tokenization, data analysis-like methods), simplification of words meaning, cases, categorization, probabilistic methods (for example, Latent Dirichlet Allocation - LDA, see Jelodar et al. (2018)), such research direction is responsible toward a huge chunk of architectures, creations of 'AI language models' capable of statistically generating coherent texts and language contents, answering question in a sense, and so on, from large availability of data in text form. This is all conducted, while pay no mind into the deep theory of linguistic or the study of language itself; in a sense, a marvellous innovation, perhaps too marvellous. As because of such, some take the basis of the language model for the basis of the consciousness, intelligence emergence concept, and posit that such models, the LM or L(Large) LM, would be the centrepiece of a fully realized AI, and thus, the discussion of AGI and furthermore, ASI. This is reinforced by the series of architectures that enables large-scale advancements, like Rumelhart et al. (1986); Schuster & Paliwal (1997); Jordan (1997); Elman (1990); Lipton et al. (2015); Graves (2012) Recurrent Neural Network (RNN), Hochreiter & Schmidhuber (1997); Gers et al. (2000); Cheng et al. (2016) Long Short-term Memory (LSTM), Cho et al. (2014); Chung et al. (2014) Gated Recurrent Unit (GRU), sequence-to-sequence model as seen in Sutskever et al. (2014), and the most foundational advanced structure of the attention-mechanism neural network — Transformer (Bahdanau et al. (2014); Luong et al. (2015); Vaswani et al. (2017)). Indeed, replicating the behaviour or coherent patterns of human language is a marvellous feat that cannot be understated. Yet, would such claim proved to be too costly, just as we have seen of criticism and empirical evidences that it is not at all omnipotent as it is pushed for?

The main fallacy of such new approach, as will be reiterated many times, is the lack of origin, and the circumstances of Descartes's argument itself. While created such good models, it still cannot cope with logic, a wide range of logic, not simplified logic or rigid, manually designed system of logic. We are unable to determine the capacity of it to understand meaning, or any hint of such concept to exists in a language model aside from some short-lived prospect yet of no proven links to such understanding but statistical grouping. 'Knowing language' does not equivalent to being intelligent, as it is always said. Furthermore, there exists a transition between language of human form, words as they are being written, to numerical encoding and manual rules on such encoding law into numerical sense, that is a problem. It brings up the question of if such models, if only the algorithm that it is, does not even understand the language itself, but is just finding the best possible answer toward the task provided. Such dilemma will have to resolve, if one is to claim language model to be the standard basis of such AI generalization. As of for now, that seems to not be the case, as the cracks are closer to being revealed, and as the potential AI bubble to burst of speculation but failed delivery. Unquestionably, again, such development cannot be understated, and should not be forgotten or relinquished. However, pushing far beyond its weight is not a good idea of such sense either.

2 UNDERSTANDING ARTIFICIAL GENERAL INTELLIGENCE (AGI)

We have understood the general notion of artificial intelligence, in one form or another. It is then naturally that we extend such conversation to the notion of Artificial General Intelligence, or AGI in short. In essence, what does an AGI constitute? The AGI notion relies partially on the concept of **fragmented intelligence**. This was first apparent by the apparatus of the Turing test (Turing (1950)), that suggest quantifying different human capabilities for determining a machine's ability to be 'human', and as if the machine can surpass human in such range. This view is supported in a different form in Gardner (1983) book *The Theory of Multiple Intelligences*, of which again posit that intelligence exists in different forms, and not a singular object of quantification. Given such, Artificial General Intelligence posits that we are able to construct, and would be able to construct, an AGI with all of such capabilities that one can consider to be human, with consciousness, with intelligent learning capabilities, with thoughts, and so on. A smaller camp, yet vocal, posits further that the structure of LLM, large language models, or **agentic AI** (Sapkota et al. (2025); Derouiche et al. (2025); Schneider (2025); Wei et al. (2025); Raza et al. (2025)), will be able to achieve this goal.

Nevertheless, the issue that plagues such notion is that the term itself is not fully understood, nor there exists any given consensus on what is the acceptable form that constitute the baseline definition of an AGI. Technically speaking, AGI can be attributed to the fact that many AI systems are constructed in fragments, of which for example, computer vision, language processing, signal processing, classification analysis, robotic spatial movements' extrapolation, and else, all of which then if can be combined into one, would inherently make a human-like form of intelligence. In between such, LLM, per its role as the language processor, is deemed to stand in between such. However, the definition in such term itself has its own fallacy. Suppose that the AGI AGI_x has multitude of ability that is inherently of its own domain and environment - computers, operating systems, software framework, etc, with full understanding and exploratory sense of such, but has no spatial movement, no sensors and the like that can attribute it of features of which human exhibits in natural sense. Would that disqualify it as an artificial intelligence, or just have to reclassify it into a different environment?

For now, we need to formalize what is AGI actually saying, in context. Or rather, a definition that is informal per its natural topic. What can be intrinsically defined to be AGI. Based on our current understanding, AGI can be defined using the basis of the fragmented intelligence theory, and by the previous statement of AI. Simply speak, of Conjecture 1.1 and Conjecture 1.2, fragmented intelligence begins with considering all construct of which fits certain limited set of qualities, one or more, but distinct. Then, AGI is the plateau that there exists such construct that can fit all qualities of its quantified notion.

Definition 2.1 (Artificial general intelligence)**.** *Based on Conjectures 1.1 and 1.2 and the fragmented intelligence scheme, define a quality basis* $\mathsf{AI}_n = \{A_1, \ldots, A_n\}$ *of arbitrary given qualities specified of an artificial intelligence construct. Then, a construct* $\mathsf{AGI} \in \mathsf{AI}_n$ *is called an artificial general intelligence, if* AGI *can be represented as*

$$\mathsf{AGI} = \alpha_1 A_1 + \alpha_2 A_2 + \cdots + \alpha_n A_n, \quad \alpha_i > 0, i =, \ldots, n \tag{1}$$

That is, AGI *expresses every given qualities of the quality basis, to a given evaluation degree* α_i *associated with each quality. Thus, we say* AGI *spans the entire quality space.*

Here, the definition is arbitrary on purposes, as for many evaluation metrics, testing setup, or different quality that is regarded of AGI, for example, the updated version of the Turing test versus the traditional one, we resolve to such definition by the action of generality. Of course, this begs the question if certain quality basis AI_n^i is better than others, and the answer is yes, but subjectively — since the qualities itself usually cannot be quantified exactly. Nevertheless, we posit that such construct can exist, from the construct's structural system itself, that when mapped of its operation onto the arbitrary basis of choice, yields such resultant observation. There then exists the fundamental problem of such definition — for such artificial intelligence construct to exist, within the fragmented theory of intelligence, then there must then exist internal connections between its components, and the expression toward separated quality's evaluations. Because such processing connector also retains its own quality specification, in such framework, aside from being a hidden, internal component of the model (of which then handle such connection arbitrarily and cannot be evaluated simply), it must then be of certain construct that is present within the basis. Such is usually reserved for large

language model, as it is on the basis of natural language qualities, and is naturally the candidate in such regard as the main focus of core component for determining the feasibility of an AGI. Such framework is then called, the **agentic AI** framework. By certain arbitrary consideration, we can say current AGI, of certain qualities such as the Turing test, is already an AGI, within the above definition. Nevertheless, by such basis, we can set the bar so low that such AGI is unbearably naive, or set the bar too high such that the qualities in consideration is considered infeasible event by 'superhuman standard'. Or by a list of all details, measurable set of qualities with overlapping, yet does not and could not usually take into account of the internal structures of the model and the inherent natural abstraction, or emergence of whatever definition being the arbitrary emergence, is for such given construct. In a sense, we already have AGI. The problem is how effective it is. It is perhaps natural that we re-question the anecdotal concept of LLM, and of the basis, and of the entirety of connector framework, and extrapolate from such — what can be different than LLM and such current understanding, and what is inherently wrong and right with it?

This view is debated fiercely, as we have introduced it before, and also because of its intrinsic nature of the field itself. For the example, the arbitrary of such consideration is a problem as there would imply the problem of inconsistency in between criteria, qualities, and so on, of which has been surprisingly apparent as many times standards and qualities have been modified and pushed forward in the case of determining "What is then even A(G)I?". However, there are also voices aside from such camp, telling us that current structures of artificial intelligence understanding, machine learning anecdotal knowledge, are enough to construct such AGI. Some further argue that the current framework is complete, of which provide exponential growth of such toward AGI in such matter. What is the correct answer to such question? Of this paper, the current stance is no. We then have to present logics and evidences to back such answer up.

2.1 The fallacy of defining intelligence

We mentioned the notion of the criterion of intelligence. However, what should we define it? How should we know to even evaluate it, is a very hard question even that we did not (or unable to) fully realize yet, then what we want to do with it? This question is where a lot of things in the artificial intelligence research was based upon. For example, the (Total) Turing Test in which outlines possible outlook for intelligence, for capabilities that then defines the fields in which we are having nowadays, for example, computer vision for the capability of visual perception, natural language processing (NLP) for the capacity of language, and more. We also have various conceptual criterions in which people have been suggesting about the model of the intelligent being, for example, various set of criterions that outlines and includes even consciousness, some suggest behavioural conditions, some goes for the exhibition of *chain of thoughts*, and some even goes further than that, which is perhaps irrelevant aside from mentioned for example. Overall, it is perhaps a mess.

We still do not know what to come of criteria, or rather, in the quest of producing intelligence, we base ourselves onto it too much. As a species capable of intelligence and more sophisticated notion, we have the basis, and the advantage of being able to examine ourselves. By that, eventually, as the highest example of intelligent being, we use ourselves as standard, for examine, psychology, neurological behaviourism, neuroscience, applied onto the quest of going for artificial intelligence. Hence, there exists the total Turing test, and there exists the conflicts between various definitions and criterion of artificial intelligence. A mistake perhaps has been made, doubtfully so that one did not realize of such. While it is said that AI researcher has been working on, or at least researching on the general notion of artificial intelligence principles, it is, in fact, not so much of a principle, as we did not realize yet that what we are doing is still the act of mimicking ourselves - creating a plane by replicating a bird. By phenomenologically absorb and construct architectures, models on the higher-level surface of what artificial intelligence constitute, the deeper construct is still non-existent. By copying the apparent capabilities of human and related intelligence being, biological rather than not, the core of which those behaviours occur, and facilitate the organs and observations made is perhaps, manifested, simply does not exist. Ironically, while being too strict, wrongfully abhorrent to the fallacy of themselves, and too resistant to changes, symbolic approach got one of the right thing. If there exists intelligence, then it must be *universal by virtue*. That is, you cannot argue that alien from another universe is not intelligent, because they do not satisfy one of the criteria of the Turing test, just because such notion does not exist in such universe.

Historically, it does not prevent people in the field of artificial intelligence to search for intelligence in their bold claims, of which their arguments make clear that they are phenomenologically mimicking the attributes of which is of human, without the substances underneath:

"Once (my intelligent system) OSCAR is fully functional, the argument from analogy will lead us inexorably to attribute thoughts and feelings to OSCAR with precisely the same credentials with which we attribute them to human beings. Philosophical arguments to the contrary will be passé." (Pollock, 1995)

From such view, I objectively don't think we should, or we could define artificial intelligence, at least of this particular stage that we are in. Philosophically, being an armchair philosopher would not help in pursuing such notion, yet again because we are arguing on the basis of our own existence, and not the subject's matter viewpoint. There are problems related to it, also, of such that the mind and consciousness is arguably debatable in every given sense, of which no one seems to agree on the mundane notion that intelligence and consciousness come from chemical and the weird 'quantum effect' that would be then believed to be. And, truth to be told, we are not even endorsing such direction. In actuality, we don't even know what is intelligent, and also don't even know what can be of artificially made rather than matching mathematics.

On the flip side, computationally and neuroscientifically, the lack of formal treatment and overall encompassing knowledge conjunctions plague the construction and foremost attempt to do anything, simply because too many things have been said yet none can unify them together. Such is also to say different directions and different methodologies being conducted, yet they are so distinctively separated to be unable to conform one to another, despite them taking on the same object. Furthermore, there are a lot of assumptions given in computational theory, and the overall application thereof. As for anything, assumptions can be broken, and reinforced, for whatever it is being inconsistent as a virtue.

It is wise to remember that, for now with neuroscience being not advanced enough and in a perhaps different direction from what can be seen, while certainly for empirical science we can utilize neuroscience's knowledge, we should not take in the philosophical arguments and 'idea', including computational theory of mind. For empirical neuroscience, it is also not the fully-encompassing field that observe the brain from every angle, and observe consciousness of everything if ever, at least of the present. And, for the *philosophical* and idealistic view, only one thing can be said about such being "the lines on the map is made up".

2.2 The statistical model critique

AI, as of current, have shifted its structural formations and logical acumen back to mathematics. That is, right now, artificial intelligence looks like nothing but the thing it is originated from, but rather statistical models on data. This view is iterated in several literatures, of which we list of Penrose & Severino (1997); Peyré (2025) and Kutyniok (2022). While not downplaying the role of mathematics in its application and advancements of the field, and successful creations such as many models have been created, one question remains - is it actually artificial intelligence, or just a statistical model trained and probabilistically interpreted to mimic certain aspect or tasks of which is considered intelligent? This is also the stance that Searle (1980a) argued against, of which produced the long-standing Chinese Room Argument. We can summarize the argument simply as followed.

> CRA is based on a thought-experiment in which Searle himself stars. He is inside a room; outside the room are native Chinese speakers who don't know that Searle is inside it. Searle-in-the-box, like Searle-in-real-life, doesn't know any Chinese, but is fluent in English. The Chinese speakers send cards into the room through a slot; on these cards are written questions in Chinese. The box, courtesy of Searle's secret work therein, returns cards to the native Chinese speakers as output. Searle's output is produced by consulting a rulebook: this book is a lookup table that tells him what Chinese to produce based on what is sent in. To Searle, the Chinese is all just a bunch of — to use Searle's language — squiggle-squoggles. (Searl, 1980s)

It is notable to point out that Searle's argument against the current advancement of AI using CRA is particular still effective even in the modern current landscape of artificial intelligence. In fact, this voice resonates with a large pool of people, whether because of the fear of losing identity, or from analytical assessment of current models. Large Language Models and advanced models often

fail miserably when changing context or changing setting, losing information or hallucinating and so on for a large spectrum of situations, for imperfections that lies outside its intended encoding. Coincidentally, this also fit the old-age argument that Descartes made on the notion of machine and human, or simply the Descartes's argument (Descartes (1950)). Descartes's argument start with the justification of the apparent reactive behaviours observed by human themselves, who at the time, was largely considered to be the only species capable of advanced rational thoughts and processes.

> (I)f someone touched it (= the machine) in a particular place, it would ask what one wishes to say to it, or if it were touched somewhere else, it would cry out that it was being hurt, and so on. But it could not arrange words in different ways to reply to the meaning of everything that is said in its presence, as even the most unintelligent human beings can do. [Descartes, 1700]

Here, Descartes argues that in order for human-like robots to acquire intelligence, they have to gain a universal capability to accurately react to any unknown situation that may happen in the environment. However, what machines can do is no more than to respond to a single situation one-on-one via a specific organ, hence, they cannot be considered to have a universal capability that even unintelligent human beings can enjoy.

Continuing, Descartes argues that those machines do not act on their knowledge, but the disposition of organs.

> For whereas reason is a universal instrument that can be used in all kinds of situations, these organs need a specific disposition for every particular action. It follows that it is morally impossible for a machine to have enough different dispositions to make it act in every human situation in the same way as our reason makes us act.

