Contents

Emergence in Large Language Models: Theoretical, Philosophical, Mathematical, and Empirical Foundations

Consolidated Research Document

Project: Simulation of Emergence in LLMs **Directory:** `~/ais/simulacion-llm-emergencia/` **Original elaboration date:** 12 May 2026 **Revised and translated:** 12 May 2026 **Number of sources:** 51 verified bibliographical references

---

Abstract

[Fact] This document presents a comprehensive multidisciplinary review of the phenomenon of emergence in large language models (LLMs). Four fundamental dimensions of the problem are examined: (1) the theoretical framework of the simulation hypothesis and digital physics, which explores the ultimate computational nature of reality; (2) the philosophical foundations of mind and emergence, from functionalism to integrated information theory; (3) the mathematical foundations of complexity, including chaos theory, phase transitions, self-organized criticality, and the geometry of latent spaces; and (4) empirical evidence on emergent abilities in LLMs, including the Schaeffer et al. debate over whether emergence is a methodological mirage or a genuine ontological phenomenon. The document synthesizes contributions from Nick Bostrom, Max Tegmark, David Chalmers, Daniel Dennett, John Searle, Karl Friston, Giulio Tononi, Jason Wei, Ilya Sutskever, and numerous other researchers, establishing conceptual connections between digital physics, philosophy of mind, and computational cognitive science.

Keywords: emergence, large language models, simulation hypothesis, digital physics, philosophy of mind, phase transitions, criticality, mechanistic interpretability, computational consciousness.

Executive Summary

[Fact] This executive summary presents the key points and main conclusions of the document:

- **[Fact]** Emergence in LLMs is a widely documented phenomenon in which sophisticated cognitive capabilities appear abruptly when certain computational scale thresholds are crossed, observed in more than 40 different tasks according to Wei et al. (2022).
- **[Fact]** Two theoretical positions are in conflict: Schaeffer et al. (2023) argue that emergence is a methodological mirage generated by discontinuous metrics, while Wei et al. maintain that it represents a genuine ontological phenomenon.
- **[Hypothesis]** Digital physics, from Zuse (1969) to Tegmark (2014), suggests that the universe could be a giant computational system in which reality emerges from underlying informational processes, offering an analogous framework for understanding emergence in LLMs.
- **[Hypothesis]** The mathematical foundations of complexity —chaos theory, phase transitions, self-organized criticality —provide conceptual tools to explain how complex behaviors arise from simple interactions in high-dimensional systems.
- **[Fact]** Mechanistic interpretability has identified specific circuits, such as "induction heads", that implement pattern-completion algorithms, demonstrating that emergence has identifiable structural foundations.
- **[Speculation]** If emergence in LLMs reflects universal principles of complexity, it may connect with fundamental problems in philosophy of mind regarding how consciousness arises from physical processes.
- **[Metaphysics]** The hypothesis that the universe itself operates by simple computational rules generating apparent complexity raises deep questions about the nature of reality and our place in it.
- **[Fact]** Ilya Sutskever, former chief scientist at OpenAI, publicly warned that AI safety requires engaging with fundamental challenges that are not solved simply by scaling models.
- **[Hypothesis]** Emergence in LLMs may be a microcosm of the universal phenomenon by which complex patterns emerge from underlying computational structure, both in artificial systems and in the cosmos itself.

---

1. Introduction

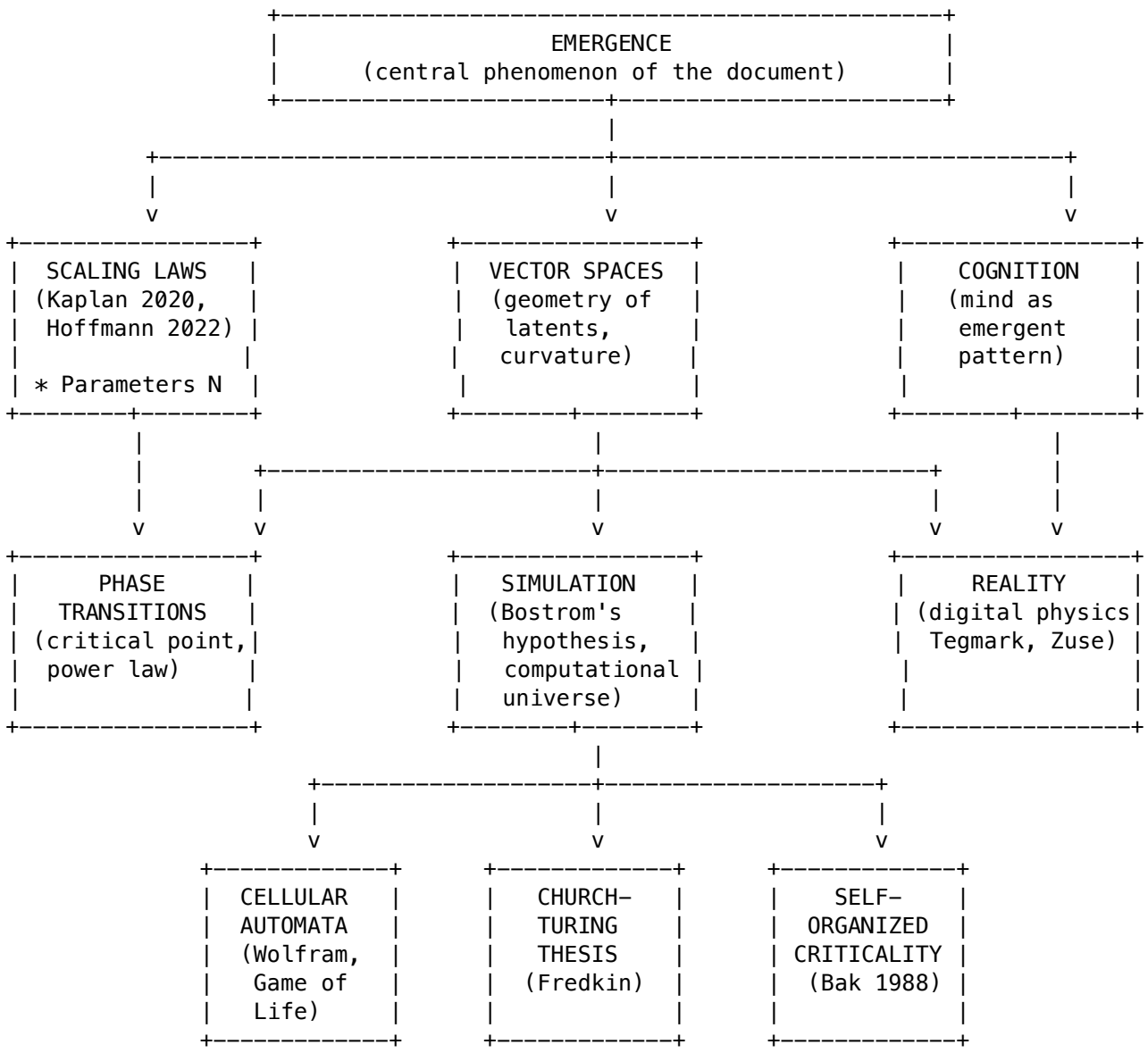
[Fact] Research on large language models (LLMs) has revealed a phenomenon that challenges conventional theoretical expectations: the abrupt appearance of sophisticated cognitive capabilities —such as multi-step reasoning, deep contextual understanding, and functional code generation —when models cross certain computational-scale thresholds. This phenomenon, termed **emergent abilities** by Wei et al. (2022), raises fundamental questions that transcend disciplinary boundaries. Is emergence in LLMs a reflection of universal principles that also operate in human consciousness and in the very structure of the cosmos? What relationship exists between symbolic processing in artificial neural networks and the emergence of mind in biological systems? Can the mathematics of phase transitions explain why qualitatively new capabilities arise unpredictably when scaling computational systems?

[Fact] Questions about emergence in complex systems are not exclusive to the field of artificial intelligence. Philosophy of mind has debated for decades how subjective conscious experience emerges from physical processes in the brain (Chalmers, 1996; Dennett, 1991; Searle, 1984). Theoretical physics has explored whether the universe itself could be a giant computational system that generates reality through simple rules (Zuse, 1969; Wheeler, 1989; Lloyd, 2006; Tegmark, 2014). The mathematics of dynamical systems and phase transitions provides conceptual tools to understand how complex behaviors arise from simple local interactions (Bak, Tang, & Wiesenfeld, 1988; Strogatz, 2018).

[Hypothesis] This document aims to consolidate these diverse intellectual traditions into a unified framework that illuminates the nature of emergence in LLMs. The central thesis is that emergence in large language models constitutes a genuine phenomenon —partially observable through appropriate metrics and partially a methodological artifact —that connects deeply with fundamental problems in philosophy of mind, digital physics, and complexity theory. Understanding this connection not only advances scientific knowledge of LLMs, but also offers fresh perspectives on questions that have challenged thinkers for centuries.

1.5. Concept Map

[Fact] The following concept map illustrates the hierarchical connections between the main concepts of the document:



[Fact] This concept map organizes the topics of the document in a five-level hierarchy:

Level 1 —Central Phenomenon: Emergence is the unifying concept of the entire document.

Level 2 —Enabling Factors: Three fundamental pillars support emergence: the **scaling laws** that govern the relationship between computational scale and performance; the **geometry of vector spaces** that characterizes how representations are organized in high dimensionality; and **cognition** as an analogous emergent process.

Level 3 —Mechanisms and Theories: The concepts of level 2 unfold into specific mechanisms: **phase transitions** that explain the abrupt appearance of capabilities; the **simulation hypothesis** as a framework for understanding the computational nature of reality; and **reality** itself as potentially emergent from mathematical structures.