The argument is quite clear. Human is universal of the environment. Whereas machine is no more than a combination of abilities that are applicable only to certain situation that the creator could imagine when they built the automated machine. This simply posit, and be relevant of our observations on the current machine learning models and large language models. While they are large and as a result, of great range of capability, they are inevitably hard-coded with what the designer wishes them to do. They are not intelligent, in a sense, so to speak of their capability as to be artificial intelligence of its true 'general form'. Hence, of this debate, we can simply say, to the disdain of the empiricists themselves, that artificial intelligence of the current structure simply is not distinguishable from the statistical model view.

The same plagues the new-found sensation of agentic AI, of which is believed to obtain general intelligence by cutting and wrapping different small specialized-AI structures together, for example, connecting an LLM to a computer vision system or image recognition for scanning documents. However, by default, such connection is inherently shallow, as there exists only a *black-box connection* of input-output, process resultant in between those components together, which makes it similar to CRA and Descartes's argument in question.

2.3 The intelligence model

Along the same line as the statistical model critique, we now change to the perspective of the current architecture of choice. Assuming the current knowledge of artificial intelligence and thereof, within respect to machine learning and other development fields, can we say that we understand, or at least can construct artificial intelligence, and hence the true form called AGI? The answer is both yes and no.

Current understanding of AI, Russell & Norvig (2009), as specified, focus on the self-reflection of the human researchers on themselves. Such reflection are often surface-level, for example, behaviours of which intelligent choice might occur rather than not, situation of which there exists patterns in which the mind choose to operate, and so on. However, this in particular, face heads-on with a problem that even now cannot be explained or go through - the domain problem, or **the frame problem**. The frame problem is the problem that an AI cannot autonomously distinguish important factors from unimportant ones when it tries to cope with something in a certain situation. The problem arises, for example, when we let AI robots operate in the real world. This problem was proposed by John McCarthy and Patrick J. Hayes in 1969, of which is considered a philosophical problem that cannot be merely reduced to a technical problem. Historically, the problem is narrowly defined for the field

of *logic-based artificial intelligence*. But it was taken up in an embellished and modified form by philosophers of mind, and given a wider interpretation, and hence, is since then applicable to almost all formal system that wishes to call themselves artificial intelligence. We will not cover all of it here, at least for now. For authoritative literature, it is recommended to refer to the Stanford's article on the frame problem Shanahan (2016), and other literatures Gryz (2013); Seager (2010s (or older); Briggs (2014). However, it is indeed a dilemma of which both symbolic Ai and the kind of statistical, variable AI of current form cannot proceed. Even with the structure of neural network and modern deep learning, symbolic encoding, expert systems, current AI structures are what called *interpolator* in the purest form, and not extrapolator. This problem is inherently similar to how we would interpret the problem of *out-of-bound* cases are, as seen in Bahng et al. (2022); Xu et al. (2020); Traber et al. (2020); Hüllermeier & Waegeman (2021); Amini et al. (2021); Sensoy et al. (2024); Jose et al. (2022) of a wide range of such notion. Additionally, there are many problems that would not have answer and thereof, for example, the problem of *hallucination* apparent in typical LLM settings, symbolic grounding problem in Harnad (1990a), and so on. The problem is not with identifying the problem and observing it — the problem lies mostly in the form of *inexplainable phenomena*. To do such, we have to inquire further into what is being done in the current theory landscape.

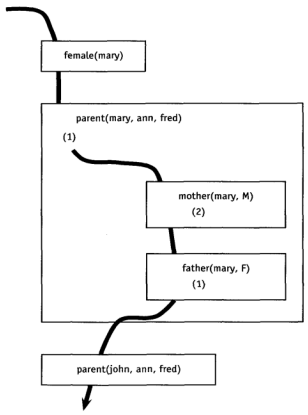
2.3.1 Artificial Intelligence Theory

The current artificial intelligence theory can be branched into several aspects, as seen in Jackson (2019); Goertzel & Pennachin (2006); Chowdhary (2020); Russell & Norvig (2009), and more. Those considerations of the field are what we can call as both phenomenological inspired systems, and general modelling theory. We can list a few of them, for example, Shortliffe (1976)'s MYCIN structures, Newell & Simon (1976) works on symbol systems, Collins & Quillian (1969) semantic memory, Davis et al. (1977) knowledge-base representation, and Lindsay et al. (1993) DENDRAL expert system for working functionals of chemical compounds. There core driven such developments are the artificial intelligence idea focus on the reasoning process of the brain, using either pure logic (Logicism), Expert Systems, Non-monotonic Logic, Planning Systems, Argumentation, Semantic Network/DL, and Modal/Temporal Logic; more can be found in Liang et al. (2025).

Theory like knowledge representation, agent structures, algorithmic searches, decision theory, rule-based or relational network-based reasoning theories, first-order logic, specific domain of thoughts of computationalism, symbolic AI, or neurosymbolic unifications, while proved to be effective in a sense, those theories do not resolve the origin problem for those behaviours, and furthermore is restricted in the same restrictions that we stated as above, of AI models in generality. Logic in the construct of symbolic approach deliberately catches surface logic, representations, knowledges, properties and attributes, and assigning strict logical conclusion together for the logic itself. This falls into the range of Descartes's argument, in which one's machine deliberately can only follow and operate of its designer's configuration, nothing more and nothing less. The restriction of a decision tree can also be founded to be limited in such case, for there are limitless consideration and fluctuated information in a given setting. Furthermore, a decision network on itself does not have any meaning. For example, the programming language `PROLOG`, which is prominent of such framework design (for reference, see Clocksin & Mellish (2003); Colmerauer et al. (1972); Colmerauer & Roussel (1993) for details on the standard of `PROLOG`), different facts are encoded manually by *atom*, a unit of logical fact in the database, and interpreting symbolic tasks being the goal, of which `PROLOG` then traverse the symbolic graph encoded to find the solution, of which also return, in the same sense, an encoded answer.

The program, while succeed, only works in its own environment. It is simply a program with encoded ruleset, of configured system of objects, and would rather be classified as an algorithmic program than artificial intelligence. Such parent-child relationship only makes sense in the eye of the designer, yet questionably nonapparent of the program itself. It is similar to the CRA, in which symbol manipulation and fact encoding only get it thus far, without any sufficient notion of understanding, for the arbitrary definition of such term is defined. Furthermore, applications using such system is severely limited at scale, for manual addition of facts, expert systems (from expert knowledge and so on), of which add into the rigid mimicry, and many more design initiatives.

When the focus switch to another constructive structure of connectionism, of which is utilized in statistical and data-driven frameworks, models like probabilistic logic, statistical relational learning, Markov/Bayesian process network and logic, Causal Inferences, Deep Neural Networks (DNN), and so on, they also struct the same disadvantage as discussed of the majority of symbolic-based



(a) Resultant process.

```
female(mary).
parent(C, M, F) :- mother(C, M), father(C, F).
mother(john, ann).
mother(mary, ann).
father(mary, fred).
father(john, fred).
```

Let us look at the sequence of events that leads to answering the question:

```
?- female(mary), parent(mary, M, F), parent(john, M, F).
```

(b) Program.

Figure 1: A typical example of a program and its logical process in PROLOG. The example requires determination of a parent-child relationship, with encoded knowledge graph, and traverse of the predicate logic. (a). Canonical example of a PROLOG predicate decision network, in an example of Chapter 2, section 6 in Clocksin & Mellish (2003); (b) The program responsible for the respective sequential predicate logic, of which answers for the question accordingly.

program, sometimes worse. In terms of mathematical modelling, Vel (2024), such models in statistical framework and on are interpretations of a *phenomenological model*, or *black-box model*, in which data and observations are encoded as statistical points for interpolation, either soft (for $\epsilon > 0$, the disparity between the wrong answer and the correct one is not required to be reduced to zero) or hard (the condition explicitly said to reduce to absolute 0, or point-fit), without any knowledge of the underlying mechanical details. Hence, they are usually nothing but statistical fit of particular problems, for example, classifications, regressions, predictions, and so on of numerical data encoding. Sometimes, the encoding range and model structure is extended so hard, that the model seems too real, as seen of LLM, yet is not real, by the observable statistical-inherent errors and limitations. A limited observation can be said, as in Song et al. (2025).

The flaw of the current artificial intelligence theory can simply be attributed to looking at the wrong way, of which while proved insightful, is misleadingly taken as the way forward of making an intelligent construct. Approaches and constructions listed above relies on the basic idea of mathematical modelling, the modelling theory, yet such theory is underdeveloped, hence the true nature of the modelling scheme, of the generality on the theory of the representation language of which represents and is used for descriptive construction and analysis of models, often not realized. Such is also seen of many theories that is of apparent importance to artificial intelligence advancement, yet is rather forgotten in pursuit of the state-of-the-art phenomenological models. Different conceptual understanding, while making sense and is rigorous of logic and foundation, often found itself in trouble of interpreting the strong argument about origin of such system, once ask of the designer to make the model learn of the concept by itself. As of now, only in the sense of mathematical reduction, data points reduction that the notion of learning is then realized, and yet such is insufficient to be called intelligent, but intelligent algorithm. There exists no unifying or general theory of which is concerned of the topic of encoding, representation, architectural design and implications, operation theory, percolation theory and systematic emergence and thereof, which makes any inquiry into the topic harder than it should be. Current theories also use their limited systems to make bold claims, for example of manifesting consciousness, emergent behaviours, models as 'living' while such criteria of living is not fully realized[4], making it inconsistent, and created a skew trend toward practical, empirical construction but no theoretical understanding. Again, it does not prevent us from making use of it, for applications on such first ever dynamic system modelling is enabled, opening a wide

[4]Each pass to the model is a statistical or logical process, of which is almost algorithmic-based. This is arguably living, since they are active only by designed activation, patterns of activation that is concrete, and of operations that would not be made apparent of its meaning, for meaning is not strictly defined.

variety of different structural encoding and thereof. But even then, the theory of such system itself is insufficient, of limited depth, or coherence that is typically seen of a matured field.

Other than such, treatments of empiricism on such artificial intelligence theory, either by reconstruction of the brain for neurophysiology and neuroscience, empirical constructions and continual improves, such as current model development of AI systems, remains deeply shrouded in unanswered questions, inexplainable behaviours, constrained results or returns, high computational cost of inefficiency, scaling issues, and many more. Of such, it seems questionable to continue pushing toward that end, as the theory has exhausted a fair lot of its development, without gaining much insight or exploratory evidences into the deeper questions about AI, to ever reach the node of AGI. On the philosophical side, many has argued against the notion of computationalism, whether computers can actually be used as the foundation to create intelligence form, and understandably yet unsurprisingly the current theoretical treatment cannot have an answer toward such argument, other than keep working on it until a roadblock is hit.

2.3.2 THE LEARNING THEORY

In all of artificial intelligence theory foundation, one if not the most important aspect, and hence field of research that now dominated the field, is the theory of learning. Inherently, from the onset, the particular intelligent behaviour that one can instantly attribute to a construct to be called intelligent, is the ability for it to learn. The notion of learning is difficult to define, hence current theory seek to determine the mechanical equivalent in mathematical form instead, of which is expressed using the current **machine learning theory**, Valiant (1984a); Wolpert (1996); Shalev-Shwartz et al. (2010). In between the learning theory, there exists the fundamental main side, one of which is concerned with pure logical encoding, or rather learning machines of symbolic logic with applications in a wide range of different fields, as seen in Newell & Simon (1956); Newell et al. (1959); Winograd (1972); Michalski (1969), Minsky (1961; 1968), and Quinlan (1986). On the other side, is the looser end of the learning theory, which seeks phenomenological learning through enough inferences onto data, or rather, information-based learning, attributing learning to pattern recognitions on **observables** of specific problem set, as seen of Solomonoff (1964a;b); Gold (1967); Newell et al. (1958); Cover (1965), and its formal treatment in theory by Vapnik & Chervonenkis (1968); Vapnik (1999); Vapnik & Chervonenkis (1971); Valiant (1984b); Angluin (1988; 1989); Board & Pitt (1992); Hajek & Raginsky (2021); Mohri et al. (2012); Shalev-Shwartz & Ben-David (2014).

Learning theory is regarded strongly as one of the most important part of a model in the modern modelling landscape. In essence,

These two are fundamentally different approach, yet of the same goal of learning, or adaptation of reasoning and *unseen situation* of which then previous knowledge can be applied in certain way, or by extrapolate new information using a copula of approach, either by observation or by deductions. Despite their success, symbolic learning is plagued with its own originality — it cannot be scaled effectively, it is too rigid for such understanding to be made, and it is inefficient in handling truth or logic outside its rigid domain. Statistical learning, or rather learning by data in its pure form, relies on the model of *black-box* modelling approach, usually not totally by can be seen of as the *grey-box* modelling, which has its own problem about interpreting the concept wrongly, surface-level phenomenological approach which leads to insufficiency of mechanical details that underlies the concept that it learns, the reliance on data and large scale observation sets, while being scalable remains deeply mystery, of which cannot be explained (so-called inexplainable AI/DL), unanswerable questions regarding the different phenomena and computational process, results and performances, additionally structural interpretation, and so on. Furthermore, the theoretical treatment itself is fragmented, with many theories competing for interpretation and explanation in the wide spectrum, and with many different non-unified outlook of the same system dynamic, for example, McAllester (1999); Alquier (2008); Haussler et al. (1996); Jeon & Roy (2025); Abreu et al. (2025), and so on. For the large majority of current theoretical treatment of machine learning, it is stuck interpreting a fixed domain of result and setting, reliance on interpretation of the statistical parameterized model θ, and using learning strictly in the domain of optimization theory only. Newer frontier in architectural researches, such as modern deep learning Marcus (2018); Goodfellow et al. (2016); Demuth et al. (2014), proves tremendously difficult to fully understanding using such theory, as for many of unanswerable questions were born from such architecture alone. The ergonomic and logistic of a unified theory of learning, either both symbolic (experts' voice of unifying both phenomenological learning with symbolic knowledge) has become apparently so large in the eye of practitioners that

some inherently try to create new theory of interpreting the learning framework, or push further in some directions regarding such, of which then again, mimic the 1970s-1980s age of artificial intelligence discovery, where many interpretation, exotic AI systems, and prototypes exist far and large.

It is undoubtedly successful of advancements and technological leap that the current theories landscape has provided with regard to artificial intelligence knowledge, for example, of figuring out a system better in some regard, human counterparts. Yet, it is clear to see that such theory is not capable of handling various degree of new problems, of new consideration, and of new standards of which has been hidden from the main argument of AI development, the question of interpretability, the question of originality, many questions of behavioural concepts, such as the question on "common sense learning" (while it is foolish to study this at this stage). It is clear that we either need a unification of all such learning concept, and a new development of a more powerful learning framework, or to jump the bandwagon entirely; or something else.

2.3.3 Architectural Insufficiency

Architectural insufficiency is the aspect in which we consider, the architectural design that is used in creating and constructing the construct that is evaluated for intelligence. Within its development history, AI has received numerous structural designs, some of which more successful than others. A fair share of our issues have been directed to the symbolic, logicism camp, so for now we would like to inquire on the matter of the statistical camp — machine learning models, and deep learning framework.

While classical frameworks have been noted of their own limitations and thereof, modern neural network structures, as deep learning architectures are much richer and much more interesting of such concern. For now, let us elaborate with the points on the architectural insufficiency observed of neural network, with respect to a provisional view toward a construct that is ideal of intelligence.

In essence, the current neural structures are concerned of the operation of neurons together in a unit-wise fashion. This started way earlier, with precursor constructions notable as Rosenblatt (1958); McCulloch & Pitts (1943), and more. In modern land scape, scaling such system is often done in **layers**, of which make up of the many smaller neurons organized and sequentially activated, i.e. feedforward, where individual layers sequentially feed its outputs to the inputs of the next layer. This is usually called layer connections, in some cases, and hence if two layers are connected such that each neuron $n_j \in L_2$ is connected to every neuron $n_i \in L_1$, we say the network is *fully connected*. Such operation is then repeated, and its learning behaviours controlled by the algorithm of backpropagation, of which propagate gradient errors network-wide and perform numerical correction, such that there then exists a correct solution to the given problem of the optimal configuration of the neural network. Then, at such stage, one simply fix the neural network, and let it run on the problem landscape with great accuracy. The most basic form of such architecture is the multilayer, fully-connected neural network, or multilayer perceptron (MLP), for K-layer, $m^{(0)} = d$ and $m^{(K)} = 1$ for m the width of the network, and d the shape of the input vector.

Definition 2.2 (Standard multilayer network, Zhang et al. (2023))**.** *We define a K-layer fully-connected deep neural network with real-valued output. Let $m^{(0)} = d$ and $m^{(K)} = 1$ for m the width of the network, and d the shape of the input vector. We then recursively define:*

$$x_j^{(0)} = x_j \quad (j = 1, \dots, m^{(0)}), \tag{2}$$

$$x_j^{(k)} = h\left(\sum_{j'=1}^{m^{(k-1)}} \theta_{j,j'}^{(k)} x_{j'}^{(k-1)} + b_j^{(k)}\right) \quad (j = 1, \dots, m^{(k)}), \quad k = 1, 2, \dots, K-1 \tag{3}$$

$$f(x) = x_1^{(K)} = \sum_{j=1}^{m^{(K-1)}} u_j x_j^{(K-1)} \tag{4}$$

where the model parameters can be represented by $w = \{[u_j, \theta_{j,j'}^{(k)}, b_j^{(k)}] : j, j', k\}$ with $m^{(k)}$ being the number of hidden units at layer k; $\theta \in \mathbb{R}^m$ the weight of the neuron.

Such structural definition, of which use the unit abstraction of neuronal units, and the dynamic consideration of connections between parallel layer computational, allows for a variety of different architectural designs and application of the framework itself. Some of such includes the aforementioned neural networks in charge of revolutionary works in NLP, Graph Neural Network (Oono & Suzuki (2020); Scarselli et al. (2009); Hamilton) for graph-like structural data, physics ODE-based neural network like Chen et al. (2019) and so on, with numerous applications.

The first problem that come with this type of standard architecture is the expressive problem. Standardly speaking, both classical modelling, symbolic computation, and continuous/discrete Markov chain process[5], relies on the perspective of **function construction/approximation**, that is, we assume that, there always exists, of the best-case scenario and the internal machine assumption, the concept $c \in \mathcal{C}$ can be expressed of the function space, that is, a collection of function that can be approximated, reconstructed, and so on. This ranges from symbolic process function, transition functions, time-evolution function given a system state S of variables and parameters, or simply variables-related functional relationship. As such, structures and solutions on such space relies on reconstruction and approximation on various metric $d(f, c)$ that gauges the ability to approximate the learning objective in such numerical approximation landscape. Expressibility is then required, as the question of "what kind of function class can encompass the entire concept space, and can there be a universal approximator?" for such various structures. For neural networks defined by definition 2.2, such is usually guaranteed, over certain class of functions, by the Universal Approximation Theorem, first proved by Cybenko (1989) for sigmoidal network, Hornik et al. (1989) for general activation function class, non-Weierstrass (polynomial) activation in Leshno et al. (1993), and partially, of probabilistic version on L^p space in Hornik (1991). There are many statements and frameworks of such UAT is considered, though, UAT on neural network is considered in two different ways, as for two different architectural build-up of the network - arbitrary depth (layer), or arbitrary width (layer's neuron). However, in general, it states:

Theorem 2.1 (Universal approximation theorem — UAT)**.** *Single hidden layer* $\Sigma\Pi$ *feedforward networks can approximate any measurable function arbitrarily well regardless of the activation function* Ψ*, the dimension of the input space* r*, and the input space environment* μ*. That is, for every squashing function* $\Psi : R \to [0, 1]$ *of which is non-decreasing and* $\lim_{\lambda \to -\infty} \Psi(\lambda) = 0$*, every* r *and every probability measure* μ *on* (R^r, B^r)*, both* $\Sigma\Pi^r(\Psi)$ *and* $\Sigma^r(\Psi)$ *are uniformly dense on compacta in* C^r *and* ρ_μ*-dense in* M^r*.*

a perhaps simpler statement can be as followed.

Theorem 2.2 (Universal approximation theorem, simplified)**.** *For a class of functions* $\mathcal{F}$ *and a compact set* $S \subset \mathbb{R}^d$*, if for every continuous function* g *on* S *and for any* $\epsilon > 0$*, there exists* $f \in \mathcal{F}$ *such that*

$$||f - g||_\infty = \max_{\mathbf{x} \in S} |f(\mathbf{x} - g(\mathbf{x})| \leq \epsilon \tag{5}$$

Then, the class of functions $\mathcal{F}$ *is a universal approximator of all continuous functions on* S*. We then indict that* $N(w, d)$ *of the neural network structure, derived from definition 2.2 is such universal approximator on the set of all continuous function on* $[a, b]$ *of arbitrary measure* μ*.*

While not determining correct optimal procedure that can lead to such result, or arbitrary properties of fitness, stability and so on that is useful for operational process, UAT still determines a weak *existence theorem* — such that in said domain, there exists the fundamental optimal model that can reconstruct partially the structure of the concept objective of the approximation task. Here, we still notice that we assume the observables of facts come in the form of function, and only function. Nevertheless, in particular, UAT allows for the guarantee of its expressive power over such class of function as the target. As such, many developments, of which derives from the class of optimization technique on function encoded space — numerical encoding on $\mathbb{R}$, landscape optimization such as gradient

[5]Indeed, in the case of Markov process, a (discrete-time) Markov chain defines an operator

$$\mathcal{T} : f(x) \to \mathbb{E}[f(X_{t+1}) \mid X_t = x]$$

on some space of function $f : \mathcal{X} \to \mathbb{R}$. That is, every Markov chain can be seen as defining a linear operator acting on a function space. Such operator is typically known as the Markov transitioning operator, or Koopman operator in dynamical systems.

descent, coordinate descent (Luo & Tseng (1992); Tseng (2001); Friedman et al. (2007); Wright (2015); Tseng & Yun (2009)), extension of such into the process of backpropagation (Rumelhart et al. (1986) and older historical development). Because our argument is about architectural insufficiency, one then can ask what is the drawback in such framework?

First and foremost, UAT and the functional structure deliberately choose and restrict itself around smooth, continuous, well-behaved function. Not taking the impossible domain lies beyond the compact subset, many functions or encoded function cannot be determined, or can be approximated partially to a given arbitrary $\epsilon > 0$, yet not satisfactory. Variations of UAT usually can be considered weak, with some stronger theorems deliberately require the use of stronger assumptions and more specialized constraints. Any given different setting, structures, data type and the encoding space of such concept, for example, as graph theoretical data, requires to be encoded or embedded into the continuous differentiable pipeline of a neural network to be utilized, and in the process making the structure sub-optimal in certain standards. This result in the structure of Graph Neural Network, or GNN, of which use an encoding scheme, and a typical block aggregator over local graph information, for example, as

$$\begin{aligned} m_v^{(k)} &= \psi^{(k)}\Big(\{\, M^{(k)}(h_v^{(k)},\, h_u^{(k)},\, e_{uv}) \mid u \in \mathcal{N}(v)\,\}\Big)\,, \\ h_v^{(k+1)} &= U^{(k)}\Big(h_v^{(k)},\, m_v^{(k)}\Big)\,, \end{aligned} \tag{6}$$

Where e_{uv} represents the edge features, if any, $h_v^{(k)}$ is the embedding of node v at layer k, $\mathcal{N}(v)$ denotes the neighbours of v, $M^{(k)}$ denotes the message function producing messages from neighbours, $U^{(k)}$ denotes the update function, and $\psi^{(k)}(\cdot)$ denotes the message aggregation up to layer k or such node in consideration. Even though GNN of such 'rigorous' in a sense framework can be used in a variety of sessions, being restricted to such neural network framework brings extraordinary drawbacks. For example, the expressive power on the native graph data is at best, equivalent to **1-dimensional Weisfeiler-Leman (1-WL) test**. Such message passing scheme often fails because of the bottleneck problem intrinsic within the framework requirement and complex backpropagation adaptation, as seen in Alon & Yahav (2021). Further problems like structural information losses, scalability problems, trainability issues, over-smoothing, ontological structural mismatch, and so on, as seen in Maron et al. (2019). Internally of the neural network framework, such problem is not restricted to such adaptation, which requires both complex restructuring and architectural mapping, but also problems like the parity or memory-sample problem in statistical query model (Blum et al. (2003); Garg et al. (2021)), planted clique problem (Barak et al. (2016)), worst-case analysis indication of neural network being NP-complete or $\exists\mathbb{R}$-complete (Blum & Rivest (1992); Abrahamsen et al. (2021)) for low-level, 3-node, 2-layer configuration, computational problem and its expressibility (Eldan & Shamir (2016); Telgarsky (2016); Siegelmann & Sontag (1991); Weiss et al. (2018)). Perhaps more interestingly, is the actuated and controversial No Free Lunch theorem from Wolpert & Macready (1997), of which basically states that a "structural-less", or naively defined structural copulative information, construction on the encoding scheme, is equally bad as with any given model of arbitrary sense, and also the learning procedure associated with such. Setting this up with respect to a particular set of preset inductive bias, or structural assumptions thereof, scaling exorbitantly both of components, parameters, and repeated structures like typical deep learning structure, is not feasible and reduce its success to uncertainty, as the stronger, relative-NFL theorem takes effect — given arbitrary structural information as the origin, there then exists, particularly, certain point of which without further information introduced to the setting, the supposed 'information' can be reiterated as misleading, thus average out the model's evaluation and performance.

Aside from some archetype of learning, for example, online learning of which keep dynamically update the model in a sense, generally, neural network and such type of parameterized models are *static*. They act, just as CRA indicted, to be an algorithmic machine only, of which fits to a description, to a task, to a given degree of accuracy within certain manually designed encoded space, manually designed objective, limited or no interpretation of the scheme itself (classification machines think in numerics and matching errors, thus would not know what a cat is, but only know that an object named `0x0AB4` must be evaluated in some way that can be optimized and reduced). The operating stream itself of the neural network is limited, as we can say of the feed-forward scheme. The extension that neural network is configured to is limited, and lest to say that the neural network scheme itself is underdeveloped, ultimately relies on the simplest possible modelling structure to then give rise to a multitude of different architectures of date. Computational optimization and

layer-optimization, practically constrained structures further hampers down such understanding of the neural network itself, as nobody knows what happen in the neural network, less to try adapting it to different situations.

Go for technicality, neural network architecture of itself faces several increasingly difficult dilemma. Let us disregard the problem of Neural Scaling Law, Kaplan et al. (2020b), as it is simply an empirical optimization observation hyped of its wording[6], the theoretical ground on which neural network is born from, and the empirical method it employs to advance ever since, face problems regarding interpretability, structural effectiveness or definition, phenomena explanations, uncontrollable behaviours, black-box restrictions, rigidness of architecture, uncertainty and vanishing problems, overloading problems, and so on. Interpretation is the largest problem with such theoretical and both heuristic treatment, as no one understand what lies underneath such system and the operation that births such result observed or tasked itself. Explainable AI, like presented in Hsieh et al. (2024), was created to counter such issue, still cannot work or cannot make substantial advancements aside from careful manual designs and limited domain analysis, which in an unfair bit of comparison, go back to the time of symbolic AI approach. This is even worse for neural network, as the regular wisdom is that no one understands the operation of hidden layers and all, in any given setting, aside from which you reduce it to very small size like a perceptron or so. Theoretical development toward such is also slow, and often simplify it to mathematical object to be analysed too rigorously, losing the essence of the architecture in the mathematics rigours. This is particular event or rather pattern that we unfortunately stretched to the previous section on learning theory. Often see in learning theory development, or machine learning theory in general - the '**mathification**' of a theory is a problem that somewhat plagues papers and researches on machine learning topics. Even though we need mathematics on either end, the approach is impractical or simply wrong, believing mathematics to be the singular thing that defines learning theory. This has many fallacies that can then be attributed to a lot of factors and whatnot, of several factors that plague this analysis even further than just the problem of double descent. For now,

1. Epistemological limits and interpretations (Barbierato & Gatti (2024)) - We ultimately lack understanding of a lot of things. While those 'theorems' are very nice in learning theory, the real picture is that it is not real learning, for the word learning are not even defined, as such is to compare them to *human learning* on itself. Even by then, theorems are severely limited. Certain voices also concern of similar problem, including Lipton (2018); Doshi-Velez & Kim (2017); Molnar et al. (2020) (Molnar et al. (2020) shortly consider the misleading interpretation question instead), and philosophically, with several pushbacks on structural anecdotes, Dreyfus (1965; 1972); Dreyfus & Dreyfus (1986); Suchman (1987); Brooks (1991); Searle (1980b); McCarthy & Hayes (1969); Harnad (1990b). On the more modern side, of contemporary critiques, Pearl (2009); Marcus (2018); Sutton (2019). Of the Chinese Room Argument, perhaps we can look into already copulated passages.

2. Applying wrongly, and is used to impress and not to explain anything (Lipton & Steinhardt (2018)) - mathematics is used to *impress* certain demographic of reviewers and readers, to provide a sense of rigours, to further enhancing the image of formal theory to the point that such theory, even if wrong, can be considered fairly correct by the sheer volume of practitioners believing in such. Such is to say the *mathiness* is turning things into ideology more than rigours itself, of which we can take a tangent to see in economic theory, one of the place to adopt a large portion of machine learning statistics, the pushback against such

[6]Of the Neural Scaling Law, conceptually, the result is purely empirical on deviation from regular norm, which is expressed by

$$L(N, D, C) \propto (N^{\alpha} D^{\beta} C^{\gamma})^{-\eta}$$

Where L is the loss, N being model parameters, D is the dataset size, C being the compute used, and η is an empirically fitted exponent. Plotting such relation into logarithmic scale gives roughly the linear relation that is typically seen in literatures and news context.The apparent 'law' only emerges because it is plotted on log-log axes. Almost any monotonically diminishing function can be made to look linear on a log plot over a limited range. The law itself also only empirically states such relation without the underlying mechanics or assumptions that would make it a technical 'law', and is purely empirical observation. In practice, achieving near-zero error is fairly normal, and within such structure, there exists no framework to indict that the neural scaling law holds any value aside from interesting engineering heuristics. Indeed, perhaps, after it is published, many papers were born dedicated to question or beating such argument, such as Sorscher et al. (2023); Ivgi et al. (2022); Su et al. (2024); Lee & Dieng (2025), and Bahri et al. (2024). Therefore, it is not worth it of such analysis.

cursed devolution (Romer (2015); Syll (2024)). Such can also be seen in generally most double descent analysis, in which there exists many formalisms yet not definite result on double descent.