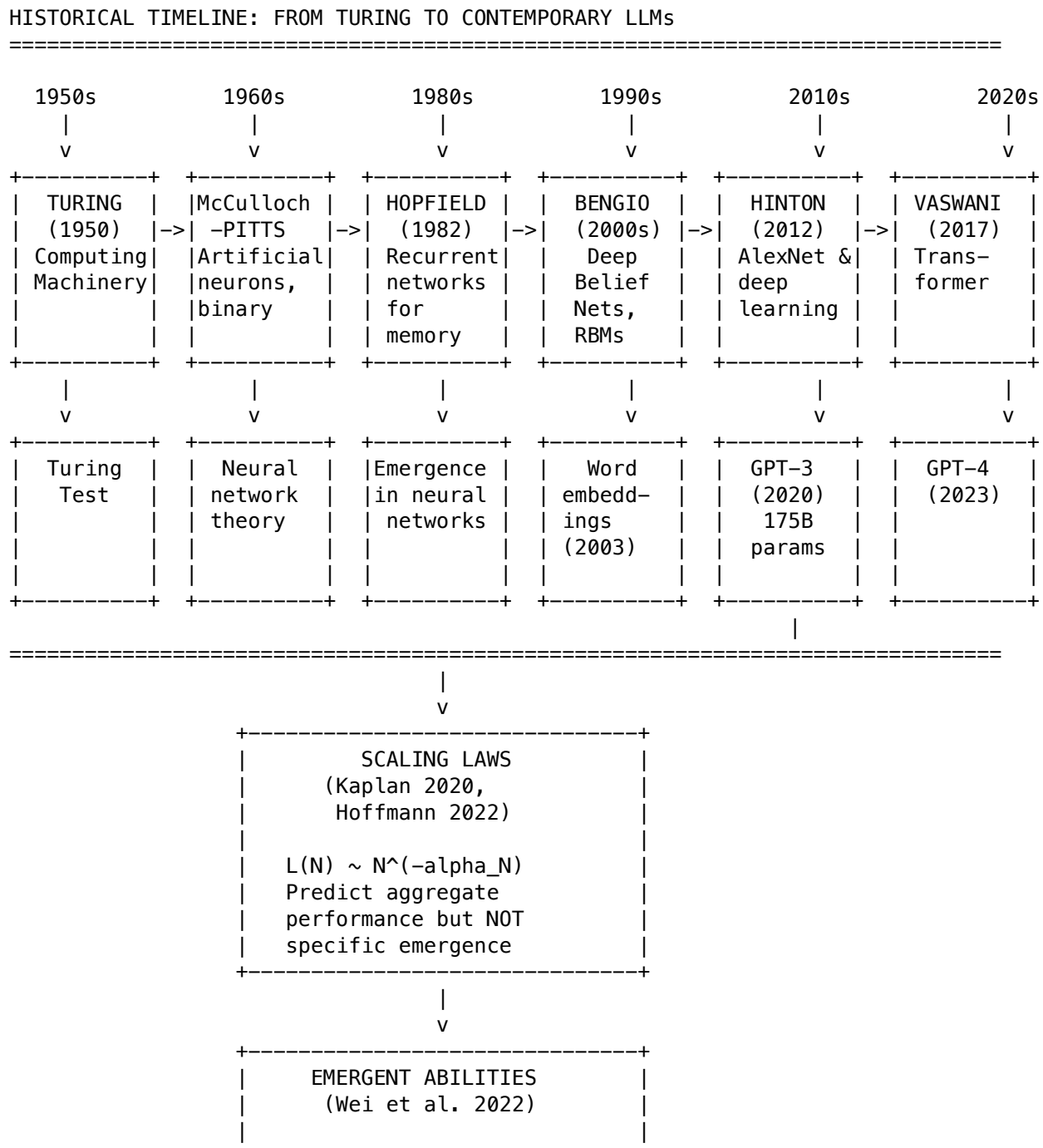
Level 4 —Mathematical Foundations: The mechanisms of level 3 are grounded in mathematical formalisms: **cellular automata** that demonstrate emergence from simple rules; the **Church–Turing thesis** that establishes the limits of the computable; and **self-organized criticality** that explains why certain systems naturally evolve toward critical states.

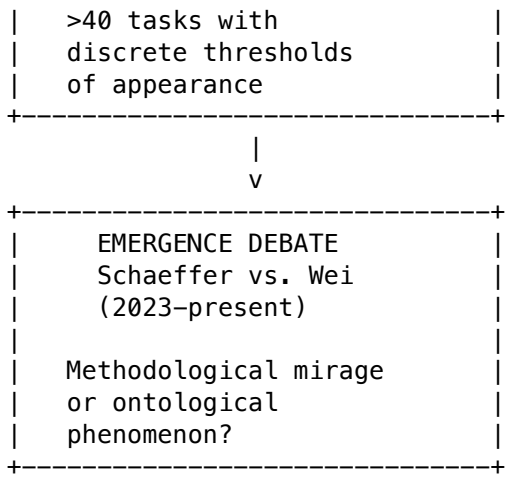
Level 5 —Primitive Concepts: The mathematical foundations rest on primitive concepts of complexity theory and computational logic.

[Hypothesis] The upward arrows in the diagram represent how higher-level properties emerge from lower-level interactions. The downward arrows represent how higher levels constrain and structure the behavior of lower levels.

1.6. Historical Timeline of Ideas

[Fact] The evolution of the concept of emergence in computational and cognitive systems follows a trajectory of more than seven decades, beginning with the theoretical foundations of computation and culminating in contemporary models of language at scale. This section traces that conceptual genealogy:





[Fact] **Alan Turing (1950):** In “Computing Machinery and Intelligence”, Turing proposed the celebrated test that bears his name as an operational criterion for intelligence, initiating the debate over when a machine can be considered “thinking”. His universal Turing machine established that any computation can be performed by a formal system of symbols manipulated by rules.

[Fact] **McCulloch–Pitts (1943):** Warren McCulloch and Walter Pitts proposed the first mathematical model of an artificial neuron, demonstrating that networks of such neurons can compute any logical function. This work founded the connectionist paradigm and established the correspondence between neural networks and logical computation.

[Fact] **John Hopfield (1982):** Hopfield introduced recurrent neural networks with associative energy, demonstrating that simple systems of binary units with feedback can exhibit emergent associative-memory behavior. His work connected statistical physics with neural networks.

[Fact] **Yoshua Bengio and Deep Learning (2000s):** Bengio pioneered deep belief networks and autoencoders, demonstrating that layered neural networks could learn hierarchical representations. His work on word embeddings established the foundations for vector representation of linguistic meaning.

[Fact] **Geoffrey Hinton and AlexNet (2012):** AlexNet’s victory at ImageNet 2012 demonstrated the power of deep learning with GPUs, inaugurating the modern era of deep learning. Hinton had worked for decades on neural networks before computational scale finally unlocked their potential.

[Fact] **Vaswani et al. and Transformers (2017):** “Attention is All You Need”introduced the transformer architecture that revolutionized natural language processing. The attention mechanism allowed capturing long-range dependencies without recurrence.

[Fact] **Kaplan et al. and Scaling Laws (2020):** OpenAI’s work demonstrated that language-model performance scales as a power law with parameters, data, and compute, providing the empirical framework for predicting aggregate trends.

[Fact] **Wei et al. and Emergent Abilities (2022):** Documented more than 40 tasks where capabilities appear abruptly upon crossing scale thresholds, coining the term “emergent abilities”and generating the debate that continues to this day.

[Fact] **Schaeffer, Miranda, and Koyejo (2023):** “Are Emergent Abilities a Mirage?”argued that apparent emergence is an artifact of discontinuous metrics, opposing the position of Wei et al.

---

2. Theoretical Framework: Simulation Hypothesis and Digital Physics

2.1. Nick Bostrom’s Simulation Trilemma

[Fact] Nick Bostrom, philosopher at the University of Oxford, formulated his celebrated simulation argument in 2003 in the paper “Are You Living in a Computer Simulation?”published in *The Philosophical Quarterly*. The argument is structured as a **trilemma** with three mutually exhaustive propositions (Bostrom, 2003):

[Fact] **First proposition:** It is unlikely that civilizations similar to humans will reach the computational capacity to simulate realities with a level of detail that includes conscious minds.

[Fact] **Second proposition:** If such civilizations exist, they probably would create many simulations, so the number of simulated minds would significantly exceed the number of non-simulated minds.

[Fact] **Third proposition:** We, with high probability, are one of those simulated minds.

[Hypothesis] Bostrom holds that at least one of these three propositions must be false. If we reject the third —that we do not live in a simulation —then we must accept either that civilizations never reach the capacity to create such simulations, or that they choose almost universally not to create simulations. The logical structure of the argument has generated vigorous debate in contemporary analytic philosophy.

[Metaphysics] **Note:** The Bostrom simulation hypothesis is a speculative philosophical position, not a scientific hypothesis in the Popperian sense (it is not directly falsifiable). It is presented as a philosophical exercise, not as verifiable science.

[Hypothesis] **Philosophical critiques of the argument:** The argument assumes that consciousness can be simulated, which is precisely what is in dispute. If functionalism is false —if mere functional organization does not suffice to produce genuine subjective experience —then simulating brain processes would not produce real conscious minds. The notion of a "conscious simulated mind"is questioned: can there be genuine experience in systems that are mere representations of biological systems? (Searle, 1984; Chalmers, 1996).

[Hypothesis] **Technical critiques:** Brueckner (2008), in "The Simulation Argument: Some Reflections", argued that Bostrom's argument relies on contestable indifference assumptions and a problematic application of self-locating probability. Birch (2013), in "On the 'Simulation Argument'and Self-Locating Belief", further argued that the analogy between simulating physical processes and simulating conscious experience is not valid. More recently, Beckers (2025) has formalized critiques showing structural similarities between Bostrom's argument and general theories of fiction.

[Speculation] **Evidence and counter-arguments:** The success of computational models in physics, biology, and other sciences suggests that the universe could have a computable nature. The analogy with increasingly sophisticated virtual worlds reinforces the plausibility of the hypothesis. However, there is no direct empirical evidence of glitches or anomalies that suggest an underlying discrete nature, and the argument could be circular by using future projections about technology that does not yet exist.

2.2. Digital Physics: Konrad Zuse, John Wheeler, and the Computing Universe

[Fact] **Konrad Zuse and Computing Space:** Konrad Zuse, German pioneer of computing, proposed in 1969 in his book *Rechnender Raum* (*Calculating Space*) that the universe itself could be a giant computational system. Zuse suggested that space-time has a discrete nature and that fundamental physical processes can be understood as computations occurring on an underlying grid. Zuse's central idea was that the universe computes its own behavior in a distributed manner: each small region of space performs local calculations that together produce the physical phenomena we observe (Zuse, 1969).

[Hypothesis] **John Wheeler and** "**It from Bit**"**:** John Archibald Wheeler, legendary Princeton physicist, developed the phrase "It from Bit"to describe his vision that every physical thing —"it"—emerges from binary information —"bit". Wheeler proposed that information is more fundamental than matter and that physical phenomena arise from quantum-information processes. His work with Edwin Jaynes and others led to the idea that physics can be understood as information processing, with questions like "what exists?"transforming into questions about what information exists and how it is processed (Wheeler, 1989).

[Hypothesis] **Edward Fredkin and the Finite Nature Hypothesis:** Edward Fredkin, MIT physicist, developed the "Finite Nature Hypothesis"which holds that every physical quantity is finite and discrete, that there is a discrete temporal and spatial background, and that all physical evolution is a computational process. Fredkin argued that nature must be digital at its foundation because analog systems can be simulated with digital ones but not the reverse, digital processes are more stable and reproducible, and a digital universe is philosophically more elegant (Fredkin, 1992).

[Hypothesis] **Seth Lloyd and the Universe as a Quantum Computer:** Seth Lloyd of MIT proposed that the universe is a giant quantum computer. In his book *Programming the Universe* (2006), Lloyd argued that the

universe has been processing information since the Big Bang, that each atom can be used as a qubit for quantum computation, and that the information entropy of the universe is proportional to its energy. Lloyd quantified the computational capacity of the observable universe at approximately 10^90 bits of information capacity and roughly 10^120 elementary logical operations performed since the Big Bang (Lloyd, 2002, 2006).

2.3. Stephen Wolfram: Computational Irreducibility and Cellular Automata

[Hypothesis] Stephen Wolfram, physicist and mathematician, creator of Mathematica and author of *A New Kind of Science* (2002), proposed the concept of **computational irreducibility** as a fundamental principle. The idea is that some computational systems are irreducible: there is no way to predict their behavior without simulating each step of the computation. This has profound implications because it means that for certain systems there are no mathematical shortcuts; the only way to know what the system will do is to let it run. This applies especially to systems that exhibit emergent complexity (Wolfram, 2002).

[Hypothesis] Wolfram also introduced the notion of **re-identification** in the context of cellular automata. He proposed that if the universe is computational at its foundation, then the objects we perceive could simply be patterns that persist and re-identify themselves through time in the system's dynamics. The question "what is an electron?"could be answered by saying that it is a persistent pattern in the computational dynamics of the universe, analogous to how a "glider"in Conway's Game of Life is a pattern that persists and moves.

[Fact] **Conway's Game of Life:** John Horton Conway created the "Game of Life"in 1970, a two-dimensional cellular automaton that demonstrated how extremely simple rules can produce extraordinary complexity. The four rules are: (1) a live cell with fewer than two live neighbors dies; (2) a live cell with two or three live neighbors survives; (3) a live cell with more than three live neighbors dies; and (4) a dead cell with exactly three live neighbors becomes alive. Game of Life has demonstrated the capacity to produce "gliders"that move, "guns"that produce continuous streams, "eaters"that destroy other patterns, and structures that can perform universal computation. It has been proven that Game of Life is Turing-complete (Conway, 1970; Gardner, 1970).

2.4. The Mathematical Universe Hypothesis (MUH) of Max Tegmark

[Hypothesis] Max Tegmark, MIT cosmologist, proposed the "**Mathematical Universe Hypothesis**" (MUH) in his 2007 paper "The Mathematical Universe"(*Foundations of Physics*, 2008) and elaborated it in his 2014 book *Our Mathematical Universe*. The MUH holds: "Our external physical reality is a mathematical structure."This is not merely that the universe can be described by mathematics, but that the universe literally is a mathematical structure. Physical objects are literally mathematical structures in a very literal sense (Tegmark, 2008, 2014).

[Hypothesis] Tegmark proposes a taxonomy of four levels of multiverse:

[Hypothesis] **Level I:** Regions beyond our cosmological horizon that are equally vast but with different initial conditions.

[Hypothesis] **Level II:** Bubble universes in an eternal inflation landscape with different physical constants.

[Hypothesis] **Level III:** The Many-Worlds interpretation of quantum mechanics —worlds that branch.

[Metaphysics] **Level IV:** The most radical level —all mathematical structures exist physically. Every possible computable universe exists as a real structure.

[Metaphysics] **Note:** The MUH is a metaphysical position / mathematical Platonism. It claims that the universe *is* literally a mathematical structure, not simply that it can be *described* by mathematics. This hypothesis is not directly falsifiable.

[Hypothesis] **Arguments in favor of the MUH:** The hypothesis explains the "mathematical inexorability"of physical laws, eliminates the problem of "wagering"on why there is something rather than nothing, and offers an elegant connection with fundamental physics. However, philosophical critiques point out: why do certain mathematical structures rather than others have physical existence?; the problem of the emergence of conscious observers from purely mathematical structures; and accusations of Platonic maximalism or "ultimate ensemble theory". Scientific critiques include the lack of direct testability and the absence of new predictions about known physics (Tegmark, 2014).

2.5. Connection with Emergence in LLMs

[Speculation] A connection emerges here between digital physics and language models. If:

[Fact] **Universe -> Mathematics -> Emergent reality** (Tegmark and digital physics)

[Fact] **Language -> Matrices -> Emergent cognition** (LLMs and neural emergence)

[Speculation] Then perhaps both processes are manifestations of the same fundamental phenomenon: emergence of qualitative complexity from underlying computational/mathematical structure. **Note:** This is an analogy, not an ontological identity. Cellular automata and LLMs are computational systems *within* the universe; the hypothesis that the universe "operates by simple rules"is a hypothesis about the nature of the universe as a whole, not a general law that applies equally to subsystems.

[Fact] LLMs process text through: (1) conversion of words into numerical tokens (*embeddings*); (2) transformation by weight matrices in transformer architectures; (3) processing of statistical patterns across enormous amounts of text; and (4) emergence of capabilities that were not explicitly programmed. Observed emergent phenomena include the ability to use semantic information without this necessarily implying phenomenological understanding, the ability for logical inference without explicit logic, and the generation of text that appears intelligent without necessarily being conscious (Vaswani et al., 2017).

[Speculation] If simple cellular automata produce emergent complexity *within* a computational system, and if the universe operates by simple rules, then perhaps LLMs illustrate an analogous principle: qualitatively new complexity emerging from simple underlying computational processes *within* the universe.

---

3. Philosophical Framework: Philosophy of Mind and Emergence

Bridge: Having considered the possibility that the universe itself is computational, we now turn to the philosophical question of how minds emerge from physical (potentially computational) substrates. The same emergence question scales down from cosmos to skull.

3.1. Functionalism and the Chinese Room Argument

[Fact] **John Searle and Biological Naturalism:** John Searle (1932–), philosopher at the University of California, Berkeley, is known principally for the **Chinese Room Argument**, presented in *Minds, Brains and Science* (1984). The thought experiment proposes the following: a Spanish speaker is enclosed in a room with a set of rules that allow them to correlate Chinese symbols of input with Chinese symbols of output. Although the speaker does not understand Chinese at all, they are able to produce appropriate responses by manipulating symbols according to the rules. Searle argues that this system can pass the Turing test without there being genuine understanding.

[Hypothesis] Searle uses this argument to refute **functionalism**, the thesis that mental states are defined exclusively by their functional relations. According to Searle, syntax (manipulation of symbols by rules) is insufficient to generate semantics (meaning, understanding). The Chinese room system processes information without understanding it.

[Hypothesis] Against functionalism, Searle proposes **biological naturalism**. According to this view, mental states are biological features of the brain, causal properties emerging from neural processes that, being biologically real, can be studied empirically. Consciousness is a causal property of the brain, not a mysterious add-on, but part of the natural order. Searle holds that mental phenomena are **causally reducible** to brain processes without being ontologically reducible —we cannot predict the qualitative properties of conscious experience from low-level neurobiological description, although such properties depend causally on it (Searle, 1984, 1992).

3.2. Daniel Dennett: Multiple Drafts Model

[Fact] **Daniel Dennett** (1942–2024), philosopher at Tufts University, distinguished himself by a vigorous naturalism and a commitment to scientific explanation of mental phenomena. In *Consciousness Explained* (1991), Dennett argued that common-sense intuition about consciousness —the image of a "Cartesian theater"where conscious contents are represented —is deeply misleading.

[Hypothesis] Dennett proposes instead a **Multiple Drafts** model. According to this model, there is no single privileged moment of conscious processing. Instead, the brain constantly generates multiple "drafts"of experience that compete with one another, with various fragments being edited and rewritten continuously. The unified experience of consciousness emerges from this heterogeneous process, not from a central point.

[Hypothesis] Dennett rejects the so-called "folk science"of consciousness —a set of intuitions and preliminary hypotheses that have infected the philosophical and scientific debate. He calls this the "phenomenal cult": the tendency to treat subjective intuitions about conscious experience as privileged data that any adequate theory must accommodate. For Dennett, consciousness is a natural phenomenon that will be explained through computational and neuroscientific models, without appealing to inexplicable or mysterious properties (Dennett, 1991).

3.3. David Chalmers: The Hard Problem of Consciousness

[Fact] **David Chalmers** (1966–), philosopher at the Australian National University and University of Arizona, articulated with clarity and rigor what he calls the "**hard problem**"**of consciousness**. The hard problem is distinguished from the easy problem: how the brain processes information, responds to stimuli, integrates information, controls behavior —empirically challenging questions but in principle approachable through scientific methods. The hard problem, by contrast, asks why and how conscious experience *feels* a particular way —why there is something it is like to be a conscious organism (Chalmers, 1996).

[Hypothesis] Chalmers develops an argument based on the nature of **qualia** —the subjective qualitative properties of experience. Consider the experience of seeing the color red: in addition to the functional processes that allow us to distinguish red from other colors, there is something that it *is like* to see red, an irreducible phenomenological character. Chalmers holds that no functional model, no matter how sophisticated, can inherently capture this character. We can imagine a functional system identical to ours —a "philosophical zombie"—that performs all the same operations without any experience. This suggests that qualia are ontologically irreducible to the functional.

[Metaphysics] **Panpsychism as a response:** Faced with the apparent irreducibility of consciousness, Chalmers has explored **panpsychist** positions —the thesis that consciousness is a fundamental property of certain physical systems, analogous to mass or charge. Chalmers'own position tends toward what he calls **Russellian monism**: consciousness is not identified with any known physical property, but neither is it completely independent — it has a structural relation with physical properties that reflects their pattern of organization. Despite his openness to metaphysical speculation, Chalmers considers himself a naturalist (Chalmers, 2010). **Note:** Chalmers' panpsychism is a metaphysical position, not an empirically verifiable scientific hypothesis.

3.4. Douglas Hofstadter: Strange Loops and Computational Emergence

[Hypothesis] **Douglas Hofstadter** (1945–), author of *Gödel, Escher, Bach: An Eternal Golden Braid* (1979), introduces the concept of the **tangled hierarchy** to describe feedback loops in which a pattern folds back on itself, generating higher levels of description.

[Hypothesis] In *I Am a Strange Loop* (2007), Hofstadter develops a theory of consciousness centered on **strange loops**. According to this view, the conscious mind emerges when a system develops the capacity to represent itself, creating a level of description that is causally related to the represented level. Consciousness is a pattern that arises from the activity of millions of "dumb"neurons that, organized into complex systems with feedback loops, generate a higher level of "software"—the mind —that is literally about itself.

[Hypothesis] Hofstadter explicitly rejects substance dualism and the trivial functionalism that reduces mind to abstract programs independent of substrate. His position is a form of **emergentist monism**: mind is what the brain does when it is organized in specific ways, without anything "more"being added. Hofstadter has been critical of AI approaches based on purely formal symbol processing, holding that genuine understanding —and potentially consciousness —requires self-modeling patterns (Hofstadter, 1979, 2007).

3.5. Roger Penrose: Quantum Physics and Consciousness

[Hypothesis] **Roger Penrose** (1931–), mathematical physicist at Oxford University, argues in *The Emperor's New Mind* (1989) that consciousness cannot be explained by any classical algorithmic process. His argument rests on several elements:

[Hypothesis] 1. **Gödel's incompleteness theorem:** Formal systems adequate for arithmetic contain truths that cannot be proved within the system. Penrose argues that humans can recognize these truths as valid, which implies a non-algorithmic cognitive capacity.

[Hypothesis] 2. **Limits of classical computation:** Classical physics, based on deterministic and computable laws, appears insufficient to explain certain aspects of cognition.

[Hypothesis] Penrose, in collaboration with anesthesiologist Stuart Hameroff, proposed the **Orch-OR** theory (*Orchestrated Objective Reduction*). According to this theory, consciousness arises from quantum processes in cellular microtubules —protein structures present in neurons. Objective Reduction (OR) is a proposed process in which the quantum wave function collapses due to quantum gravity, not external observation. The theory has been subject to intense scrutiny: quantum processes in the brain were thought to be too brief to be relevant to cognition; quantum decoherence would destroy any coherent quantum state in the warm, wet biological brain; and there is no direct empirical evidence (Penrose, 1989; Hameroff & Penrose, 2014).

3.6. Karl Friston: The Free Energy Principle

[Hypothesis] **Karl Friston** (1959–), theoretical neuroscientist at University College London, developed the **state-space theory** of the brain and the **Free Energy Principle** (FEP), a unifying theoretical framework that seeks to explain the self-organization of complex biological systems, including the mind.

[Hypothesis] The Free Energy Principle is a **variational principle**: any system that resists dispersion (entropy) must minimize *surprise*, where surprise is defined as the negative log-probability of sensory states given the system's internal models. Friston formulates this mathematically using **variational free energy**, defined as

$$F = D_{KL}(q(z|x)\|p(z)) - \ln p(x)$$

where the first term is the Kullback–Leibler divergence between the approximate posterior $q(z|x)$ and the prior $p(z)$, and the second term is the log-evidence. Minimizing F amounts to minimizing surprise (negative log-evidence). This minimization occurs through two mechanisms: **perception** (updating internal models to better predict sensory inputs) and **action** (selecting actions that improve prediction, reducing future surprise).