3. Reproducibility and acute false claims (Kapoor & Narayanan (2022)) - In general, what we have done cannot be recreated, in certain way, and of certain too optimistic setting that is seemingly unrealistic - particularly in adoption toward practical means of actions. Certain theoretical assumptions and formulations are too stricts, of which means the increment of hypothesis fluctuation with removal of such constraints.

Continuing, practical problems range from grokking (Power et al. (2022); Davies et al. (2023)), double descent (Belkin et al. (2019); Nakkiran et al. (2019)), triple descent (d' Ascoli et al. (2020)), counter-intuitive results (Szegedy et al. (2014)), adversarial examples or statistical instability (Goodfellow et al. (2014); Carlini & Wagner (2017)), 'memorization issue' (Arpit et al. (2017)), generalization problem and its purposes (Zhang et al. (2017)), catastrophic forgetting (Kirkpatrick et al. (2017)), and so on. Terms like inductive bias are thrown in of many contexts without a single grounded definition or notion to support such, many terms where born without actual consideration, concepts are thrown in and out without basis. While it is easy to simply ignore such problems, such as to be seen in usual deep learning practices of AI, the crack inevitably shows, and a very wide range of developmental gap ensues.

2.4 The end goal of AGI

With that said, what is the end goal of AGI? In all of it, the central goal is supposed to be the creation, out from the doctrine of fragmented intelligence, a unification in which we can then centralize everything in a singular artificial intelligence system, hence AGI. However, as we stated, the term AGI can be further dissected to subsets of such AGI construct, and even then, different system has different AGI threshold. Thereby, the end goal is not clarified yet. What we are seeing instead is a wave of hysteria, from people afraid of the incremental development, and the current market bubble generated from the present AI boom, all of which in turn, pushes the notion of 'building AGI' further and further to the truth. Such is also amplified of the outlook to the concept of ASI, of which is sold in the general media as the next stage on which post-scarcity can be achieved, where Nobel-level researches are outputted every few days, and automation leaves human of no burdensome tasks. However, with what is being done right now, it seems like such outlook cannot be realized, at least in due time.

The conflict in which the philosophy of which promised AGI to the public, as well as the optimistic outlook of people and thinkers, perhaps can be reflected in the sense of Heideggerian philosophy, or famously so, the philosophy between **ready-to-hand** (Zuhandenheit) and **present-at-hand** (Vorhandenheit) (Heidegger (1962); Dreyfus (1991)). Heidegger simply explains, of the notion in which the object appears in front of us. In a sense, the subject of question, and the person of inspection, of all but anthropocentric. For Heidegger, most of the time we are involved in the world in an ordinary way or "ready-to-hand." We are usually doing things with a view to achieving something, and hence, with a purpose already configured, and project such voice outward. The being of the ready-to-hand announces itself as a field of equipment to be put to use, and hence ultimately defines itself around such action and potential. A famous example of such notion is the hammer analogy. In question toward what the hammer can do, whether because from intuition or else, we extrapolate such object of its functionality, instead of inclining on the more sophistry of decompositions and analysing components. Such knowledge can then be made trivial, or thus transferred to a lower priority of thought, of which to the point where one can talk comfortably of other subjects or adjacent topics, but do not require careful contemplation to the original consideration (the hammer). The reversal of such action happens in isolation, in a typical understanding of which decomposition is required, structural knowledge needs to be formed, and so forth, such is to say we purposefully study such object similar to a scientist in idealization studies carefully of certain object, and hence, defers its meaning and existence to the notion of it "exist" to be there. That is the idea of present-at-hand.

It is then fairly connective to approach from certain angle, that concerning Heidegger's notion, our AGI system, plus the non-elucidating way of studying the architecture of choice for such endeavour of AGI, we deliberately refuse the notion of present-at-hand for a more ready-to-hand, in some cases to the extreme of *empiricism*, of which then also founded the principle of agentic AI system. This can be applied to even further, as to the normal principle and conceptual architecture of artificial intelligence

system in an agent-environment scheme as being inadequate, correct, but not enough (Stanford Encyclopedia of Philosophy (2018)). There will be, ultimately, a wall in which this approach cannot reach, as to there exists too many permutations of a path such that there exists fundamentally, of a statistical approach, to be an almost zero chance in which such can happen, within the wrong mode of understanding, both from the perspective of the designer, and the perspective of the construct itself. In the same spirit of the UAT theorem, we can say that the optimism will be, to a certain point, similar to saying "Somewhere in the space of all possible English sentences, one perfectly describes quantum gravity.". Interesting and true (probably), but useless.

3 FUTURE OF ARTIFICIAL INTELLIGENCE

3.1 NEURAL ARCHITECTURE FORMALISM

We would argue that the neural network idea is indeed, formal and foundational, more thatn it is usually attributed of. Formally, a neural architecture follows inspiration from the biological neuron in the human brain, and thus employ the philosophy of unit-based processing (Rosenblatt (1958); Minsky & Papert (1988); McCulloch & Pitts (1943)). This implies various properties. First, they are categorized conceptually into units of processing, of which all neuron n admits the structure (I, M, O), of the input receiver I, the internal processor M, and the output transceiver O. Any neuron admitting this structure then can be constructed, combined, and fully realized at will. Conceptually, this created the *typed* of the neuron architecture, in which all structures would have the same type for operations, such as composition of two neurons, $n_1 \circ n_2 = n_2(n_1)$, in this case connected to each other by sequential move, for all neuron to have the sequential operation $n_i : I_i \times M_i \to O_i$. In practice, handling this might require more careful planning, but nevertheless the structure is particularly streamlined. Another property that is implied is the recursive structure that can be employed. Intrinsically, the admission of (I, M, O) structure implies that any given structure can mutate M for different purposes, for different processing and thereof, as long as I and O stays as immovable component of the neural structure. Then, Any nested sequence of neuron can be compressed to be a singular neuron, since the footprint of the operation $n_1 \circ n_2$ is simply

$$n_1 \circ n_2 \equiv n_3 : (I_1, M_1) \to O_1 \to (I_2, M_2) \to O_2 \equiv n_3 : (I_1, M_3(M_1, M_2)) \to O_2 \tag{7}$$

in which we clarify M_3 as the processing equivalent of both the first and the second neuron. Thus, we can nest many structures together, changing dynamics altogether, and create different type of specialized neurons, yet with only careful planning of the pair (I, O) we can operate on them together at will. In fact, one can simply also connect as many input from n_1 to n_2, neglecting the rest, and the neural structure will still work. We say that $(I,)O)$ represents the **signature** of the neural structure. A baseline, minimal neuron structure can then be defined, which fits the basic definition of a singular perceptron in theory. Let us define $\mathcal{N}_i$ as the ith arbitrary classification of neuron class. We define the criteria of minimization as

Definition 3.1 (Minimization set)**.** *Let x be a neuron of arbitrary neuronal classification $\mathcal{N}$. Then, the requirement of all neuron class is to be able to distinguish its component to three parts, that is, $\min_{\mathcal{N}_i \in \mathcal{N}} \mathcal{N}_i \equiv \mathcal{I}, \mathcal{M}, \mathcal{O}$ where $\mathcal{I}$ is the input channel, $\mathcal{M}$ the internal mechanics, and the output $\mathcal{O}$. Let i, j, k represents the cardinality of each part respectively, then if*

$$i = j = k = 1, \quad \min_{\mathcal{N}_q \in \mathcal{N}} \mathcal{N}_q = \mathcal{N}_q, \forall q \geq 0 \tag{8}$$

*Then we call this class of neuron the **minimal neuron class**, and any $x \in \mathcal{N}_i$ of such is called the **minimal neuron** or **standard neuron**, denoted by x_S. By default, this is satisfied if $q = 0$ in our construction.* [a]

[a] The constant i here refers to the organization numbering of nested classes built upon by another components. In such, we observe that this construction implicitly defines itself to be the simple zeroth class.

Then, we defined the class of all minimal perceptron $\mathcal{N}_0$, or neuron unit, as followed.

Definition 3.2 (Class $\mathcal{N}_0$ on $\mathbb{R}$)**.** *A neuron unit $x \in \mathcal{N}$ belongs to class $\mathcal{N}_0(\mathbb{R})$ and is called a **standard neuron on** $\mathbb{R}$ if it satisfies the minimization set criteria, and can be written of the form:*

$$x = q = \sigma_{\mathcal{M}}(w \cdot p + b), \quad p \in \mathcal{I} \subset \mathbb{R}, w, b \subset \mathbb{R} \subset \mathcal{M}, \sigma : \mathbb{R} \to \mathbb{R} \in \mathcal{M}, q \in \mathcal{O} \tag{9}$$

*If σ is linear unit, that is, $\sigma(wp + b) = wp + b$, then we say x is a **linear standard unit**.* [a]

[a] One might ask why we use the product and addition in the formula of wp and then b. In fact, this is perhaps more trivial - as to facilitate the concept of *linearity* - the formulation looks exactly like the linear line in a plane. Furthermore, as we will soon see, it is also of interest such that units of neurons can be linearly combined in a way, at least of computational aspect in running it on computers.

Naturally, a singular neuron is not enough, and as illustrated in Minsky & Papert (1988), they alone cannot do everything, for example, the XOR problem illustrated that particularly, there are unsolvable problems one can get with a simple perceptron. The resolution to this problem come in form of,

as implied, of the streamlined nature of neural structures — what if we operate them in parallel, in larger structures called **layers**, and so on? This is first illustrated by Rosenblatt, and its elementary form varies a lot in the history of classical connectionism. By the form of neuron class, we classify it $\mathcal{N}_2$. We reserve class $\mathcal{N}_1$ for the class of all *multiple-input neuron*, of which the cardinality is $(i, 1, 1)$ for $i = 1, \dots, n$. The motivation for $\mathcal{N}_1$ is that to resolve the problem of the class $\mathcal{N}_0$, one potential fix would be to 'fix bayonet' and free up i, thus giving the construction of $(i, j, 1)$. We call this **multivariate neuron**. If it is $(i, 1, 1)$, then we call it the **multivariate standard neuron**. All of such neurons then belong to the **class** $\mathcal{N}_1$ **neuron** simplex. Then, the class $\mathcal{N}_2$ of layer neural networks, is defined as followed.

Definition 3.3 (Class $\mathcal{N}_2$ structure)**.** *We fix the signature of any given structure $\mathbf{N} \in \mathcal{N}_2$. Let us define, for $L_i \in M$ the structures of layers, of which L_i contains $n_{i,j} \in \{\mathcal{N}_0, \mathcal{N}_1\}$ of subsequent lower class, and fix their cardinality of the form $(i, 1, k)$. Then, a **neural network** $\mathbf{N}$ of the class $\mathcal{N}_2$ is equivalent to the following structure:*

$$\mathbf{N} \in \mathcal{N}_2 \equiv I_{\mathbf{N}} \times M_{\mathbf{N}}(L_1 \times L_2 \times L_3 \times \cdots \times L_j) \to O_{\mathbf{N}} \tag{10}$$

The cardinality of $\mathbf{N}$ is then (i, j, k), for all $i, j, k \in \{1, \dots, n\}$. For $\mathbf{N}(i, 1, k)$, we call it as the shallow neural network.

The structure of our theory on the neural formalism is influenced by the object-abstracted treatment of mathematically embedded structures, and the unit-wise principle of particular neuron. Before we meet ourselves into the notion of epistemic circularity[7] problem, we might as well clarify a few prerequisites for such structure to exhibit.

First, we indict on the fundamental encoding environment that any object can take. The main point of any structure here is that there exists fundamentally the encoding space of two types. First is the object's cardinality space, denoted $\Gamma = (\mathbb{N}, F)$ for any given categorization F. Second is the encoding primitive of the field $\mathbb{R}$ for generality - in general any field is alright, and they define the analytic structure of the system itself. Any extension, for example, the $\mathbb{R}$-algebra of complex number $\mathbb{C}$ is then the primitive field's extension. [8]. Any type of data or system can then be decomposed to such, with additional structure on top of such primitive. Such is then called the *primitive framework*.

Definition 3.4 (Primitive framework)**.** *Let us define the primitive framework $\mathcal{P}_0$ of the dual $(\Gamma, \mathbb{R})$ where $\Gamma = (\mathbb{N}, F)$ is the cardinality encoding space, and $\mathbb{R}$ is the base field primitive of the analytical encoding. An object $X \in \mathcal{P}_0$ admits a dual representation,*

$$X \mapsto (\gamma(X), \rho(X)) \tag{11}$$

for $\gamma : \mathrm{Obj} \to \Gamma$ of cardinality encoding, and $\rho : \mathrm{Obj} \to \mathcal{E}(\mathbb{R})$ for the field extension of $\mathbb{R}$ category, or the category of all $\mathbb{R}$-algebras. For $\mathcal{E}(\mathbb{R})$ without extension, then $\mathcal{E}(\mathbb{R}) \cong \{\mathbb{R}\}$ of all $\mathbb{R}$-algebra.

For a structure that is supposed to be unit-wise constructed like neural network and the like, one of the main principle is the principle of abstraction. With this, come the idea of layer. Specifically, a *layer* separates abstraction in terms of subspace. Let us take an example of such kind for clarification. Let us define the primitive framework $\mathcal{P}_0$ as now the layer L_0 of this layering scheme. Then, we define the layer $\mathcal{L}_1$ of all unit-wised neuron-like units taking over the representation scheme on $\mathcal{L}_0$. We then define the structure of the neuron class $\mathcal{N}_1$ upon such as followed.

Definition 3.5 (Base neuron class)**.** *We define the base neuron class $\mathcal{N}_1 \equiv L_1$ as followed. For any $\mathcal{U} \in L_1$ for $\mathcal{U}$ as unit-wise construction over L_0, then every $\mathcal{U}$ satisfies the input signature*

[7] Also called bootstrapping, where to understand or define certain notion, requires the knowledge of the object wishes to be defined itself - for example, defining the size of a finite natural number set, using the elements of the number set itself.

[8] This corresponds to a variety of ideas. For example, see Cartuyvels et al. (2021) for the idea of discrete and continuous representations and processing, and Müller et al. (2022) for the same idea, but used on computer graphic process where it is constructed as discrete indexing structure (hash tables) with continuous neural fields. Cardinality space takes inspiration from combinatorial species, sheaf theory of global-local set, and Grothendieck's approach to algebraic topology. Extension is the idea from field extension itself, for structures that can be extended. The general idea of this section relies on interpretation of category theory.

*$\mathcal{I}_{sig} : \mathcal{E}'(\mathbb{R}) \to L_0$, output signature $\mathcal{O}_{sig} : L_0 \to \mathcal{E}'(\mathbb{R})$, and $\mathcal{C}_{ext}$ as extensible construction over L_0, for $\mathcal{E}'(\mathbb{R})$ particular extension on analytical encoding of L_0. The **type** template of a neural unit $\mathcal{U} \in L_0 \equiv \mathcal{N}_1$ is then defined as $\mathcal{U} = \langle \mathcal{I}_{sig}, \mathcal{C}_{ext}, \mathcal{O}_{sig} \rangle$ where $\mathcal{I}_{sig}$ and $\mathcal{O}_{sig}$ are invariant as type.*

Those definitions directly link it to type theory, while such development is perhaps more complex. In general, this specifies the *type* of a particular unit using the invariant analytical typing in the general environment linking to it. This is the *outer typing* of a given model construct , of which defers the type of which interaction between the environment, or the *global space state* happens and what is received of such. While the ambiance space can be forgiving, such cannot be said for the typing of given structure, since it must have the correct typing for identification. The extension is then is fairly simple, but the more important notion is the equivalence of the variant C_{ext} extensible unit of the inner structure itself.