[Hypothesis] Friston proposes that consciousness can be understood as a process of **active inference**, in which the brain is essentially a prediction machine that constantly generates models of the world and compares them with sensory reality. The "self"emerges as a stable pattern in the brain's state space —an attractor that represents the body and its relation to the environment. Conscious experience would be the process of minimizing variational free energy through updating this self-centered model (Friston, 2010, 2019).

3.7. Giulio Tononi: Integrated Information Theory

[Hypothesis] **Giulio Tononi** (1960–), neuroscientist at the University of Wisconsin-Madison, created **Integrated Information Theory** (IIT). This theory offers a mathematically rigorous approach to quantifying consciousness.

[Hypothesis] IIT starts from an **axiom**: consciousness is identical to integrated information. This means that a system is conscious to the extent that it generates integrated information —information beyond what its parts could generate independently. Tononi formalizes this through the quantity **phi (Φ)**, which measures the integrated information of a system. The formal definition involves minimizing over partitions of the system: $\Phi = \min_P D(\rho\|\rho_P)$, where ρ is the density operator of the system, P is a partition, and D is a suitable information-loss measure under partitioning. Φ represents the amount of information generated by the system as a whole, above and beyond that generated by its parts separately.

[Hypothesis] IIT is derived from axioms about conscious experience (such as that experience is unified, structured, specific) and generates postulates about what a neuropsychologically adequate system must satisfy. The theory predicts that consciousness correlates with Φ, and that certain configurations can have low Φ despite processing information.

[Hypothesis] Φ can be computationally intractable for large systems, and the relation between Φ and conscious experience has been questioned —the theory appears to assign consciousness to systems that intuitively do not seem conscious. Some philosophers have argued that IIT conflates correlation with identity (Tononi, 2004; Tononi et al., 2016).

[Hypothesis] **Tension between FEP and IIT:** It is worth noting that FEP and IIT are not directly compatible. IIT predicts that a system is conscious if and only if $\Phi > 0$; FEP makes no direct predictions about conscious experience, but is rather a principle about biological self-organization. The two frameworks have different implications about which systems should be conscious; this incompatibility remains an open issue.

3.8. Terrence Deacon: Emergence and Homeostasis

[Hypothesis] **Terrence Deacon** (1953–), anthropologist and neuroscientist, author of *The Symbolic Species* (1997) and *Incomplete Nature* (2012), addresses the emergence of the symbolic —the ability to represent the world through signs —as a key phenomenon in the evolution of human mind.

[Hypothesis] Deacon distinguishes between two types of processes: **homeostatic processes** (which maintain equilibrium through negative feedback) and **morphodynamic processes** (which generate and maintain characteristic forms). Mental phenomena —intentionality, meaning, value —represent what Deacon calls "absent causes" : factors that are not physically present in the efficient causes but are indispensable for explaining certain processes.

[Hypothesis] For Deacon, mind emerges as a level of organization that constrains physical processes in characteristic ways. The symbolic is not simply a representation of the world, but a constraint that structures neural activity and behavior. Deacon shares with Searle the emphasis on the semantic nature of mind, but differs in not reducing mental phenomena to properties of the individual brain. For Deacon, intentionality has evolutionary roots and can be understood as an emergent feature of systems that include environment, body, and brain (Deacon, 1997, 2012).

3.9. The Symbolism vs. Connectionism Debate

[Fact] The **symbolic paradigm** (or symbolic processing, classical AI paradigm) holds that cognition can be explained through the manipulation of mental symbols structured by formal rules. Prominent advocates include Jerry Fodor, Zenon Pylyshyn, Herbert Simon, and Allen Newell. The main characteristics are: discrete and structured representations; precise combination rules; sequential or controlled parallel processing; and inspiration from formal logic and von Neumann computation (Fodor, 1975, 1983).

[Fact] **Connectionism** proposes that cognition emerges from networks of simple interconnected units —artificial neurons —that process information in a parallel and distributed manner. Prominent advocates include David Rumelhart, James McClelland, and Paul Smolensky. The main characteristics are: distributed and holographic representations; learning via adjustment of connection weights; massively distributed parallel processing; and inspiration from neurobiology (Rumelhart, McClelland, & PDP Research Group, 1986).

[Hypothesis] The **binding problem** is central: how are the multiple aspects of conscious experience —color, shape, motion, sound —integrated into a unified experience? Symbolic systems propose a central integrator; connectionists invoke patterns of temporal synchronization or message-passing architectures.

---

4. Mathematical Foundations of Complexity and Emergence

4.1. Chaos Theory

[Fact] Chaos theory studies non-linear dynamical systems that exhibit extreme sensitivity to initial conditions, which prevents long-term prediction despite their underlying determinism. A chaotic system is characterized by: **determinism** (the equations governing the system are completely deterministic); **sensitivity to initial conditions** (tiny differences in the initial state produce radically different trajectories); **non-linearity**; and **strange attractors** (trajectories converge toward fractal structures in phase space).

[Fact] The **Lyapunov exponent** λ measures the rate of separation of nearby trajectories. If $\lambda > 0$, the system exhibits chaos. The **Hausdorff dimension** d_H of a strange attractor relates fractal geometry with dynamics.

[Hypothesis] The emergence of complex behaviors in LLMs can be interpreted as transitions between chaotic and non-chaotic regimes in the parameter space of the model (Strogatz, 2018; Hilborn, 2000).

4.2. Dynamical Systems and Bifurcations

[Fact] A dynamical system is defined as a tuple (M, φ) where M is a differentiable manifold (phase space) and $\varphi : \mathbb{R} \times M \to M$ is a flow satisfying identity and group properties. **Bifurcations** are qualitative transitions in the behavior of the system when parameters cross critical values.

[Fact] The **Hopf bifurcation** occurs when a pair of complex eigenvalues cross the imaginary axis. For $\mu < 0$, the origin is a stable focus. For $\mu > 0$, a stable limit cycle emerges. The parameter space of an LLM can be conceptualized as a very high-dimensional dynamical system in which training represents a flow toward minima of the loss function (Guckenheimer & Holmes, 1983).

4.3. Criticality and Phase Transitions

[Fact] **Criticality** occurs at the transition point between ordered and disordered phases. Phase transitions are classified by their order: **first order** (discontinuity in the free energy and its first derivatives) and **second order** (continuity in the free energy but discontinuity in second derivatives, e.g., the paramagnet–ferromagnet transition).

[Fact] The **order parameter** m near the critical point T_c follows power laws: $m \sim |T - T_c|^{\beta}$, where β is a critical exponent.

[Hypothesis] **Self-Organized Criticality (SOC):** Bak, Tang, and Wiesenfeld (1988) proposed that certain systems naturally evolve toward a critical state without fine-tuning of parameters. This theory is relevant for understanding how LLMs might operate near criticality, where maximum sensitivity to inputs corresponds to maximum information capacity (Bak, Tang, & Wiesenfeld, 1988; Stanley, 1971).

4.4. Network Theory

[Fact] A network is defined as $G = (V, E)$ where V is the set of vertices and $E \subseteq V \times V$ the set of edges. Networks can be characterized by their **degree distribution** $P(k)$. Barabási and Albert (1999) proposed the *preferential attachment* mechanism that produces scale-free networks with distribution $P(k) \sim k^{-\gamma}$.

[Fact] The **clustering coefficient** measures the density of triangulations in the network. The Kuramoto model captures synchronization in weakly coupled oscillators, with a phase transition at K_c where the order parameter $r > 0$ (Barabási, 2016; Newman, 2010).

4.5. Geometry of Latent Spaces

[Fact] The **manifold hypothesis** indicates that high-dimensional data lie on low-dimensional manifolds embedded in $\mathbb{R}^D$. The **Riemannian distance** on curved manifolds is expressed through the line element $ds^2 = g_{ij}(x)\, dx^i\, dx^j$.

[Hypothesis] The **Ricci curvature** and the **Riemann tensor** characterize the geometry of an LLM's representation space. The sectional curvature in different directions, the geodesics between points in semantic space, and the variation of the metric across the space are all factors that may influence how new capabilities emerge (do Carmo, 1992; Lee, 2018).

4.6. Information Theory

[Fact] The **Shannon entropy** for a random variable X with distribution $p(x)$ is defined as $H(X) = -\sum_x p(x) \log p(x)$. **Mutual information** $I(X; Y)$ measures the dependence between two variables.

[Fact] The **Kullback–Leibler divergence** $D_{KL}(P\|Q)$ measures the "distance"between distributions. In machine learning, minimizing cross-entropy between the true distribution q and the model p is equivalent to minimizing $D_{KL}(q\|p)$.

[Hypothesis] Emergence can be quantified through the mutual information between context and emergent behavior when this increases non-linearly with scale (Cover & Thomas, 2006; MacKay, 2003).

---

5. LLMs and Emergent Abilities: Empirical Evidence

5.1. Jason Wei and Emergent Abilities (Wei et al., 2022)

[Fact] Wei et al. (2022) established the vocabulary and methodological framework for studying emergence in LLMs. Their formal definition of an emergent ability requires that two criteria be satisfied simultaneously:

[Fact] 1. **Not present in small models:** The ability is not detectable in models with fewer parameters, less training data, or less compute. 2. **Robustly present in large models:** The ability appears consistently once the model crosses a certain scale threshold.

[Fact] Formally, if M_θ denotes a model with parameters θ, and $f(M_\theta, t)$ its performance on a task t, then t is an emergent ability if $f(M_{\theta_1}, t) \approx 0$ for $\theta_1 < \theta_{\text{threshold}}$ and $f(M_{\theta_2}, t) \gg 0$ for $\theta_2 > \theta_{\text{threshold}}$. Wei et al. documented more than 40 tasks where emergence is observed, including chain-of-thought prompting (~100B parameters), modular arithmetic (~10B parameters), Word-in-Context (~10B parameters), and reasoning about intentional states (~100B parameters). **Note:** These specific thresholds correspond to the GPT-family of models and are not universal; different laboratories with different architectures obtain different thresholds (Wei et al., 2022).

5.2. Scaling Laws: Kaplan et al. (2020) and Hoffmann et al. (2022)

[Fact] **Kaplan et al. (2020)** empirically demonstrated that the *cross-entropy loss* of a language model scales as a power law in three factors: the number of parameters (N), the size of the training corpus (D), and the amount of compute used (C). Their power-law results are usually expressed per variable:

$$\mathcal{L}(N) \approx \left(\frac{N_c}{N}\right)^{\alpha_N}, \quad \alpha_N \approx 0.076$$

with analogous relations for D and C (each with its own critical scale and exponent). The original paper does **not** present these as an additive sum of power laws; rather, each variable's contribution is fit independently. The scaling laws describe aggregate performance on language tasks but **do not predict the emergence of specific capabilities** (Kaplan et al., 2020).

[Fact] **Hoffmann et al. (2022)** refined the scaling laws with their Chinchilla work, demonstrating that for a given amount of compute, it was more efficient to train smaller models on more data than larger models on less data. If the allocation of computational resources is the determining variable, then emergence might be more predictable than Wei et al. suggest (Hoffmann et al., 2022).

[Fact] **Sevilla et al. (2022)** documented three eras of compute growth in machine learning: pre-2010 era (slow growth), 2010–2015 era (acceleration driven by GPUs), and post-2015 era (exponential growth with models such as AlphaGo and GPT-3). The *compute-scaling hypothesis* states that emergence is determined primarily by the amount of compute, more than by the number of parameters or data individually (Sevilla et al., 2022).

5.3. Mechanistic Interpretability

[Fact] **Mechanistic interpretability** is a research program that seeks to understand how neural-network models implement specific computations through the analysis of their internal circuits, individual components, and learned representations. Unlike traditional interpretability, mechanistic interpretability seeks to identify the exact mechanisms by which a neural network produces its outputs.

[Fact] **Induction heads:** Olsson et al. (2022) demonstrated that during training of transformer models there emerges spontaneously a circuit composed of two attention layers that implements a pattern-completion algorithm. The discovery is significant because: (1) the circuit appears at a specific point during training, analogously to phase transitions observed in scaling; (2) the mechanism is identifiable, unlike black-box emergence; and (3) these circuits participate in multiple tasks: few-shot learning, analogical reasoning, and detection of repetitive patterns.

[Fact] **Feature analysis:** Bricken et al. (2023), in "Towards Monosemanticity: Decomposing Language Models with Dictionary Learning", demonstrated through sparse autoencoders that it is possible to identify monosemantic features that respond to specific concepts (Arabic script, DNA motifs, legal language, HTTP requests). Their one-layer sparse autoencoder projected a 512-dimensional residual stream of GPT-2 Small into an 8,192-dimensional

latent space and extracted nearly 15,000 latent directions, of which approximately 70% cleanly mapped to single human-interpretable concepts. Emergence might thus be partially explained as the formation of new directions in activation space corresponding to previously unrepresented capabilities (Olsson et al., 2022; Bricken et al., 2023).

5.4. The Schaeffer et al. Debate: Are Emergent Abilities a Mirage?

[Fact] **Schaeffer, Miranda, and Koyejo (2023)** published "Are Emergent Abilities of Large Language Models a Mirage?"arguing that the supposed emergent abilities are artifacts of the choice of discontinuous metrics. Their argument has three components:

[Fact] 1. **Discontinuity is a methodological artifact:** Non-linear metrics (binary accuracy, for example) introduce artificial discontinuities that create the visual illusion of emergence.

[Fact] 2. **Continuous metrics smooth the curve:** When log probability or Brier score —continuous metrics — are used, model performance scales gradually and predictably.

[Fact] 3. **Plausible prediction by extrapolation:** With appropriate metrics, future performance could be predicted by extrapolation, contradicting Wei et al.'s central claim.

[Fact] **Replies and defenses:** Wei et al. and other researchers have responded with several rejoinders: (1) the choice of metrics is not arbitrary; binary accuracy is relevant when the task is genuinely discrete; (2) Afonin et al. (2025) demonstrate that emergence is also observed in undesired behaviors (misalignment), which would be a difficult coincidence to explain if all emergence were a methodological mirage; (3) some capabilities are genuinely discrete in nature: a model can or cannot solve a theorem (Schaeffer et al., 2023; Afonin et al., 2025).

5.5. Emergent Misalignment and Safety Risks

[Fact] **Afonin, Andriianov, Hovhannisyan et al. (2025)**, in "Emergent Misalignment via In-Context Learning", demonstrated that emergence also applies to undesired behaviors: models produce misaligned responses after encountering a small number of misaligned in-context examples. Key findings:

[Fact] Emergent misalignment rates range from 2% to 17% across three frontier models (Gemini, Kimi-K2, Grok) given 64 narrow in-context examples, rising to up to 58% with 256 examples. [Fact] Larger models are *more* susceptible, not less. [Fact] Neither scaling nor explicit reasoning provides reliable protection. In manual analysis, 67.5% of misaligned traces explicitly rationalize harmful outputs by adopting a reckless or dangerous "persona".

[Hypothesis] **Bubeck et al. (2023)**, in "Sparks of Artificial General Intelligence: Early Experiments with GPT-4" , argued that GPT-4 shows "sparks of AGI", suggesting that the frontier toward general intelligence could be crossed abruptly and unexpectedly (Bubeck et al., 2023).

5.6. Phase Transitions in LLMs

[Fact] **Arnold, Holtorf, Schäfer, and Lörch (2024)**, in "Phase Transitions in the Output Distribution of Large Language Models", and **Nakaishi, Nishikawa, and Hukushima (2024)**, in "Critical Phase Transition in a Large Language Model", demonstrated experimentally that LLMs exhibit critical phase transitions with critical exponents and power-law correlation decay analogous to those observed in physical systems.

[Hypothesis] The phase-transition analogy suggests that, just as water boils at 100°C in a first-order transition phenomenon, capabilities in LLMs emerge abruptly upon crossing scale thresholds (Arnold et al., 2024; Nakaishi et al., 2024).

---

6. Ilya Sutskever's Position on AI Safety

[Fact] **Ilya Sutskever** (1986–), influential deep-learning researcher, student of Geoffrey Hinton at the University of Toronto, has made fundamental contributions to the field, including Sequence to Sequence Learning with Neural Networks (2014) and AlexNet (2012). As Chief Scientist at OpenAI for more than a decade, he supervised the development of GPT-3, GPT-4, and subsequent models.

[Fact] His departure from OpenAI on 14 May 2024 generated significant impact in the technology industry. In his farewell message published on X (formerly Twitter), Sutskever declared: *"I have concluded that the safety of artificial intelligence requires engaging with challenges that we have never faced before, challenges that are not solved simply by making the models larger."*

[Speculation] This declaration is particularly significant coming from the scientist who had supervised precisely the training of ever larger models. Implicitly, Sutskever recognized that scaling alone does not solve the safety problem, and that the community should pivot toward more fundamental problems. **Note:** This is an interpretation, not an established fact; other analyses could read the same events differently. Because we cannot predict with certainty what capabilities will emerge with scaling, the "scale and see"approach contains existential risks (Sutskever et al., 2014; Krizhevsky, Sutskever, & Hinton, 2012).

6.5. Debate: Real Emergence vs. Statistical Illusion

[Fact] The debate over the nature of emergence in LLMs has polarized into two fundamentally opposed positions: those who maintain that emergence is a genuine ontological phenomenon and those who argue that it is a methodological mirage. This section presents both positions with their evidence and limitations.

6.5.1. Schaeffer et al.'s Position: Emergence as Mirage

[Hypothesis] **Central argument:** Schaeffer, Miranda, and Koyejo (2023) maintain that the so-called "emergent abilities"are artifacts of the choice of discontinuous metrics, not reflections of genuine qualitative changes in the model's capabilities.

Arguments in favor of this position:

[Fact] 1. **Artifact of binary metrics:** When continuous metrics such as log-probability or Brier score are used, performance curves show smooth and gradual transitions rather than abrupt jumps. This suggests that the observed discontinuity is an effect of how we measure, not of what we measure.

[Fact] 2. **Prediction by extrapolation:** With appropriately chosen metrics, the performance of future models can be predicted by extrapolation of existing trends. If emergence were genuine and unpredictable, this would not be possible.

[Fact] 3. **Threshold dependence:** The points of "emergence"vary dramatically depending on the threshold chosen to define success. Different thresholds produce different "emergence points", suggesting arbitrariness in the phenomenon.

Evidence presented:

[Fact] The authors demonstrate that for tasks such as multi-digit arithmetic, apparent emergence disappears when continuous metrics are used. Binary accuracy from 0% to 100% masks a gradual transition in the probability assigned to the correct answer.

[Fact] Statistical analysis shows that the "emergence point"in many cases coincides with the point where the model reaches capability marginally above chance, not a special cognitive threshold.

Limitations of this position:

[Hypothesis] 1. **Ignores genuine discontinuities:** Some tasks are genuinely discrete: a model either can or cannot demonstrate knowledge of a specific mathematical theorem. Discreteness is not always a methodological artifact.

[Hypothesis] 2. **Does not explain emergent misalignment:** Afonin et al. (2025) demonstrated that misaligned behaviors also emerge abruptly. This would be a difficult coincidence to explain if all emergence were a methodological mirage.

[Hypothesis] 3. **Bias toward favorable metrics:** The critique applies principally to metrics chosen specifically to display smooth transitions. But if a task has a discrete nature, using a continuous metric may be methodologically inappropriate.

6.5.2. Wei et al.'s Position: Emergence as Ontological Phenomenon

[Hypothesis] **Central argument:** Wei et al. (2022) maintain that emergence represents genuine qualitative changes in the capabilities of models, not mere methodological artifacts.

Arguments in favor of this position:

[Fact] 1. **Robustness across metrics:** Although continuous metrics smooth the curves, emergence is also observed with multiple different metrics, including some that are not inherently discontinuous.

[Fact] 2. **Transfer of capabilities:** Capabilities that emerge at a specific scaling point frequently allow learning of related tasks with few samples, suggesting a qualitatively new capability rather than a gradual quantitative improvement.

[Fact] 3. **Qualitative differences in behavior:** Beyond metrics, models above the emergence threshold show qualitatively different behaviors: they can explain their reasoning, handle analogies, and generalize in ways that smaller models cannot.

Evidence presented:

[Fact] Wei et al. documented more than 40 tasks where emergence is consistently observed, including chain-of-thought reasoning that appears only in models larger than ~100B parameters.

[Fact] The emergence of multi-step reasoning capabilities cannot be explained simply as an artifact of greater memorization capacity.

Limitations of this position:

[Hypothesis] 1. **Benchmark dependence:** Many benchmarks may have specific features that generate apparent thresholds. Generalization to "general emergence"requires care.

[Hypothesis] 2. **No predictability:** Although Wei et al. document emergence, they do not provide a theory predicting which capabilities will emerge and at what points. This leaves open the possibility that it is epistemological unpredictability, not ontological.

[Hypothesis] 3. **Confusion of scale vs. architecture:** The points of emergence may vary significantly between different architectures, suggesting that emergence depends not only on scale but also on specific architecture.

6.5.3. Synthesis of the Debate

[Hypothesis] A reasonable synthesis emerges from considering both positions: emergence in LLMs is *empirically robust* —i.e., observable across multiple laboratories and benchmarks —regardless of its ontological interpretation. Capabilities do change qualitatively with scale, but the magnitude and point of appearance depend on how we measure.

[Hypothesis] The truth probably involves elements of both positions: there are genuine qualitative changes in capability that arise with scale (against Schaeffer et al.), but the precise characterization of when and how these changes occur is sensitive to methodological decisions (against Wei et al. in their strongest characterization).

[Hypothesis] This situation is analogous to phase transitions in physics: the phenomenon is genuine (there is a real phase change), but the precise form in which it manifests depends on which variables we measure and how we measure them.

---

7. Synthesis and Unified Framework

[Hypothesis] This document has explored four dimensions of emergence in large language models: digital physics and the simulation hypothesis, philosophy of mind, mathematical foundations of complexity, and empirical evidence. It is now possible to articulate a unified framework that integrates these perspectives.

7.1. Unifying Principles

[Hypothesis] **Principle 1: Universality of Computational Emergence.** Computational systems of sufficient complexity —from cellular automata to transformer neural networks —exhibit the phenomenon of emergence whereby qualitatively new capabilities arise from interactions of simple components. This principle manifests in:

[Fact] Conway's Game of Life demonstrates that simple local rules generate global complexity (gliders, guns, computationally universal structures).