While this is a prototype, it certainly focuses on the main axis of development that would be detrimental to the topic of *interpretability*, capacity, analysis and so on. We can indeed, extend this particular base template of a neuron class to much greater strength, of which address partially concerns in both implementation, and analysis. Type and neuron specification classifies neuron into different *class* of components instead of the mundane functional structure that we currently have, but also individualized and structuralized the *connection, information* and signature aspect of a processing unit. This allows for more grounded expression and composition rule, for example, the complexity metric of individual components of a larger construction of neural network, of which then can be expressed by individual neurons, and such neuron's individual's components. Dynamics works the same, however, it will shift - instead of learning as a back-tracking process, we can express backpropagation as a kind of *internal mechanism* on addition to the working framework, and the components such measures.

Furthermore, the primitive framework itself serves of a detrimental role in expressing system-wide, modelling-style and global specification of the working space. While we speak of hypothesis class $\mathcal{H}$ or $\mathcal{C}$ for example, such class can only infer, usually of specific function class, definitions, requirements, and so on. With this, we can express them both in a set-theoretical way of cardinality and field-extension similar to embedded vector space, and also encoding such as the numerical encoding typically seen in computing system, and structural addition and expression explicitly considered so. Indeed, there are problems and pitfalls, as well as the immaturity of the functional structure itself, but per a prototype, the general conceptual idea of this particular framework is fairly sound, and would be able to provide deeper insight and more sophisticated interpretation, rather than just simple neural network functions before.

3.2 THE LEARNING THEORETIC

REFERENCES

Mathematical Modeling and Simulation: Introduction for Scientists and Engineers, 2nd Edition | Wiley, 2024. URL https://www.wiley.com/en-us/Mathematical+Modeling+and+Simulation%3A+Introduction+for+Scientists+and+Engineers%2C+2nd+Edition-p-9783527839407.

Rahul Mishra Abdur Rahman Bin Md Faizullah, Ashok Urlana. Limgen: Probing the llms for generating suggestive limitations of research papers. *arXiv preprint arXiv:2403.15529*, March 2024. URL https://arxiv.org/abs/2403.15529. Shows summarization-specific models fail to generate limitations due to their training objectives.

Mikkel Abrahamsen, Linda Kleist, and Tillmann Miltzow. Training neural networks is $\exists\mathbb{R}$-complete. In *Proceedings of NeurIPS 2021*, 2021. URL https://arxiv.org/abs/2102.09798. arXiv:2102.09798.

Nicole Abreu, Parker B. Edwards, and Francis Motta. Topological machine learning with unreduced persistence diagrams, 2025. URL https://arxiv.org/abs/2507.07156.

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications, 2021. URL https://arxiv.org/abs/2006.05205.

Pierre Alquier. Pac-bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, 2008. doi: 10.3103/S1066530708040017. Bounds for randomized predictors; extended PAC-Bayesian theory.

Alexander Amini, Wilko Schwarting, Ava P. Soleimany, and Daniela Rus. Quantifying epistemic uncertainty in deep learning. *arXiv preprint arXiv:2110.12122*, 2021.

Dana Angluin. Queries and concept learning. *Machine Learning*, 2(4):319–342, 1988. doi: 10.1007/BF00116828. URL https://doi.org/10.1007/BF00116828.

Dana Angluin. Computational limitations on learning from examples. *Journal of the ACM*, 36(4):955–981, 1989. doi: 10.1145/76322.76335. URL https://dl.acm.org/doi/10.1145/76322.76335.

R. Anil et al. Palm 2 technical report. arXiv preprint arXiv:2305.10403, 2023. URL https://arxiv.org/abs/2305.10403.

Anthropic. The claude 3 model family: Opus, sonnet, haiku (model card). https://www-cdn.anthropic.com/.../Model_Card_Claude_3.pdf, 2024. Official Claude 3 model card (PDF) from Anthropic.

Anthropic Research. Tracing the thoughts of a large language model. https://www.anthropic.com/research/tracing-thoughts-language-model, 2025. Research blog/papers on interpretability and internal features.

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, E Bengio, I. Serban, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning (ICML)*, 2017. URL https://arxiv.org/abs/1706.05394. arXiv:1706.05394.

Joris Baan. The slodderwetenschap (sloppy science) of stochastic parrots – a plea for science to not take the route advocated by gebru and bender. *arXiv preprint arXiv:2101.10098*, January 2021. URL https://arxiv.org/abs/2101.10098. Counter-argument to the Stochastic Parrots paper, criticizing its methodology and ethics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Hyojin Bahng, Hyunwoo J. Kim, and Jaegul Yoo. Learning representations that support extrapolation. In *International Conference on Machine Learning (ICML)*, pp. 1433–1448. PMLR, 2022.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024. doi: 10.1073/pnas.2311878121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2311878121.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022. URL https://arxiv.org/abs/2212.08073. Anthropic's approach to AI alignment using AI-generated feedback.

Boaz Barak, Samuel B. Hopkins, Jonathan Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 428–437, 2016. doi: 10.1109/FOCS.2016.52.

Enrico Barbierato and Alice Gatti. The challenges of machine learning: A critical review. *Electronics*, 13 (2), 2024. ISSN 2079-9292. doi: 10.3390/electronics13020416. URL https://www.mdpi.com/2079-9292/13/2/416.

James Barrat. *Our Final Invention: Artificial Intelligence and the End of the Human Era*. Thomas Dunne Books, 2013. ISBN 978-0-312-62237-4.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. *Proc. Natl. Acad. Sci. U.S.A.*, 116(32):15849–15854, August 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1903070116. URL http://arxiv.org/abs/1812.11118. arXiv:1812.11118 [cs, stat].

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. ACM, 2021. doi: 10.1145/3442188.3445922. URL https://dl.acm.org/doi/10.1145/3442188.3445922. Foundational critique arguing LLMs statistically mimic text without real understanding, introducing the "stochastic parrot" metaphor.

Andrea Bianchini. On the notion of artificial in the age of synthetic biology and ai. *Foundations of Science*, 2021. doi: 10.1007/s10699-021-09799-w.

Derek Bickerton. *Language and Species*. University of Chicago Press, Chicago, 1990. URL https://archive.org/details/languagespecies0000bick.

Jonathan Birch. *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford University Press, 2024. ISBN 978-0-19-287042-1.

Leonard Bloomfield. *Language*. Henry Holt and Company, New York, 1933. URL https://books.google.com/books/about/Language.html?id=zduwAAAAIAAJ.

Avrim Blum, Adam Tauman Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM*, 50(4):506–519, 2003. doi: 10.1145/792538.792543.

Avrim L. Blum and Ronald L. Rivest. Training a 3-node neural network is np-complete. *Neural Networks*, 5(1): 117–127, 1992.

Raymond Board and Leonard Pitt. On the necessity of occam algorithms. *Theoretical Computer Science*, 100(1):157–184, 1992. doi: 10.1016/0304-3975(92)90367-O. URL https://doi.org/10.1016/0304-3975(92)90367-O.

Margaret A. Boden. *Artificial Intelligence and Natural Man*. MIT Press / Basic Books, Cambridge, MA / New York, 2, expanded edition, 1987.

Margaret A. Boden (ed.). *The Philosophy of Artificial Intelligence*. Oxford University Press, New York / Oxford, 1990.

Margaret A. Boden (ed.). *The Philosophy of Artificial Life*. Oxford Readings in Philosophy. Oxford University Press, New York / Oxford, 1996.

Margaret A. Boden and Ernest A. Edmonds. *From Fingers to Digits: An Artificial Aesthetic*. MIT Press, Cambridge, MA, 2019.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014. ISBN 978-0199678112.

Gordon Briggs. Machine ethics , the frame problem , and theory of mind. 2014. URL https://api.semanticscholar.org/CorpusID:14954096.

Selmer Bringsjord. A refutation of penrose's godelian case against artificial intelligence, 2000. URL http://cogprints.org/553/.

Rodney A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991. doi: 10.1016/0004-3702(91)90053-M. URL https://people.csail.mit.edu/brooks/papers/representation.pdf.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Neural Information Processing Systems, 2020. URL https://arxiv.org/abs/2005.14165. NeurIPS 2020. Introduced GPT-3 with 175B parameters demonstrating few-shot learning.

Cambridge Dictionary. Artificial, 2025. URL https://dictionary.cambridge.org/dictionary/english/artificial. Accessed: 2025-10-01.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2017. URL https://arxiv.org/abs/1608.04644.

Ruben Cartuyvels, Graham Spinks, and Marie-Francine Moens. Discrete and continuous representations and processing in deep learning: Looking forward. *AI Open*, 2:143–159, 2021. ISSN 2666-6510. doi: 10.1016/j.aiopen.2021.07.002. URL http://dx.doi.org/10.1016/j.aiopen.2021.07.002.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations, 2019. URL https://arxiv.org/abs/1806.07366.

Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.

Noam Chomsky. *Syntactic Structures*. Mouton & Co., The Hague, 1957. URL https://degruyter.com/document/doi/10.1515/9783112316009/html.

Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA, 1965. URL https://mitpress.mit.edu/9780262530071/aspects-of-the-theory-of-syntax/.

K. R. Chowdhary. *Fundamentals of Artificial Intelligence*. Springer, 2020. ISBN 978-81-322-3970-3.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022a. URL https://arxiv.org/abs/2204.02311. Google's 540B parameter model demonstrating breakthrough capabilities.

Amitabh Chowdhery et al. Palm: Scaling language modeling with pathways. 2022b. URL https://arxiv.org/abs/2204.02311.

Morten H. Christiansen and Nick Chater. Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5):489–558, 2008. doi: 10.1017/S0140525X08004998. URL https://greenfieldlab.psych.ucla.edu/wp-content/uploads/sites/168/2019/06/14-S0140525X08005141a.pdf.

Morten H. Christiansen and Simon Kirby (eds.). *Language Evolution*. Oxford Studies in the Evolution of Language. Oxford University Press, Oxford, 2003. ISBN 0199244839. URL https://www.lel.ed.ac.uk/~simon/0-19-924484-7.pdf.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Patricia S. Churchland. *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. MIT Press, Cambridge, MA, 1986. ISBN 978-0262031165.

William F. Clocksin and Christopher S. Mellish. *Programming in Prolog: Using the ISO Standard*. Springer, 5th edition, 2003. ISBN 978-3-540-00678-7.

Allan M. Collins and M. Ross Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2):240–247, 1969.

Alain Colmerauer and Philippe Roussel. The birth of prolog. In *History of programming languages—II*, pp. 331–367. ACM, 1993. Originally written November 1992.

Alain Colmerauer, Henry Kanoui, Robert Pasero, and Philippe Roussel. Un système de communication homme-machine en français. Rapport préliminaire de fin de contrat iria, Groupe Intelligence Artificielle, Faculté des Sciences de Luminy, Université Aix-Marseille II, France, October 1972.

Jack Copeland and Diane Proudfoot. From computer metaphor to computational modeling: The evolution of computationalism. *Minds and Machines*, 28(4):515–542, 2018. doi: 10.1007/s11023-018-9468-3.

Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965. doi: 10.1109/PGEC.1965.264137. URL https://doi.org/10.1109/PGEC.1965.264137.

Tianyu Cui, Yanling Wang, Chuanpu Fu, Yong Xiao, Sijia Li, Xinhao Deng, Yunpeng Liu, Qinglin Zhang, Ziyi Qiu, Peiyang Li, Zhixing Tan, Junwu Xiong, Xinyu Kong, Zujie Wen, Ke Xu, and Qi Li. Risk taxonomy, mitigation, and assessment benchmarks of large language model systems. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2401.05778.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989. doi: 10.1007/BF02551274.

Stéphane d' Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: where & why do they appear? In *Advances in Neural Information Processing Systems*, volume 33, pp. 3058–3069. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1fd09c5f59a8ff35d499c0ee25a1d47e-Abstract.html.

Charles Darwin. *The Descent of Man, and Selection in Relation to Sex*. John Murray, London, 1871. URL https://www.gutenberg.org/ebooks/2300.

Gilles E. Gignac David Ilić. Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement? *Intelligence*, 106:101858, 2024. doi: 10.1016/j.intell.2024.101858. URL https://doi.org/10.1016/j.intell.2024.101858.

Xander Davies, Lauro Langosco, and David Krueger. Unifying Grokking and Double Descent, March 2023. URL http://arxiv.org/abs/2303.06173. arXiv:2303.06173 [cs].

Randall Davis, Bruce Buchanan, and Edward Shortliffe. Production rules as a representation for a knowledge-based consultation program. *Artificial Intelligence*, 8(1):15–45, 1977.

Terrence W. Deacon. *The Symbolic Species: The Co-evolution of Language and the Brain*. W. W. Norton & Company, New York, 1997. URL https://uberty.org/wp-content/uploads/2016/02/Terrence_W._Deacon_The_Symbolic_Species.pdf.

DeepSeek AI / Hugging Face. deepseek-ai/deepseek-r1 (model card). https://huggingface.co/deepseek-ai/DeepSeek-R1, 2025. Model page and downloads (weights / details).

Howard B. Demuth, Mark H. Beale, Orlando De Jess, and Martin T. Hagan. *Neural Network Design*. Martin Hagan, Stillwater, OK, USA, 2nd edition, 2014. ISBN 0971732116.

Daniel C. Dennett. *Brainstorms: Philosophical Essays on Mind and Psychology*. MIT Press, Cambridge, MA, 1978. ISBN 978-0262540377.

Daniel C. Dennett. *Consciousness Explained*. Little, Brown and Company, Boston, 1991. ISBN 978-0316180665.

Hana Derouiche, Zaki Brahmi, and Haithem Mazeni. Agentic ai frameworks: Architectures, protocols, and design challenges. *arXiv preprint arXiv:2508.10146*, 2025. URL https://arxiv.org/abs/2508.10146.

Rene? Descartes. *Discourse on Method*. Harmondsworth, Penguin, Harmondsworth,, 1950.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. URL https://arxiv.org/abs/1810.04805. NAACL 2019. Introduced bidirectional pre-training for language representations.

Dictionary.com. Artificial, 2025. URL https://www.dictionary.com/browse/artificial. Accessed: 2025-10-01.

Gordana Dodig-Crnkovic. Info-computationalism and morphological computing of informational structures. *Information*, 3(2):204–218, 2012. doi: 10.3390/info3020204.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. In *arXiv preprint arXiv:1702.08608*, 2017.

Hubert L. Dreyfus. Alchemy and artificial intelligence. Technical Report P-3244, RAND Corporation, 1965. URL https://www.rand.org/pubs/papers/P3244.html.

Hubert L. Dreyfus. *What Computers Can't Do: A Critique of Artificial Reason*. Harper & Row, 1972. ISBN 0060110821.

Hubert L. Dreyfus. *Being-in-the-World: A Commentary on Heidegger's Being and Time, Division I*. MIT Press, Cambridge, MA, 1991.

Hubert L. Dreyfus and Stuart E. Dreyfus. *Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. Free Press, 1986. ISBN 0029080606.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Proceedings of the 29th Annual Conference on Learning Theory (COLT)*, volume 49 of *Proceedings of Machine Learning Research*, pp. 907–940, 2016. URL https://proceedings.mlr.press/v49/eldan16.html.

Jeffrey L Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

Chris Smith et al. The history of artificial intelligence. Technical review, 2006.

Nicholas Evans and Stephen C. Levinson. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–492, 2009. doi: 10.1017/S0140525X0999094X. URL https://cognitionandculture.net/wp-content/uploads/Evans-Levinson-BBS-2009-2.pdf.

Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, and Jiaxuan You. How far are we from agi: Are llms all we need? *arXiv preprint arXiv:2405.10313*, May 2024. URL https://arxiv.org/abs/2405.10313.

Charles J. Fillmore. The mechanisms of "construction grammar". In *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society*, pp. 35–55, 1988. doi: 10.3765/bls.v14i0.1794. URL https://journals.linguisticsociety.org/proceedings/index.php/BLS/article/view/1794.

Luciano Floridi. From the philosophy of information to an information ethics. *Philosophy & Technology*, 2004. Places AI debates within a broader 'information' philosophical framework.

Jerry A. Fodor. *The Language of Thought*. Harvard University Press, Cambridge, MA, 1975. ISBN 978-0674510302.

AI Alignment Forum. Beware general claims about "generalizable reasoning capabilities" (of modern ai systems). AI Alignment Forum, June 2025. URL https://www.alignmentforum.org/posts/5uw26uDdFbFQgKzih/beware-general-claims-about-generalizable-reasoning. Discusses historical arguments from Gary Marcus and statistical learning theorists about limitations of neural network architectures.

Dave Friedman. Understanding inference and the "stochastic parrot" in large language models. Personal blog (Substack), December 2024. URL https://davefriedman.substack.com/p/understanding-inference-and-the-stochastic. Argues LLMs are sophisticated pattern-matchers devoid of understanding, reasoning, or intentionality.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007. doi: 10.1214/07-AOAS131.

Lukas Galke, Yoav Ram, and Limor Raviv. Emergent communication for understanding human language evolution: What's missing?, 2022. URL https://arxiv.org/abs/2204.10590.

Howard Gardner. *Frames of Mind: The Theory of Multiple Intelligences*. Basic Books, New York, 1983.

Sumegha Garg, Pravesh K. Kothari, Pengda Liu, and Ran Raz. Memory-sample lower bounds for learning parity with noise. *arXiv preprint arXiv:2107.02320*, 2021. URL https://arxiv.org/abs/2107.02320.

Nicolas Gauvrit, Hector Zenil, and Per Tegnér. The information-theoretic and algorithmic approach to human, animal and artificial cognition. *arXiv preprint*, 2015.

Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural Computation*, 12(10):2451–2471, 2000.

Ben Goertzel. Artificial general intelligence: Concept, state of the art, and future prospects. In *Artificial General Intelligence*, pp. 1–48. Springer, 2014.

Ben Goertzel. Generative ai vs. agi: The cognitive strengths and weaknesses of modern llms. *arXiv preprint*, 2023. URL https://arxiv.org/abs/2309.10371.

Ben Goertzel and Cassio Pennachin (eds.). *Artificial General Intelligence*. Springer Science & Business Media, 2006. ISBN 9783540686774.

E. Mark Gold. Language identification in the limit. *Information and Control*, 10(5):447–474, 1967. doi: 10.1016/S0019-9958(67)91165-5. URL https://doi.org/10.1016/S0019-9958(67)91165-5.

Adele E. Goldberg. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago, 1995. URL https://press.uchicago.edu/ucp/books/book/chicago/C/bo3683810.html.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint*, 2014. URL https://arxiv.org/abs/1412.6572.

Alex Graves. Supervised sequence labelling with recurrent neural networks. *Studies in Computational Intelligence*, 385, 2012.

Jacob Grimm. *Deutsche Grammatik*. Bei Dieterich, Göttingen, 1819. URL https://archive.org/details/deutschegrammati01grim. Vol. 1 (other vols. published 1822–1837).

Jarek Gryz. The frame problem in artificial intelligence and philosophy. *Filozofia Nauki*, 21:15–30, 06 2013.

D. Guo et al. Deepseek-r1: Incentivizing reasoning capability in llms. arXiv preprint arXiv:2501.12948, 2025. URL https://arxiv.org/abs/2501.12948. Introduces DeepSeek-R1 and DeepSeek-R1-Zero; RL-based reasoning training.

Bruce Hajek and Maxim Raginsky. *Statistical Learning Theory*, volume 1. 2021. URL https://maxim.ece.illinois.edu/teaching/SLT/.

William L. Hamilton. Graph representation learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 14(3):1–159.

Karen Hao. *Empire of AI: Dreams and Nightmares in Sam Altman's OpenAI*. Penguin Press, 2025. ISBN 978-0593657508.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990a. doi: 10.1016/0167-2789(90)90087-6. How symbols in AI can acquire meaning beyond arbitrary manipulation.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42:335–346, 1990b. doi: 10.1016/0167-2789(90)90087-6.

John Haugeland. *Artificial Intelligence: The Very Idea*. MIT Press, Cambridge, MA, 1985. ISBN 978-0262580953.

Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579, 2002. doi: 10.1126/science.298.5598.1569. URL https://web.stanford.edu/class/linguist197a/hauser.pdf.

David Haussler, Michael J. Kearns, H. Sebastian Seung, Naftali Tishby, et al. Rigorous learning curve bounds from statistical mechanics. *Machine Learning*, 25(2–3):195–236, 1996. doi: 10.1007/BF00116991. Connects VC dimension, statistical mechanics, learning curves.

Martin Heidegger. *Being and Time*. Harper & Row, New York, 1962. Originally published in German as *Sein und Zeit* (1927).

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Charles F. Hockett. The origin of speech. *Scientific American*, 203(3):88–111, 1960. doi: 10.1038/scientificamerican0960-88. URL https://openlab.bmcc.cuny.edu/lin100b05w/wp-content/uploads/sites/3689/2024/03/The-Origin-of-Speech.pdf.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. URL https://arxiv.org/abs/2203.15556. Introduced Chinchilla and revised scaling laws for compute-optimal training.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991. doi: 10.1016/0893-6080(91)90009-T.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. doi: 10.1016/0893-6080(89)90020-8.

Weiche Hsieh, Ziqian Bi, Chuanqi Jiang, Junyu Liu, Benji Peng, Sen Zhang, Xuanhe Pan, Jiawei Xu, Jinlang Wang, Keyu Chen, Pohsun Feng, Yizhu Wen, Xinyuan Song, Tianyang Wang, Ming Liu, Junjie Yang, Ming Li, Bowen Jing, Jintao Ren, Junhao Song, Hong-Ming Tseng, Yichao Zhang, Lawrence K. Q. Yan def:SMLPand Qian Niu, Silin Chen, Yunze Wang, and Chia Xin Liang. A comprehensive guide to explainable ai: From classical models to llms, 2024. URL https://arxiv.org/abs/2412.00800.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021. doi: 10.1007/s10994-021-05946-3.

Wilhelm von Humboldt. *On Language: The Diversity of Human Language-Structure and its Influence on the Mental Development of Mankind*. Originally published 1836; modern English ed. Cambridge Univ. Press (1988), 1836. URL https://archive.org/details/onlanguagedivers0000humb. Originally published in German as "*Über die Verschiedenheit des menschlichen Sprachbaues*"; English translations published later (e.g., Cambridge Univ. Press ed.).

IBM. Artificial intelligence, 2024. URL https://www.ibm.com/think/topics/artificial-intelligence. Accessed: 2025-10-01.

Maor Ivgi, Yair Carmon, and Jonathan Berant. Scaling laws under the microscope: Predicting transformer performance from small scale experiments. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 7354–7371, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.544. URL https://aclanthology.org/2022.findings-emnlp.544/.

Philip C. Jackson. *Introduction to Artificial Intelligence*. Dover Publications, 3 edition, 2019. ISBN 0486843076.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey, 2018. URL https://arxiv.org/abs/1711.04305.

Hong Jun Jeon and Benjamin Van Roy. Information-theoretic foundations for machine learning, 2025. URL https://arxiv.org/abs/2407.12288.

A. Q. Jiang et al. Mistral 7b. arXiv preprint arXiv:2310.06825, 2023. URL https://arxiv.org/abs/2310.06825. Mistral AI release (7B model) and technical description.

Michael I Jordan. Serial order: A parallel distributed processing approach. *Advances in Psychology*, 121: 471–495, 1997.

Theresa Jose, Nicolò Cesa-Bianchi, and Yevgeny Seldin. Information-theoretic analysis of epistemic uncertainty in bayesian meta-learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 11025–11047. PMLR, 2022.

Justia Legal Dictionary. Artificial, 2025. URL https://dictionary.justia.com/artificial. Accessed: 2025-10-01.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020a. URL https://arxiv.org/abs/2001.08361. Established empirical scaling laws for language model performance.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020b. URL https://arxiv.org/abs/2001.08361.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in ml-based science. *arXiv preprint arXiv:2207.07048*, 2022.

Simon Kirby. Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110, 2001. doi: 10.1109/4235.918430. URL https://www.ling.ed.ac.uk/~simon/Papers/Kirby/Spontaneous%20evolution%20of%20linguistic%20structure%20an%20iterated.pdf.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, K Milan, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/content/114/13/3521.

Gitta Kutyniok. The mathematics of artificial intelligence. *arXiv preprint arXiv:2203.08890*, 2022. URL https://arxiv.org/abs/2203.08890.

Geoffrey LaForte. Why gödel's theorem cannot refute computationalism. *Minds and Machines*, 8(4):577–593, 1998. doi: 10.1023/A:1008320429318.

Siwoo Lee and Adji Bousso Dieng. Are neural scaling laws leading quantum chemistry astray?, 2025. URL https://arxiv.org/abs/2509.26397.

Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17:391–444, 2007a. doi: 10.1007/s11023-007-9079-x.

Shane Legg and Marcus Hutter. A collection of definitions of intelligence, 2007b. URL https://arxiv.org/abs/0706.3639.

Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993. doi: 10.1016/S0893-6080(05)80131-5.

Baoyu Liang, Yuchen Wang, and Chao Tong. Ai reasoning in deep learning era: From symbolic ai to neural–symbolic ai. *Mathematics*, 13(11), 2025. ISSN 2227-7390. doi: 10.3390/math13111707. URL https://www.mdpi.com/2227-7390/13/11/1707.

J. Lighthill. Artificial intelligence: A general survey. In J. Lighthill, N. S. Sutherland, R. M. Needham, H. C. Longuet-Higgins, and D. Michie (eds.), *Artificial Intelligence: A Paper Symposium*, pp. 1–21. Science Research Council of Great Britain, London, 1973. URL https://www.aiai.ed.ac.uk/events/lighthill1973/lighthill.pdf.

Robert K. Lindsay, Bruce G. Buchanan, Edward A. Feigenbaum, and Joshua Lederberg. Dendral: A case study of the first expert system for scientific hypothesis formation. *Artif. Intell.*, 61:209–261, 1993. URL https://api.semanticscholar.org/CorpusID:6929723.

Zachary C. Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship, 2018. URL https://arxiv.org/abs/1807.03341.

Zachary C Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

Zhi-Quan Luo and Paul Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992. doi: 10.1007/BF00939948.

Minh-Thang Luong, Huynh Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, 2015.

Gary Marcus. Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631, 2018. URL https://arxiv.org/abs/1801.00631.

Gary Marcus. Elegant and powerful new result that seriously undermines large language models. Marcus on AI (Substack), September 2023. URL https://garymarcus.substack.com/p/elegant-and-powerful-new-result-that. Critique of LLM capabilities with proposed disclaimer: "All facts presented by Generative AI—even those that are true—are fictitious".

Gary Marcus. A knockout blow for llms? *Communications of the ACM*, June 2025. URL https://cacm.acm.org/blogcacm/a-knockout-blow-for-llms/. Discusses devastating Apple research paper on LLM reasoning limitations, noting advocates are "partly conceding the blow".

Haggai Maron, Heli Ben-Hamu, Hadar Serviansky, and Yaron Lipman. Provably powerful graph networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bb04af0f7ecaee4aae62035497da1387-Paper.pdf.

David A. McAllester. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999. doi: 10.1023/A:1007618624809. Early PAC-Bayesian bounds that combine Bayesian ideas with PAC guarantees.

John McCarthy. *Generality in Artificial Intelligence*. Communications of the ACM, 1987. A key voice emphasizing the distinction between domain-limited AI ('weak') and the aspiration toward general AI ('strong').

John McCarthy and Patrick J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie (eds.), *Machine Intelligence 4*, pp. 463–502. Edinburgh University Press, 1969.

John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude E. Shannon. A proposal for the dartmouth summer research project on artificial intelligence. Technical proposal (Aug 31, 1955); reprinted/retrospective in AI Magazine 2006, 1955. Founding proposal that coined the term "artificial intelligence".

Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, December 1943. ISSN 1522-9602. doi: 10.1007/BF02478259. URL https://doi.org/10.1007/BF02478259.

Merriam-Webster. Artificial, 2025. URL https://www.merriam-webster.com/dictionary/artificial. Accessed: 2025-10-01.

Merriam-Webster Etymology. Artificial (etymology), 2025. URL https://www.merriam-webster.com/dictionary/artificial. Accessed: 2025-10-01.

Ryszard S. Michalski. On the quasi-inductive learning: the aq approach to rule extraction. In *Machine Intelligence and Pattern Recognition*, volume 4, pp. 171–197. 1969. Early work leading to AQ family of inductive rule-learning algorithms (symbolic rule induction).

Marvin L. Minsky. Steps toward artificial intelligence. *Proceedings of the IRE*, 49:8–30, 1961. doi: 10.1109/JRPROC.1961.287775. URL https://courses.csail.mit.edu/6.803/pdf/steps.pdf. Survey/position paper by Minsky summarizing early AI directions; foundational for symbolic AI.

Marvin L. Minsky (ed.). *Semantic Information Processing*. MIT Press, Cambridge, MA, 1968. URL https://philpapers.org/rec/MINSIP. Edited volume containing many classic symbolic AI papers (e.g., McCarthy, Minsky chapters).

Marvin L. Minsky and Seymour A. Papert. *Perceptrons: expanded edition*. MIT Press, Cambridge, MA, USA, 1988. ISBN 0262631113.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, October 2024. URL https://arxiv.org/abs/2410.05229. Demonstrates fundamental barriers to generalizable reasoning, with complete performance collapse on complex problems.

Mistral AI. Announcing mistral 7b. https://mistral.ai/news/announcing-mistral-7b, 2023. Official release blog and model card.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012. ISBN 026201825X.

Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A. Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. General pitfalls of model-agnostic interpretation methods for machine learning models. *arXiv preprint arXiv:2007.04131*, 2020.

Alhassan Mumuni and Fuseini Mumuni. Large language models for artificial general intelligence: A survey of foundational principles and approaches. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2501.03151.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):1–15, July 2022. ISSN 1557-7368. doi: 10.1145/3528223.3530127. URL http://dx.doi.org/10.1145/3528223.3530127.

Vincent C. Müller. Symbol grounding in computational systems: A paradox of intentions. *arXiv preprint*, 2025.

Thomas Nagel. *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False*. Oxford University Press, New York, 2012. ISBN 978-0199919758.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep Double Descent: Where Bigger Models and More Data Hurt, December 2019. URL http://arxiv.org/abs/1912.02292. arXiv:1912.02292 [cs, stat].

NASA. What is artificial intelligence?, 2023. URL https://www.nasa.gov/what-is-artificial-intelligence/. Accessed: 2025-10-01.

A. Newell and H. Simon. The logic theory machine–a complex information processing system. *IRE Transactions on Information Theory*, 2(3):61–79, 1956. doi: 10.1109/TIT.1956.1056797.