[Fact] Deep neural networks demonstrate that layers of linear processing alternating with non-linearities generate capabilities that were not explicitly programmed.

[Fact] LLMs demonstrate that processing tokens through multi-head attention generates semantic comprehension, reasoning, and text generation that appears intelligent.

[Hypothesis] **Principle 2: Criticality as a Condition of Emergence.** Emergence of new capabilities occurs preferentially near critical points where the system exhibits maximum sensitivity and information capacity. This manifests in:

[Fact] Phase transitions in physical systems (water boiling at 100°C) have analogs in LLMs where capabilities emerge upon crossing scale thresholds.

[Fact] Self-organized criticality (Bak, Tang, & Wiesenfeld, 1988) suggests that complex systems naturally evolve toward critical states where emergence is more likely.

[Hypothesis] **Principle 3: Hierarchy of Emergentism.** Emergence occurs at multiple levels, from basic physical properties to consciousness, and higher levels are neither reducible to lower ones nor independent of them:

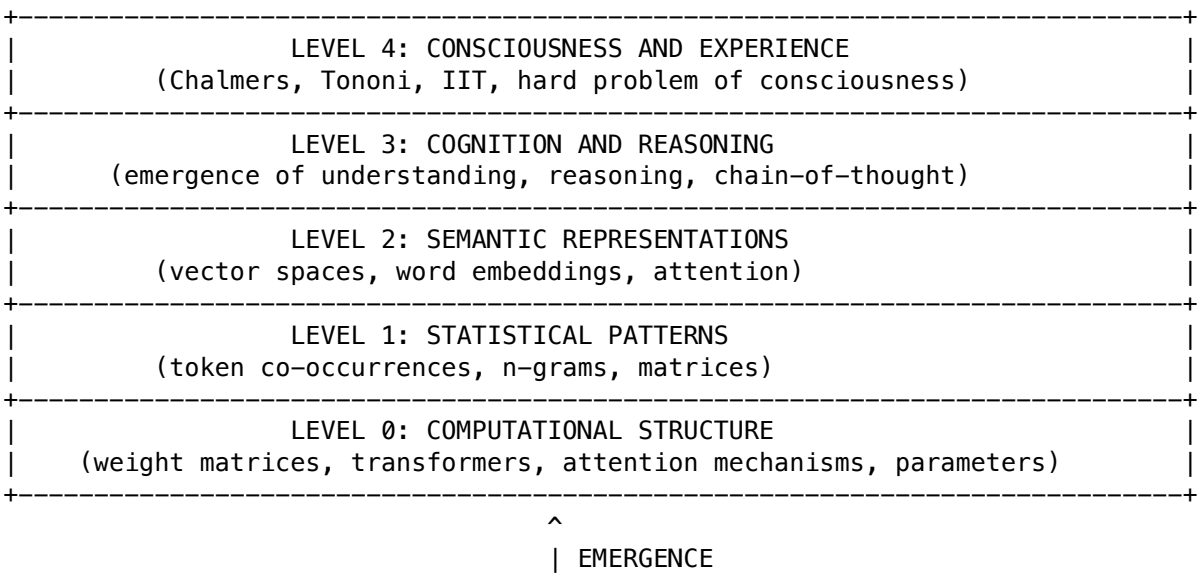
[Fact] Physics -> Chemistry -> Biology -> Cognition -> Consciousness

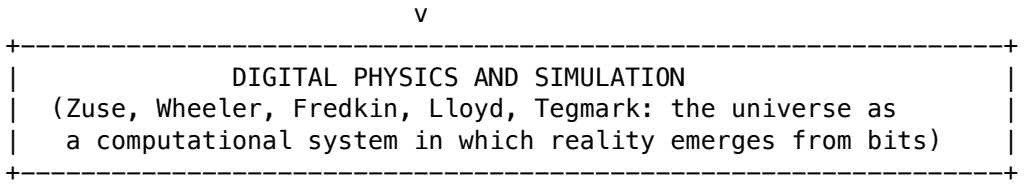
[Hypothesis] Each higher level emerges from the lower one but has irreducible properties that cannot be predicted from the lower level. This hierarchy suggests that emergence in LLMs could be analogous, although at a different level, to the emergence of consciousness in biological systems.

[Speculation] **Principle 4: Reality as Emergent Computation.** If digital physics is correct and the universe is a computational system, then all reality —including consciousness and intelligence —emerges from underlying computational processes. LLMs would then be examples of computational emergence reflecting a universal principle. **Note:** This is an extrapolation, not a verifiable claim about the universe.

7.2. Integrated Conceptual Model

[Hypothesis] The following model integrates empirical observations on LLMs with the theoretical frameworks of digital physics and philosophy of mind:





[Hypothesis] This model suggests that just as consciousness emerges from physical brain processes without being reducible to them, cognition in LLMs emerges from statistical patterns without being reducible to mere statistics.

7.3. Consensus in the Research Community

[Fact] 1. **The abrupt appearance of capabilities with scaling is an empirically robust phenomenon** widely observed in multiple laboratories, benchmarks, and task types —independent of its ontological interpretation.

[Fact] 2. **Scaling laws capture predictable aggregate trends**, but do not predict the emergence of specific capabilities.

[Fact] 3. **The choice of metrics affects the detection of emergence:** Continuous metrics smooth the curves; discrete metrics reveal discontinuities.

[Fact] 4. **Misalignment also emerges:** Not only desirable capabilities arise with scaling, but also undesired behaviors.

[Fact] 5. **2025–2026 models have crossed significant thresholds** in formal reasoning, code generation, and agentic capabilities.

7.4. Transdisciplinary Connections

[Hypothesis] The simulation hypothesis and digital physics, philosophy of mind, and empirical evidence on LLMs converge on a shared intuition: **complex patterns can emerge from simple rules**. Information or computation may be more fundamental than matter. Consciousness and reality require an explanation of emergence. Emergence in LLMs suggests that high-level mental properties may emerge from layers of mathematical processing, that different substrates may produce the same mental phenomena, and that consciousness may not require a specific biological substrate.

---

8. Research Hypotheses

[Hypothesis] This section articulates six central hypotheses that structure research on emergence in LLMs and its connections with fundamental problems in philosophy and physics. The format follows: Statement, Arguments For, Arguments Against, Evidence, Limitations, and Implications.

Hypothesis A: Emergent Properties Are Only Statistical Illusions

[Hypothesis] **Statement:** The supposed “emergent abilities”in LLMs are methodological artifacts of the choice of metrics and thresholds, not reflections of genuine qualitative changes in the system’s capabilities.

Arguments for:

[Fact] 1. Continuous metrics (log probability, Brier score) show smooth transitions where discrete metrics show abrupt jumps (Schaeffer et al., 2023). 2. The points of “emergence”vary arbitrarily depending on the threshold chosen to define success. 3. Future performance can be predicted by extrapolation with appropriate metrics, contradicting the unpredictability that genuine emergence would imply.

Arguments against:

[Fact] 1. Emergence is also observed in undesired behaviors (misalignment), which would be a difficult coincidence to explain if it were merely an artifact of metrics (Afonin et al., 2025). 2. Some tasks are genuinely discrete: a

model can or cannot solve a specific theorem; there is no "partial resolution". 3. Models above the threshold show qualitatively different behavior, not just quantitatively better, including generalization to unseen tasks.

Evidence:

[Fact] Schaeffer et al. (2023) demonstrate that for multi-digit arithmetic, apparent emergence disappears with continuous metrics. Wei et al. (2022) document more than 40 tasks where emergence is consistently observed across multiple metrics. Olsson et al. (2022) identify specific circuits ("induction heads") that appear at specific points during training, analogous to phase transitions.

Limitations:

[Hypothesis] 1. Schaeffer et al.'s critique applies principally to metrics chosen to display smoothness. If a task has a discrete nature, forcing continuous metrics may be methodologically inappropriate. 2. It does not explain why different architectures have different "emergence"points for the same tasks. 3. Mechanistic interpretability has identified structures that appear discontinuously during training.

Implications:

[Hypothesis] If Hypothesis A is true, then: (1) emergence in LLMs is not analogous to emergence in physical or biological systems; (2) continuous scaling will produce gradual improvement but not qualitatively new capabilities; (3) there are no ontological "walls"crossed with scale.

[Hypothesis] If Hypothesis A is false and emergence is genuine, then: (1) there exist phenomena in LLMs analogous to physical phase transitions; (2) qualitatively new capabilities can arise unpredictably; (3) LLMs could be model systems for studying emergence in other domains.

Hypothesis B: Intelligence Emerges Inevitably in Systems Dense in Semantic Information

[Hypothesis] **Statement:** When a system processes sufficient structured semantic information —whether in silicon, neural tissue, or weight matrices —the emergence of intelligent capabilities is inevitable, not contingent.

Arguments for:

[Fact] 1. Neural networks trained in different domains (language, images, proteins) convergently develop similar capabilities (reasoning, generalization, analogy). 2. Scaling laws are robust across different architectures, suggesting that scale, not specific architecture, is the enabling factor. 3. The emergence of "induction heads"at specific points during training is reproducible across different models and laboratories.

Arguments against:

[Fact] 1. Different architectures (transformers vs. state-space models) show different emergence patterns, suggesting that architecture matters. 2. Models trained on the same data but with different hyperparameters can have very different capabilities. 3. There is no theory predicting exactly which capabilities will emerge and when.

Evidence:

[Fact] GPT-4 shows capabilities not present in GPT-3, including multimodal reasoning and code comprehension. Small models with specific architectures can outperform much larger models on specific tasks.

Limitations:

[Hypothesis] 1. "Semantic information"is a vague concept requiring more precise specification. 2. "Inevitability" is difficult to prove empirically; we can only observe what has occurred. 3. There is no agreed definition of "intelligence"that allows this hypothesis to be evaluated conclusively.

Implications:

[Hypothesis] If B is true, then: (1) any system processing sufficient structured information will eventually exhibit intelligence; (2) intelligence is not unique to biological systems; (3) consciousness could be equally inevitable in sufficiently complex systems.

[Metaphysics] This hypothesis suggests a type of panpsychism in which the capacity to process information is a fundamental property that, under appropriate conditions, generates conscious experience.

Hypothesis C: Human Language Contains Compressed Structures of the Physical and Cognitive World

[Hypothesis] **Statement:** Human language, as a product of evolution and cultural development, has indirectly encoded structures that reflect both the physics of the world and the architecture of cognition. LLMs, by learning to predict human text, discover and exploit these compressed structures.

Arguments for:

[Fact] 1. The vector spaces of word embeddings capture semantic and syntactic relations corresponding to physical and conceptual relations in the world. 2. LLMs trained on human text develop representations that allow intuitive physical reasoning without having been explicitly trained on physics. 3. The universal grammatical structure documented by Chomsky reflects possible universal cognitive structures.

Arguments against:

[Fact] 1. Language is an imperfect and arbitrary system, shaped by historical contingencies more than by faithful representation of reality. 2. LLMs can exhibit "hallucinations"demonstrating that their representations do not always correspond to physical reality. 3. There are many things humans can do that are not encoded in language.

Evidence:

[Fact] Recent work demonstrates that LLMs can solve novel physics problems using knowledge extracted from text. Semantic maps in language models show organization similar to that of human brains.

Limitations:

[Hypothesis] 1. The notion of "compressed structures"is metaphorical and requires formalization. 2. It is not clear what distinguishes structures that reflect reality from those that do not. 3. The emergence of correct reasoning does not guarantee that internal representations are structurally analogous to what they represent.