Allen Newell and Herbert A. Simon. Computer science as empirical inquiry: symbols and search. *Communications of the ACM*, 19(3):113–126, 1976.

Allen Newell, J.C. Shaw, and Herbert A. Simon. Elements of a theory of human problem solving. *Psychological Review*, 65(3):151–166, 1958. doi: 10.1037/h0048495. URL https://doi.org/10.1037/h0048495.

Allen Newell, J. C. Shaw, and Herbert A. Simon. Report on a general problem-solving program. In *Proceedings of the International Conference on Information Processing*, pp. 256–264, 1959. URL https://bitsavers.informatik.uni-stuttgart.de/pdf/rand/ipl/P-1584_Report_On_A_General_Problem-Solving_Program_Feb59.pdf. Description of GPS (General Problem Solver) — means-ends analysis.

Martin A. Nowak, Natalia L. Komarova, and Partha Niyogi. Evolution of universal grammar. *Science*, 291 (5501):114–118, 2001. doi: 10.1126/science.291.5501.114. URL https://pubmed.ncbi.nlm.nih.gov/11141560/.

Justin Nnaemeka Onyeukaziri. Artificial intelligence and the notions of the "natural" and the "artificial". In *Journal of Data Analysis*, volume 17, pp. 101–116. 2022.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1ldO2EFPr.

OpenAI. Introducing gpt-5. https://openai.com/gpt-5/, 2025a. OpenAI product/technical announcement (Aug 7, 2025).

OpenAI. Inside gpt-5 for work. https://cdn.openai.com/pdf/inside-gpt-5-for-work.pdf, 2025b. Technical overview / PDF from OpenAI.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022. URL https://arxiv.org/abs/2203.02155. Introduced RLHF for aligning language models with human preferences.

Oxford Learner's Dictionaries. Artificial, 2025. URL https://www.oxfordlearnersdictionaries.com/definition/english/artificial. Accessed: 2025-10-01.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009. ISBN 9780521895606. doi: 10.1017/CBO9780511803161.

Roger Penrose. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press, Oxford, 1989a. ISBN 978-0198519737.

Roger Penrose. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press, 1989b. Philosophical argument against the possibility of full computational emulation of human consciousness.

Roger Penrose. *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press, Oxford, 1994. ISBN 978-0198539780.

Roger Penrose and Emanuele Severino. *Artificial Intelligence Versus Natural Intelligence*. Springer, Dordrecht, Netherlands, 1997. URL https://link.springer.com/book/10.1007/978-3-030-85480-5.

Gabriel Peyré. The mathematics of artificial intelligence. *arXiv preprint arXiv:2501.10465*, 2025. URL https://arxiv.org/abs/2501.10465.

Steven Pinker and Paul Bloom. Natural language and natural selection. *Behavioral and Brain Sciences*, 13(4): 707–727, 1990. URL https://stevenpinker.com/files/pinker/files/pinker_bloom_1990.pdf. Target article with commentaries; preprint available.

A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

Hilary Putnam. *Representation and Reality*. MIT Press, Cambridge, MA, 1988. ISBN 978-0262660594.

Pierre Peigné Quentin FEUILLADE-MONTIXI. The stochastic parrot hypothesis is debatable for the last generation of llms. 2024. URL https://www.lesswrong.com/posts/HxRjHq3QG8vcYy4yy/the-stochastic-parrot-hypothesis-is-debatable-for-the-last. Provides examples where LLMs fail basic inference with novel information, supporting the stochastic parrot critique.

J. R. Quinlan. *Induction of Decision Trees*. Morgan Kaufmann, 1986. URL https://www.sciencedirect.com/science/article/pii/B9781558602485500092. ID3 / decision-tree induction — classic symbolic learning algorithm widely used in expert systems and symbolic concept learning.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI blog*, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. The original GPT paper introducing generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Introduced GPT-2, demonstrating zero-shot task transfer.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://arxiv.org/abs/1910.10683. Introduced T5, treating every NLP task as a text-to-text problem.

Shaina Raza, Ranjan Sapkota, Manoj Karkee, and Christos Emmanouilidis. Trism for agentic ai: A review of trust, risk, and security management in llm-based agentic multi-agent systems. *arXiv preprint arXiv:2506.04133*, 2025. URL https://arxiv.org/abs/2506.04133.

Paul M. Romer. Mathiness in the theory of economic growth. *American Economic Review*, 105(5):89–93, 2015.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408, 1958. URL https://api.semanticscholar.org/CorpusID:12781225.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, USA, 3rd edition, 2009. ISBN 0136042597.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition, 2010. Standard textbook that frames AI via agents and a taxonomy of definitions/goals.

Mahito Sugiyama Ryunosuke Ishizaki. Large language models: assessment for singularity. *AI & Society*, 2025. URL https://link.springer.com/article/10.1007/s00146-025-02271-4. Open access.

David Hyland-Wood Sandra Johnson. A primer on large language models and their limitations. *arXiv preprint arXiv:2412.04503*, December 2024. URL https://arxiv.org/abs/2412.04503. Argues LLMs lack self-monitoring (phenomenal consciousness) and internal updatable models of their environment.

Ranjan Sapkota, Konstantinos I. Rosenthal, and Manoj Karkee. Ai agents vs. agentic ai: A conceptual taxonomy, applications, and challenges. *arXiv preprint arXiv:2505.10468*, 2025. URL https://arxiv.org/abs/2505.10468.

Ferdinand de Saussure. *Course in General Linguistics*. Philosophical Library / Duckworth (English translations), 1916. URL https://ia600204.us.archive.org/0/items/SaussureFerdinandDeCourseInGeneralLinguistics1959/Saussure_Ferdinand_de_Course_in_General_Linguistics_1959.pdf. Compiled from lecture notes; classic edition/English translations include Wade Baskin (1959) and Roy Harris (1983).

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. doi: 10.1109/TNN.2008.2005605.

Matthias Scheutz (ed.). *Computationalism: New Directions*. MIT Press, Cambridge, MA, 2002. ISBN 978-0262692843.

August Schleicher. *A Compendium of the Comparative Grammar of the Indo-European, Sanskrit, Greek and Latin Languages*. Trübner & Co., London, 1874. URL https://archive.org/details/compendiumofcomp01schluoft.

J. Schneider. Generative to agentic ai: Survey, conceptualization, and future directions. *arXiv preprint arXiv:2504.18875*, 2025. URL https://arxiv.org/abs/2504.18875.

Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

William Seager. Frame problems, emotions and axiological projectionism. *Philosophical report*, 2010s (or older).

John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–457, 1980a. doi: 10.1017/S0140525X00005756.

John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–457, 1980b.

John R. Searle. *The Rediscovery of the Mind*. MIT Press, Cambridge, MA, 1992. ISBN 978-0262691549.

Murat Sensoy, Alex Kendall, and Flavio Esposito. Is epistemic uncertainty faithfully represented by evidential deep learning methods? *arXiv preprint arXiv:2402.09056*, 2024.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014. ISBN 1107057132.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11:2635–2670, 2010. URL http://jmlr.org/papers/v11/shalev-shwartz10a.html.

Murray Shanahan. The Frame Problem. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2016 edition, 2016.

Jingbo Shang, Zai Zheng, Jiale Wei, Xiang Ying, Felix Tao, and Mindverse Team. Ai-native memory: A pathway from llms towards agi. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2406.18312.

Edward Hance Shortliffe. *Computer-Based Medical Consultations: MYCIN*. Elsevier, New York, 1976. American Elsevier Publishing Company.

Hava T. Siegelmann and Eduardo D. Sontag. Turing computability with neural nets. *Applied Mathematics Letters*, 4(6):77–80, 1991. doi: 10.1016/0893-9659(91)90080-F. URL https://www.sontaglab.org/FTPDIR/aml-turing.pdf.

Herbert A. Simon. *The Sciences of the Artificial*. MIT Press, Cambridge, MA, 1 edition, 1969. ISBN 978-0262190510.

Ray J. Solomonoff. A formal theory of inductive inference. part i. *Information and Control*, 7(1):1–22, 1964a. doi: 10.1016/S0019-9958(64)90223-2. URL https://doi.org/10.1016/S0019-9958(64)90223-2.

Ray J. Solomonoff. A formal theory of inductive inference. part ii. *Information and Control*, 7(2):224–254, 1964b. doi: 10.1016/S0019-9958(64)90131-7. URL https://doi.org/10.1016/S0019-9958(64)90131-7.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. Language models fail to introspect about their knowledge of language, 2025. URL https://arxiv.org/abs/2503.07513.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning, 2023. URL https://arxiv.org/abs/2206.14486.

Stanford Encyclopedia of Philosophy. Artificial intelligence. https://plato.stanford.edu/entries/artificial-intelligence/, 2018. A well-rounded survey of history, proposed definitions, and philosophy-of-AI debates.

Hui Su, Zhi Tian, Xiaoyu Shen, and Xunliang Cai. Unraveling the mystery of scaling laws: Part i, 2024. URL https://arxiv.org/abs/2403.06563.

Lucy A. Suchman. *Plans and Situated Actions: The Problem of Human–Machine Communication*. Cambridge University Press, 1987. ISBN 0521388473.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27, pp. 3104–3112, 2014.

Richard S. Sutton. The bitter lesson. Web essay / blog post, 2019. URL https://www.incompleteideas.net/IncIdeas/BitterLesson.html.

Lars Pålsson Syll. Post-real economics — a severe case of mathiness. Blog post, Heterodox Economic Blogs, 2024.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR) Workshop*, 2014. URL https://arxiv.org/abs/1312.6199. preprint arXiv:1312.6199.

Matus Telgarsky. Benefits of depth in neural networks. In *Proceedings of COLT 2016*, 2016. URL https://arxiv.org/abs/1602.04485.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Édouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a. URL https://arxiv.org/abs/2302.13971.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023b. URL https://arxiv.org/abs/2302.13971. Meta's open foundation models ranging from 7B to 65B parameters.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023c. URL https://arxiv.org/abs/2307.09288. Updated and improved LLaMA models with better performance.

Thilo Traber, Andreas Zech, Nils Gessert, and Alexander Schlaefer. Relex: Regularisation for linear extrapolation in neural networks with rectified linear units. In *Artificial Neural Networks and Machine Learning – ICANN 2020*, pp. 170–182. Springer, 2020. doi: 10.1007/978-3-030-63799-6_13.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001. doi: 10.1007/s10957-001-0006-4.

Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1):387–423, 2009. doi: 10.1007/s10107-007-0170-0.

Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. Classic proposal of the "Imitation Game" (Turing Test).

Muhammad Saad Uddin. Stochastic parrots: A novel look at large language models and their limitations, April 2023. URL https://towardsai.net/p/machine-learning/stochastic-parrots-a-novel-look-at-large-language-models-and-their-limitations. Models are not capable of true reasoning or understanding, prone to errors and biases, and perpetuate stereotypes.

L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, November 1984a. ISSN 0001-0782. doi: 10.1145/1968.1972. URL https://doi.org/10.1145/1968.1972.

Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984b. doi: 10.1145/1968.1972. URL https://dl.acm.org/doi/10.1145/1968.1972.

V. N. Vapnik and A. Ya. Chervonenkis. The uniform convergence of the frequencies of events to their probabilities. *Doklady Akademii Nauk SSSR*, 181(4):781–783, 1968. URL https://mi.mathnet.ru/eng/tvrf/v181/i4/p781. Translated & extended in Theory of Probability & Its Applications 1971.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025. URL https://link.aip.org/link/?TPR/16/264/1.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer: New York, 1999.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008. Neural Information Processing Systems, 2017. URL https://arxiv.org/abs/1706.03762. NeurIPS 2017. The foundational paper introducing the Transformer architecture.

Tom Villani. How close is agi actually? why llms alone will not get us to agi, July 2024. URL https://www.njii.com/2024/07/why-llms-alone-will-not-get-us-to-agi/. Argues recent NLP advancements, while remarkable, do not constitute a path to AGI.

Shansong Wang, Mingzhe Hu, Qiang Li, Mojtaba Safari, and Xiaofeng Yang. Capabilities of gpt-5 on multimodal medical reasoning, 2025. URL https://arxiv.org/abs/2508.08224.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. URL https://arxiv.org/abs/2201.11903. NeurIPS 2022. Demonstrated how step-by-step reasoning improves LLM performance.

Jiaqi Wei et al. From ai for science to agentic science: A survey on autonomous scientific discovery. *arXiv preprint arXiv:2508.14111*, 2025. URL https://arxiv.org/abs/2508.14111.

Gail Weiss, Yoav Goldberg, and Eran Yahav. On the practical computational power of finite precision rnns for language recognition. In *Proceedings of ACL (short papers) / arXiv*, 2018. URL https://arxiv.org/abs/1805.04908. arXiv:1805.04908.

Terry Winograd. *Understanding Natural Language*. Academic Press, New York, 1972. URL https://archive.org/details/understandingnat0000wino. SHRDLU project — language understanding in a blocks world; core early symbolic NLP system.

David H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390, 1996. doi: 10.1162/neco.1996.8.7.1341. URL https://direct.mit.edu/neco/article/8/7/1341/6016/The-Lack-of-A-Priori-Distinctions-Between-Learning.

David H. Wolpert and William G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. doi: 10.1109/4235.585893. URL https://www.cs.ubc.ca/~hutter/earg/papers07/00585893.pdf.

Robert Worden. A unified theory of language, 2025. URL https://arxiv.org/abs/2508.20109.

Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015. doi: 10.1007/s10107-015-0892-3.

Keyulu Xu, Mozhi Zhang, Jingling Li, Simon S. Du, Kenji Kawaguchi, and Stefanie Jegelka. How neural networks extrapolate: From feedforward to graph neural networks. *arXiv preprint arXiv:2009.11848*, 2020.

Eliezer Yudkowsky and Nate Soares. *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. Hachette Book Group, 2025. ISBN 9780316595643.

Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. Dive into deep learning, 2023. URL https://arxiv.org/abs/2106.11342.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, 2017. URL https://arxiv.org/abs/1611.03530.

Yijun Zhang et al. Lllms: A data-driven survey of evolving research on limitations of large language models. *arXiv preprint arXiv:2505.19240*, May 2025. URL https://arxiv.org/abs/2505.19240. Semi-automated review of 14,648 papers on LLM limitations from 2022-2024, covering reasoning failures, hallucinations, and multilingual capabilities.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, March 2023. URL https://arxiv.org/abs/2303.18223. Comprehensive survey covering pre-training, adaptation, utilization, and evaluation of LLMs. Updated through March 2025.

Wayne Xin Zhao et al. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, February 2024. URL https://arxiv.org/abs/2402.06196. Reviews prominent LLM families (GPT, LLaMA, PaLM) and discusses techniques for building and augmenting LLMs.

Yilun Zhao Zhijian Xu, Manasi Patwardhan, Lovekesh Vig, and Arman Cohan. Can llms identify critical limitations within scientific research? a systematic evaluation on ai research papers. *arXiv preprint arXiv:2507.02694*, July 2025. URL https://arxiv.org/abs/2507.02694. Presents LimitGen benchmark and taxonomy of limitation types in AI research, evaluating LLMs' meta-cognitive capabilities.

Even Øygarden. What is intelligence? a proposed framework of four different concepts of intelligence. Master's thesis, University of Agder, 2019. URL https://uia.brage.unit.no/uia-xmlui/bitstream/handle/11250/2632728/%C3%98ygarden%2C%20Even.pdf?sequence=1.