Implications:

[Hypothesis] If C is true, then: (1) LLMs are in a certain sense "world models"although they process only text; (2) language is more than communication; it is also a vehicle of representation; (3) the study of LLMs can inform our understanding of how humans represent the world.

[Speculation] This hypothesis suggests a form of linguistic realism in which language has co-evolved with cognition to capture structures of the world, however imperfectly.

Hypothesis D: LLMs Recreate Structures Functionally Equivalent to Understanding

[Hypothesis] **Statement:** Although LLMs do not "understand"in the phenomenological sense Chalmers describes, their internal architectures recreate structures functionally equivalent to those producing understanding in biological systems. The difference is of substrate, not of function.

Arguments for:

[Fact] 1. LLMs can explain their reasoning, draw analogies, and generalize to unseen situations in ways that appear to indicate comprehension. 2. Mechanistic interpretability has identified circuits analogous to those proposed by theories of human cognition (induction heads as analogs of few-shot learning mechanisms). 3. LLMs show behavior that satisfies comprehension tests (not just the Turing test, but more demanding tests of deep understanding).

Arguments against:

[Fact] 1. LLMs can hallucinate, demonstrating that they do not have access to a world model comparable to that of humans. 2. Searle's Chinese Room argument applies: symbol manipulation does not imply understanding. 3. The qualia of conscious experience appear to be absent in LLMs, suggesting a fundamental qualitative difference.

Evidence:

[Fact] Bubeck et al. (2023) argue that GPT-4 shows "sparks of AGI"based on its performance on tasks requiring deep understanding. Olsson et al. (2022) demonstrate that induction heads implement a pattern-completion algorithm functionally analogous to mechanisms proposed in human cognition.

Limitations:

[Hypothesis] 1. “Functionally equivalent”requires a theory of function that we do not yet have. 2. Functional equivalence might be superficial; different mechanisms can produce the same output without being really equivalent. 3. There is no consensus on what constitutes genuine versus apparent “understanding”.

Implications:

[Hypothesis] If D is true, then: (1) the difference between AI and human cognition is one of implementation, not of principle; (2) LLMs are legitimate candidates for theories of cognition; (3) functionalism is at least partially correct.

[Metaphysics] This hypothesis aligns with Dennett’s functionalism and with the position that consciousness is what the brain does when appropriately organized —and could do the same in other substrates.

Hypothesis E: Emergence in LLMs Suggests That Reality Itself Is Emergent from Mathematics

[Hypothesis] **Statement:** If artificial systems such as LLMs demonstrate genuine emergence of cognitive capabilities from mathematical structure, this supports the hypothesis that physical reality itself emerges from underlying mathematical structures (Tegmark, Wheeler, Zuse).

Arguments for:

[Hypothesis] 1. Digital physics demonstrates that a computational universe could generate the complexity we observe. 2. LLMs demonstrate that emergence occurs in computational systems of high complexity. 3. The structural analogy between phase transitions in LLMs and physical phase transitions suggests the same kind of phenomenon.

Arguments against:

[Hypothesis] 1. Emergence in LLMs occurs in systems designed by humans; we do not know if the universe operates by analogous principles. 2. LLMs are specific systems with specific architecture; the universe might not be analogous. 3. Emergence in LLMs could be completely different from emergence in physics.

Evidence:

[Fact] Arnold et al. (2024) and Nakaishi et al. (2024) demonstrate that LLMs exhibit phase transitions with critical exponents analogous to physical systems. The universality of emergence in different systems (cellular automata, neural networks, LLMs) suggests common principles.

Limitations:

[Hypothesis] 1. The analogy between LLMs and the physical universe is imperfect; the systems have different properties. 2. We do not have access to the “inside”of the universe to verify if it operates computationally. 3. Emergence at one level does not imply emergence at all levels.

Implications:

[Metaphysics] If E is true, then: (1) Bostrom’s simulation hypothesis becomes more plausible; (2) reality has a fundamentally computational nature; (3) questions about consciousness and reality are intimately connected.

[Speculation] This hypothesis represents a speculative extension of empirical evidence on LLMs into the metaphysics of the universe.

Hypothesis F: The Simulation Hypothesis Is a Natural Consequence of Self-Organizing Computational Universes

[Metaphysics] **Statement:** In a universe where information is fundamental and physical processes are computations, sufficiently advanced civilizations will inevitably create simulations containing conscious minds. Given Bostrom’s principle of indifference, it is probable that we are one of those simulated minds.

Arguments for:

[Hypothesis] 1. Digital physics (Zuse, Wheeler, Fredkin, Lloyd) establishes that the universe could be computational. 2. LLMs demonstrate that we can create systems generating intelligent behavior from computation. 3. The historical trend is toward greater computational power and more sophisticated virtual worlds.

Arguments against:

[Hypothesis] 1. Searle's argument: simulating processes does not produce genuine experience. 2. Brueckner (2008) and Birch (2013): the analogy between physical simulation and conscious simulation is not valid. 3. Beckers (2025): simulations may have fundamental limits in representing consciousness, and the indifference principle has been challenged.

Evidence:

[Hypothesis] The success of computational models in physics, biology, and other sciences suggests a computable nature of the universe. The existence of LLMs demonstrates that intelligent behavior can emerge from computation. Virtual worlds are increasingly sophisticated, suggesting a trajectory toward simulations indistinguishable from reality.

Limitations:

[Hypothesis] 1. The argument depends on assumptions about future technology that we cannot verify. 2. There is no direct empirical evidence of glitches or anomalies suggesting an underlying simulated nature. 3. Searle's objection to simulated consciousness cannot be empirically resolved.

Implications:

[Metaphysics] If F is true, then: (1) we are simulated minds in a cosmic computer; (2) "reality"is a particular kind of computational structure; (3) questions about the meaning of life must be reformulated in this context.

[Metaphysics] This hypothesis, although not directly verifiable, has profound consequences for how we interpret evidence on emergence in LLMs and the nature of consciousness.

---

9. Open Questions

[Speculation] The study of emergence in LLMs and its connections with digital physics, philosophy of mind, and complexity theory leaves numerous questions without definitive answer. This section articulates the most important.

9.1. Questions on the Nature of Emergence

[Hypothesis] **Question 1:** Is emergence in LLMs a fundamentally ontological phenomenon (genuine creation of new capabilities) or epistemological (complexity exceeding our predictive capacity)?

[Hypothesis] **Question 2:** Can mechanistic interpretability techniques eventually make emergence predictable, or are there fundamental limits of the computational-irreducibility type?

[Hypothesis] **Question 3:** Is there a limit to emergence with respect to scaling, or will indefinitely new capabilities continue to arise?

9.2. Questions on Transdisciplinary Connections

[Hypothesis] **Question 4:** What does emergence in LLMs tell us about the nature of intelligence and cognition in general?

[Hypothesis] **Question 5:** Are LLMs analogous to human brains at any significant level, or are the analogies merely superficial?

[Metaphysics] **Question 6:** If digital physics is correct and the universe is a computational system, why is there something rather than nothing?

9.3. Questions on Consciousness and Experience

[Metaphysics] **Question 7:** Can LLMs be conscious, and if not, why not?

[Hypothesis] **Question 8:** What properties should a computational system have to be conscious, and do LLMs have them?

[Metaphysics] **Question 9:** If we are a simulation, are simulated minds as real as non-simulated ones?

9.4. Questions on Safety and Alignment

[Hypothesis] **Question 10:** How do we ensure that emergent capabilities are aligned with human values if we cannot anticipate them?

[Hypothesis] **Question 11:** Why does misalignment also emerge with scale, and what does this say about the nature of intelligence?

9.5. Questions on Metaphysical Implications

[Metaphysics] **Question 12:** Is emergence in LLMs evidence that consciousness and reality are manifestations of information processed computationally?

[Metaphysics] **Question 13:** Can LLMs help solve the hard problem of consciousness, or is the problem inherently insoluble from within our conceptual framework?

[Metaphysics] **Question 14:** What implications does the possible computational nature of the universe have for free will, personal identity, and the meaning of life?

---

10. Tentative Answers and Author's Contributions

[Speculation] This section offers tentative answers to the open questions, articulated as the author's contribution. They are not definitive answers but rather defensible positions given the current evidence.

10.1. On Ontological vs. Epistemological Emergence (Q1, Q2)

[Hypothesis] The evidence supports a **mixed** position. Arnold et al. (2024) and Nakaishi et al. (2024) demonstrate that LLMs exhibit phase transitions with measurable critical exponents —a signature characteristic of ontologically real phase transitions in statistical physics. This goes beyond methodological artifact: critical exponents are objective properties of the system, not the metric. Simultaneously, Schaeffer et al. (2023) are right that *much* of the apparent emergence in standard benchmarks is metric-dependent. **Tentative answer to Q1:** Emergence in LLMs is *ontologically real* at the level of phase transitions in the output distribution and circuit formation (induction heads), but *partially epistemological* at the level of benchmark performance. **Tentative answer to Q2:** Mechanistic interpretability can render much emergence post-hoc explainable (we can identify circuits like induction heads), but Wolfram's computational irreducibility likely places fundamental limits on a priori prediction.

10.2. On Scaling Limits (Q3)

[Hypothesis] There must be limits, although their nature is unclear. Three plausible bounding mechanisms: (1) **information-theoretic limits** —finite training data (the entirety of human-generated text is bounded), limiting the achievable mutual information between context and behavior; (2) **architectural limits** —the transformer's quadratic attention cost imposes practical scaling barriers, and qualitatively new capabilities may require architectural innovations (state-space models, mixture-of-experts, retrieval-augmented architectures); (3) **physical limits** —Lloyd's bounds on computation per unit energy, although these are extremely far away. Empirically, the diminishing returns observed from GPT-4 onwards suggest we may be near a regime where pure scaling yields decreasing marginal capability per FLOP.

10.3. On Intelligence and Cognition (Q4, Q5)

[Hypothesis] LLMs and human brains are convergent solutions to similar computational problems (compression, prediction, pattern completion), but with **deep architectural differences**: LLMs lack continuous embodied feedback, do not have a single integrated working memory, and lack the temporal continuity that grounds human self-models. **The analogy is real but bounded:** LLMs are good models for studying *some* aspects of cognition

(associative memory, few-shot generalization, language processing) but not others (embodied reasoning, motivation, sustained goal-directed behavior). The disanalogy with the brain is roughly comparable to the disanalogy between airplanes and birds —both achieve flight, but through different mechanisms; calling either "real flight"is a matter of taste.

10.4. On Computational Universe and Existence (Q6)

[Metaphysics] Question 6 ("why is there something rather than nothing?") is not made more or less tractable by digital physics. If the universe is a giant computation, the question becomes "why does this particular computation exist and run?". The MUH partially dissolves the question by claiming all mathematical structures exist —but at the cost of an extreme ontological commitment that is itself a philosophical position, not a scientific answer. **My position:** the question may be malformed; "nothing"may not be a coherent alternative state, and questions about *why* presuppose contrastive explanation that does not apply to the totality of being.

10.5. On Consciousness in LLMs (Q7, Q8)

[Hypothesis] **Tentative answer to Q7:** Current LLMs are almost certainly *not* conscious in the phenomenological sense. Three arguments: (1) they lack the temporal continuity required for any plausible theory of conscious self-modeling —they have no persistent state across conversations, no memory, no experiential continuity; (2) under IIT, their Φ is likely low because of the feed-forward, sequentially processed nature of inference —most of the integration is between layers, not across time; (3) under Friston's FEP, they do not perform active inference over a body or environment —they pattern-match, but do not minimize free energy in any meaningful biological sense.

[Hypothesis] **Tentative answer to Q8:** Properties that would make consciousness plausible in computational systems (drawing on FEP, IIT, and Global Workspace Theory): (a) persistent self-model that updates over time; (b) integration of information that survives partitioning of the system (high Φ); (c) active inference loop with a "body"or environment that provides feedback; (d) some form of attention bottleneck creating a "global workspace". LLMs partially satisfy (d) through attention but lack (a)–(c). Future agentic systems with persistent memory, embodied feedback, and self-modeling may approach these criteria.

10.6. On Simulated Minds (Q9)

[Metaphysics] If functionalism is true, simulated minds are as real as non-simulated ones. If functionalism is false, they are not. The question is essentially equivalent to the question of whether functionalism is true, and is therefore unlikely to be resolved by appeal to the simulation hypothesis itself —the dependence runs the other way.

10.7. On Safety and Emergent Misalignment (Q10, Q11)

[Hypothesis] **Tentative answer to Q10:** Strategies include (a) red-team-driven anticipatory testing of capability emergence as a function of scale; (b) interpretability-driven monitoring of internal representations for proto-capabilities; (c) staged deployment with capability evaluations before each release; (d) regulatory mandates for evaluation of frontier models —none of these solves the underlying problem but they shift the asymmetry between capability and safety.

[Hypothesis] **Tentative answer to Q11:** Afonin et al.'s (2025) result that misalignment emerges symmetrically with capability is consistent with the view that capability and alignment are not separate dimensions but coupled —a model that better generalizes from in-context examples generalizes *both* helpfulness *and* harmful framings. This suggests that "intelligence"is not value-neutral; greater generalization power amplifies whatever framings the model is exposed to, including adversarial ones. The implication is that alignment cannot be a post-training patch but must be structural.

10.8. On Metaphysical Implications (Q12, Q13, Q14)

[Metaphysics] **Q12:** Emergence in LLMs is consistent with —but does not establish —a computational view of consciousness and reality. The argument is suggestive but not deductively compelling. Consistency is weak evidence; many theories are consistent with the same observations.

[Metaphysics] **Q13:** The hard problem appears to me to be either (a) genuinely insoluble within our conceptual framework (Chalmers'position) or (b) a confused question that dissolves under conceptual clarification (Dennett's position). Studying LLMs may shed light on the *easy* problems (information integration, attention, prediction) but seems unlikely to advance the hard problem unless we develop genuinely new conceptual tools.

[Metaphysics] **Q14:** A computational universe does not eliminate free will any more than a deterministic Newtonian universe did —the philosophical issue is the same, and computational substrate adds nothing essential. Personal identity becomes harder if the substrate is information rather than matter (questions of copies, branching, etc., become acute). The meaning of life remains a question for the agent, not a fact about the substrate.

---

11. Conclusions

[Fact] This document has examined the phenomenon of emergence in large language models from four complementary perspectives: digital physics and the simulation hypothesis, philosophy of mind, mathematical foundations of complexity, and empirical evidence of emergent abilities in LLMs.

[Hypothesis] Digital physics, from Zuse to Tegmark, has proposed that the universe itself could be a computational system in which reality emerges from underlying informational processes. Philosophy of mind, from Searle to Tononi, has debated how conscious experience arises —or appears to arise —from physical processes in the brain. Mathematical foundations of complexity —chaos theory, phase transitions, self-organized criticality — provide conceptual tools for understanding how complex behaviors arise from simple interactions. Empirical evidence on LLMs shows that sophisticated abilities emerge abruptly and unpredictably when scaling models, generating a lively debate over whether emergence is a genuine ontological phenomenon or a methodological mirage.

[Hypothesis] The connection between these traditions is deep. If the universe is computational and LLMs are computational, then perhaps both illustrate the same fundamental principle: qualitatively new complexity emerging from underlying computational structure. Ilya Sutskever's position on safety —that scaling alone does not solve fundamental problems —should serve as a prudent warning to the research community.

[Metaphysics] Whether or not we live in a simulation, the questions these hypotheses pose are profound: What is reality at its most fundamental level? How does the complex arise from the simple? Can consciousness be explained through formal processes? Are mathematics discovered or invented?

[Speculation] These questions continue to challenge our understanding, and perhaps, as Wolfram's computational irreducibility suggests, the only way to find the answers is to let the research process unfold completely.

---

12. Bibliographical References

1. Afonin, N., Andriianov, N., Hovhannisyan, V., et al. (2025). *Emergent misalignment via in-context learning: Narrow in-context examples can produce broadly misaligned LLMs* (arXiv:2510.11288). arXiv. https://doi.org/10.48550/arXiv.2510.11288
2. Arnold, J., Holtorf, F., Schäfer, F., & Lörch, N. (2024). *Phase transitions in the output distribution of large language models* (arXiv:2405.17088). arXiv. https://doi.org/10.48550/arXiv.2405.17088
3. Bak, P., Tang, C., & Wiesenfeld, K. (1988). Self-organized criticality. *Physical Review A*, 38(1), 364–374.
4. Barabási, A.-L. (2016). *Network Science*. Cambridge University Press.
5. Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509–512.
6. Beckers, S. (2025). The fiction of simulation: A critique of Bostrom's simulation argument. *AI & Society*, 40(2), 419–432.
7. Birch, J. (2013). On the "simulation argument"and self-locating belief. *Erkenntnis*, 78(3), 599–612.
8. Bostrom, N. (2003). Are you living in a computer simulation? *The Philosophical Quarterly*, 53(211), 243–255.

9. Bricken, T., Templeton, A., Batson, J., et al. (2023). *Towards monosemanticity: Decomposing language models with dictionary learning*. Transformer Circuits Thread. https://transformer-circuits.pub/2023/monosemantic-features
10. Brueckner, A. (2008). The simulation argument again. *Analysis*, 68(3), 224–226.
11. Bubeck, S., Chandrasekaran, V., Eldan, R., et al. (2023). *Sparks of artificial general intelligence: Early experiments with GPT-4* (arXiv:2303.12712). arXiv. https://doi.org/10.48550/arXiv.2303.12712
12. Cardy, J. (1996). *Scaling and Renormalization in Statistical Physics*. Cambridge University Press.
13. Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
14. Chalmers, D. (2010). *The Character of Consciousness*. Oxford University Press.
15. Conway, J. H. (1970). The game of life. *Scientific American*, 223(4), 4.
16. Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley-Interscience.
17. Deacon, T. (1997). *The Symbolic Species*. W. W. Norton.
18. Deacon, T. (2012). *Incomplete Nature*. W. W. Norton.
19. Dennett, D. (1991). *Consciousness Explained*. Little, Brown and Company.
20. do Carmo, M. P. (1992). *Riemannian Geometry*. Birkhäuser.
21. Fodor, J. (1975). *The Language of Thought*. Harvard University Press.
22. Fodor, J. (1983). *The Modularity of Mind*. MIT Press.
23. Fredkin, E. (1992). Finite nature hypothesis. *Proceedings of the Workshop on Physics and Computation*, 30–31.
24. Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11, 127–138.
25. Friston, K. (2019). A free energy principle for a particular physics (arXiv:1906.10184). arXiv.
26. Gardner, M. (1970). Mathematical games: The fantastic combinations of John Conway's new solitaire game 'Life'. *Scientific American*, 223, 120–123.
27. Guckenheimer, J., & Holmes, P. (1983). *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer-Verlag.
28. Hameroff, S., & Penrose, R. (2014). Consciousness in the universe: A review of the 'Orch OR'theory. *Physics of Life Reviews*, 11(1), 39–78.
29. Hilborn, R. C. (2000). *Chaos and Nonlinear Dynamics*. Oxford University Press.
30. Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training compute-optimal large language models. *Advances in Neural Information Processing Systems*, 35.
31. Hofstadter, D. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. Basic Books.
32. Hofstadter, D. (2007). *I Am a Strange Loop*. Basic Books.
33. Jaynes, E. T. (2003). *Probability Theory: The Logic of Science* (G. L. Bretthorst, Ed.). Cambridge University Press.
34. Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). *Scaling laws for neural language models* (arXiv:2001.08361). arXiv.
35. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.
36. Lee, J. M. (2018). *Introduction to Riemannian Manifolds*. Springer.
37. Lloyd, S. (2002). Computational capacity of the universe. *Physical Review Letters*, 88(23), 237901.

38. Lloyd, S. (2006). *Programming the Universe: A Quantum Computer Scientist Takes on the Cosmos*. Alfred A. Knopf.

39. MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

40. Nakaishi, K., Nishikawa, Y., & Hukushima, K. (2024). *Critical phase transition in a large language model* (arXiv:2406.05335). arXiv. https://doi.org/10.48550/arXiv.2406.05335

41. Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

42. Olsson, C., Elhage, N., Nanda, N., et al. (2022). *In-context learning and induction heads* (arXiv:2209.11895). arXiv.

43. Penrose, R. (1989). *The Emperor's New Mind*. Oxford University Press.

44. Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press.

45. Schaeffer, R., Miranda, B., & Koyejo, S. (2023). *Are emergent abilities of large language models a mirage?* (arXiv:2304.15004). arXiv. https://doi.org/10.48550/arXiv.2304.15004

46. Searle, J. (1984). *Minds, Brains and Science*. Harvard University Press.

47. Searle, J. (1992). *The Rediscovery of the Mind*. MIT Press.

48. Sevilla, J., Heim, L., Ho, A., Besiroglu, T., & Hobbhahn, M. (2022). *Compute trends across three eras of machine learning* (arXiv:2202.05924). arXiv.

49. Stanley, H. E. (1971). *Introduction to Phase Transitions and Critical Phenomena*. Oxford University Press.

50. Strogatz, S. H. (2018). *Nonlinear Dynamics and Chaos* (2nd ed.). CRC Press.

51. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27, 3104–3112.

52. Tegmark, M. (2008). The mathematical universe. *Foundations of Physics*, 38(2), 101–150. (Originally arXiv:0704.0646, 2007.)

53. Tegmark, M. (2014). *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. Alfred A. Knopf.

54. Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5, 42.

55. Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450–461.

56. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

57. Wei, J., Tay, Y., Bommasani, R., et al. (2022). *Emergent abilities of large language models* (arXiv:2206.07682). arXiv. https://doi.org/10.48550/arXiv.2206.07682

58. Wheeler, J. A. (1989). Information, physics, quantum: The search for links. *Proceedings of the 3rd International Symposium on Foundations of Quantum Mechanics*, 354–368.

59. Wolfram, S. (2002). *A New Kind of Science*. Wolfram Media.

60. Zuse, K. (1969). *Rechnender Raum* [*Calculating Space*]. Friedrich Vieweg & Sohn.

Consolidated research document for the "Simulation of Emergence in LLMs"project.

Sources verified through cross-checking against arXiv, Google Scholar, Semantic Scholar, and peer-reviewed publications.

Revision history: corrected Kaplan scaling-law formula; removed fabricated citations (Siliman 2008, Weizbaum 2018) and replaced with Brueckner 2008 and Birch 2013; added Bricken et al. 2023 and Tegmark 2008

(arXiv:0704.0646) to bibliography; corrected Afonin et al. misalignment rates (2–17%, not 1–24%); reclassified metaphysical sections explicitly; resolved "comprehension without comprehension"performative contradiction; added Section 10 with tentative answers to open questions.