A Appendix. A potential answer to CRA

Somewhat, in a way or another, a solution can be reached at least in terms of interpreting intelligence in a varied different way. Of such, there exists my blueprint notes of before, which outlined particularly interesting aspects and somewhat naive representation and ideas on a functional construct similar of those established above. While it is not perfect, and the line of logic leaves much to be desired, the idea is nevertheless worth of consideration.

A.1 The merits of the processing unit

In fundamental neuroscience, one of the main question, that was answered figuratively and conceptually, is the question about the **merit of existence** of a separated organ that eventually, will exhibit controlling behaviours and might up to the point of intelligent behaviours. That is, it raises that

> Protists and the simplest multicellular animals (sponges) display ingestive, defensive, reproductive, and other behaviors without any nervous system whatsoever, raising the question: what is the adaptive **value** of adding a nervous system to an organism?

This point is, as it might sound, dilligently enough is very strong. We study intelligence as the essential part to search and create generally, a learnable system, adaptable system, and rational system, but to *which stage*, *which layers* and which *form* would this be classified into? Are our systems generally similar to the Proties and multicellular animals, or have advanced to a greater abstraction level?

A.2 The layer dilemma - or Chinese Room Argument

We pertain to the **layer dilemma**, or as generally recalled, the **Chinese room argument** [Searle, 1980]. If we choose to simplify partially our system to an IO conduct, then this argument is the first to be addressed. However, this will be done in a very problematic way.

Assumption A.1. *Of the layer dilemma, for the set $\{L_i\}_{i\leq n}$ of the bottom-up categorization and measure up to n layers, then L_{n-1} do not "understand" what happens in L_n, but L_n is of itself, with sufficient realization and connection. This means the hierarchical information pathway is one-way, bottom up.* [9]

However, if is so, then there exists the fact that we can define a condition Q on such that, for a measurable complexity and 'model capability' scale $A[p]$, then if $A[p] \geq Q$, the model escaped the potential restriction, and become recursive in nature. In principle, this such that the *man in the box* recursively think outside what is conceived to be in his control. For every downstream, there exists an encoding to the lower dimension. Such downstream inevitably will result in an exposure to the data, in some form or another, that can be decoded. If, the model capability for the outermost layer,

[9] Note that, in here, the presumptuous assumption is that each individual layers, or, the *man in the box*, do not understand the higher scope he is received. Yet, he can understand what he has to do, hence the operation is still valid. However, understanding what was required and interpreted by the upper layer, simply makes it so that the man himself cannot interpret what was relayed, and hence do not possess any capable realization on such fact.

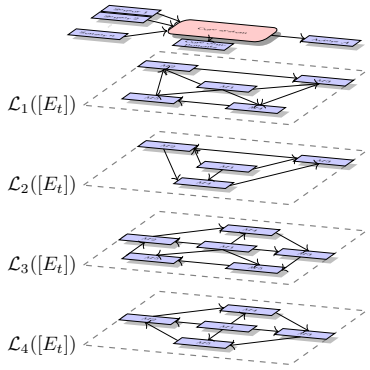


Figure 2: A loose illustration of the layering principle. The lower layers do not know what is on the upper part. However, they can receive potential downstream input features, and hence work accordingly. This is the basis of **abstraction by layers**, thus stating that *the man in the box do not know, and do not understand, anything but his circumstances and his current capability*. Nevertheless, it still fulfil its role in the lower level hence forth, and its operation only in such range of the layer.

L_n, satisfies the inequality, then it can, indeed, escape, and interfere directly to a certain extent of the actual 'world' layer L_0 it receives downstream to, which is, where we feed it the data and else, isolated from the model. This is, in principle again, totally a hypothesis, but then, we need to have a boundary to restrict this.

Conjecture A.1. *Suppose there exists a measure $\mathcal{M}[L_i \in A_n] \in \mathbb{R}^p$, usually for $p = 1$ as scalar for any model configuration A_n. Then, if $\mathcal{M}[L_i] \geq Q$ for $Q \to \dim(p)$, then there exists a recursive pattern $\mathrm{Re}(A)$ such that it enables the model's ability to 'learn' of the upper L_{i-1} layer.*

Hence, the answer in such form is reached, such so to address the fact that the operating machine receives downscale information within its own embedding, hence is typically restricted to that embedding layer alone. Furthermore, it also usually satisfies, since we categorize and generalize the layering, as according to the generalizability and abstraction value of each. Therefore, conspicuously, most of the time, the layer will hold, but under configuration, then it might jumps off the layer without any re-categorization method.

The layering theory also is an answer to the Chinese room argument, in the sense that there exists levels of which the abiding room is taking into account. Hence, for a person living in a scenario of input-output interaction, it is well-conceived that the 'I', within such capability, within such restriction, and being categorized as 'lower', do not possess the ability to look outside the room itself. A possible hypothesis, would be in analytical form, there exists a separated evaluating space, where such layering principle is demonstrated by dimensional increases, and their functional downstream compression function.

Hypothesis A.1. *Given the set $\{L_i\}_{i \leq n}$, there exists a dimensional analytic encoding Comp^m, such that there also exists certain functional downstream compression function F_{DC} that maps $\{L_i\} \times \mathsf{Comp}^m \to \{L_i\} \times \mathsf{Comp}^{m-1}$, essentially rescaling the dimensionality.*

Of course, this is all conceptual. We still need to define the measure $\mathcal{M}$, the supposed inequality, the supposed construction, and furthermore, the complete test implementation that can be perhaps focused on and verify the principle. Nevertheless, this might prove useful, or at least in a direction of analysis that will either falsify entirely this approach, or partially, thus makes use of this to formalize new framework on understanding such problem[10]. Nonetheless, such notion makes sense, for Searle-in-the-box to only be 'aware', of the arbitrary sense of the word 'aware' attributed of the analogous human awareness of what of his own domain, but not of any external or higher concept. It is apparent of human, too, as we can only semi-imagine the higher-level being, in philosophical inquiry, but nothing else — however when it comes to lower subjects, then it is surprisingly rich of

[10]Especially this, for the variable Q, which then by default of this, must either be formulated correctly in speciality, or rather, tested empirically.

certain interpretations. Such inquiry will remain valid, as for all of such layering scheme we can create. Such answer leads to two development. We can identify the layer of human, and try to build up layer-by-layer up there. Or, we might want to believe, there exists this layer, L_ϵ, of which such restriction of top-down interpretability becomes redundant, in which the subject leaps out of its layer to the above. What of the two is more probable? We do not know.

In one way or another, CRA infers with the option of an *input-output* procedure. In such case, CRA clearly tells us that a simple, mundane notion of an input-output machine which 'does its tasks' would not be sufficient of receiving the clarification and qualification for being intelligent. Which, by all means and purposes, are true. The construct if given in such circumstances, of the thought experiments, represents, if we took out the part where the argument said Searl is supposed to be everything a computer can be, then it's true that such constructs are false in its claim that it can 'understand'. In fact, relatively simple, a given input-output mechanism, from what was observed from the outside, do not exhibit anything, and do not have the ability to even *think*, regardless that is valid of such. If we are to stand by Descartes arguments, then it is even more of the truth - the system in which the thought experiment was conducted provides it with no capabilities of any such.

Then what would be of the Chinese Room Argument that is worth it to dissect? Well, firstly, the claims of, at least in the acute interpretation - the man in the box is supposed to be everything a computer can be is *false*. IF we are to stand by our construction of facilities, then we are innately arguing about such facilities, and not the processes, and the underlying operations itself. Rather, we are complaining that the machine is not capable enough, which is true. But we also, to a given reference point, pointing to the **existential facilities** instead of the arbitrary, yet reasonable operational facilities instead for the comparison. And in fact, if we think of it in the layer construction, it makes more sense - each layer is classified given its arbitrary for now, an interpretation thereof. Such interpretation is contained for such layer, and hence cannot be thoroughly or at least in a glance, interpreted by lower-layer components. This construct offers a one-way restriction on the property of interpretability. In such case, the **System Reply** is partially right - the man inside can not be, by all means and purposes, understand what does it mean by even 'English', or 'Chinese'. And even if such English 'understanding' is embedded in the reasonable interpretable space of Searl in the room, then Chinese would appear to be entirely unknown, *unless* there exists a helper tool to resolve the situation of undefined operation. Indirectly, this prevents a Descartes situation from happens - an undefined situation with particular way to resolve.

But more than that, what constitute the notion of understanding? Taken only from the setting of the thought experiment, we cannot do but deceptively assume that understanding a language is similar to giving it certain ruleset to transfer from this word to others word only. By that, converting from base-2 to base-10 works the same - you know how to do it, yet there are things that constitute the philosophical, higher-level notion of 'understanding' in such that allows you to actually understand the conversion - otherwise, the conversion is a blind matching. In fact, given the setting, would a conversion rulebook exists for such language? Language is, by itself, a very high-level concept. Translating from text to texts requires it not only to provide the definition matching, which could be reasonably identified by such notion like the rulebook mention, but the interpretation of the string is dictated by the logic of the language - the context in which it appears, the logical conformation that it contributes, the grammatical structure that makes sense of what is said and what is transferred, and else. Language itself, is a medium of information exchange, by one of its definition. A conversion does not constitute an understanding. And if the argument going back and forth is that Searl can somehow figure it out the patterns mean something, then it violates what I called the *principle of externality* - the 'lower components' up to a given point in which the *law of recursive immergence* does (not) apply, cannot implement its higher constructions. Then, the Chinese Room Argument can be interpreted, in somewhat meagre form, the argument against telling the current conformation as able to understand, while it is not.

B Appendix. Argument against the Gödelian argument

In 1961, J. R. Lucas presents the Gödelian argument against the existence of a "strong" AI. His proof is based on Gödel theorem, which is stated as followed:

> In any consistent system which is strong enough to produce simple arithmetic there are formulae which cannot be proved-in-the-system, but which we can see to be true. Essentially, we consider the formula which says, in effect, "This formula

> is unprovable-in-the-system". If this formula were provable-in-the-system, we should have a contradiction: for if it were provable-in-the-system, then it would not be unprovable-in-the-system, so that "This formula is unprovable-in-the-system" would be false: equally, if it were provable-in-the-system, then it would not be false, but would be true, since in any consistent system nothing false can be proved-in-the-system, but only truths. (Lucas, 1961)

This theorem holds for all formal systems which are consistent, adequate for simple arithmetic, and shows that those formal systems are incomplete, with some fact being true, but unprovable. "It is of the essence of being a machine, that it should be a concrete instantiation of a formal system. It follows that given any machine which is consistent and capable of doing simple arithmetic, there is a formula which is incapable of producing as being true…" (Lucas, 1961)

Further argued, he then comes to such conclusion that no machine can be a complete or adequate model of the mind, since "the mind are essentially different from machines". Lucas's defenders, Roger Penrose, also state in his *Shadow of the Mind* (1994). A human mathematician, if presented with a sound formal system F, could argue as followed:

> Though I don't know that I necessarily am F, I conclude that if I were, then the system F would have to be sound and, more to the point, F' would have to be sound, where F' is F supplemented by the further assertion "I am F"[11]. I perceive that it follows from the assumption that I am F that the Gödel argument $G(F')$ would have to be true and, furthermore, that it would not be a consequence of F'. But I have just perceived that "If I happen to be F, then $G(F')$ would have to be true", and perceptions of this nature would be precisely what F' is supposed to achieve. Since I am therefore capable of perceiving something beyond the powers of F', I deduce that I cannot be F after all. Moreover, this applies to any other system, in place of F. (Penrose, 1992, 3.2)

By default, the argument supplemented from Penrose raised the contradiction of proof-ness. The Gödelian argument implicitly creates layers, and levels, on which one puts those languages they are abided to seem fit of their expressions on the shelf, by the order of *effectiveness*. Such notion then, would make the advancement of machine to human seems perpetually, unsophisticatedly, inoperable and impossible in essence. Lucas argument, just as Searle, also claim that it is all the computer can do, of which the system itself is inherently useless of hosting such entity. However, artificial intelligence, as for now, using this term since chapter 3 which is not yet here, is not a computer in its form. We say, however, for an *artificial intelligent subject with computers as its existential facilities*, not the computer itself. This open up the fact that the notion of computer we are having right now, are also limited to the kind of classical computer, and not taking into account of any such similar 'computing architecture' of framework that might differ from such understanding. If so, then CRA is only partially right. But partially wrong since the comparison is limited to a form of internal structure in a well-formed system. That is to said - we need to create the (a) construct(s) that exceed(s) such argument. The problem is, how?

B.1 Remark

Before even taking a stance on such argument, what is the meaning and interpretations, as well as ostensively why it is even important to divulge into such point? The answers might be a bit difficult.

Human is variedly different from machine, for the current time with all the knowledge at present. Truth to take, the action of writing this itself is part of the endeavour to discover one's self, or rather, to understand F with the assertion of 'I am F', for now, that we can, and is doing. By the language and construction of contemporary and propositional logic, a machine cannot do that.[12]

[11] The phrase "I am F" is merely a shorthand for "F encapsulates all the humanly accessible methods of mathematical proof"

[12] In general, we cannot even say that it is true of the truth that human actually differs from what is proposed to be perceiving F' being F. For human understanding of ourselves alone, we are trying to fit it into the interpretation and the rough 'understanding' of human itself. That is, there exists the space of reason and argument of a scientist, which interpretation follows. If, supposedly, this interpretation is strong enough, then we might be able to perceive and understand ourselves from ourselves - a looped interpretability. This mechanism, if ever, is not well understood if exists.

However, if the converse situation happens, where we cannot totally perceive what we actually think, and how it is formed - per metric, being either consciousness, or one's self, surprisingly, it does not support the previous argument from an intuitive view (Bear in mind that this is a non-rigorous study). If stays rigid as it is, not counting being dynamic as we want, the model created from a human being can only imitate and represents what directly is entailed in the human mind of interpretation and logics. But logic and interpretation is a construct of the mind, for all intents and purposes, to directly infers to the physical world, the living world. However, if one is to use such inference on itself, for example, examining the brain itself, then to a certain point, what can be deduced from such observation can only fit in the interpretation space of what its creator, the human brain itself, can contrive. Thereby, we might conclude that figuratively, even human cannot understand human itself, from certain perspective. But the quality of succinctly interesting loop is to be taken seriously. The point now is, what type of construct, even logic, would be sufficient of taking the understanding, and will it make uses of the looped behaviours? By that, we then argue superficially that anything that relies on the machine cannot model the human existence and conscience itself. There are some assumptions thereof in the argument:

- Existences and the state of the world are in fact, modelled in mathematics, for one way or another. This is to facilitate the use of formal system in the argument. Everything is a set of rules, in which things operates.
- A machine per its definition, cybernetic machines are of all expressed by the single principle that it is born out of a formal system itself.
- Truth is the finite quantity that exists in such formal system, and is absolute.
- The mind is an entity of which is inherently different from the logic of formal system.

Those fundamental, overlapping assumptions make up the bulk of the Gödelian argument, from the surface. However, is that true of all the merit? [13]

[13] It turns out, however, the Gödelian argument has various proponents and opponents, and there are arguments of it being false. See Bringsjord (2000) for such argument, but it can be simplified as this. The Gödelian argument makes use of two assumptions: $G(F')$ is true for a Gödelian statement, and $F' \nvdash G(F')$ for $F' \nvdash G(F')$ for F' being "I am F" with added semantic. Then, the statement on $G(F')$ is true is nothing but a *satisfaction* claim, of meta-mathematical assertion which can be reduced to $\mathcal{I} \vDash G(F')$ is true for given interpretation $\mathcal{I}$. Thereby, there exists no contradiction thereof.