Bad Reasoners, the Turing Trap & the Problem of Artificial Dualism

Gonçalo Hora de Carvalho[1] and Kristinn R. Thórisson[2,1]

[1] Icelandic Institute for Intelligent Machines, Menntavegur 1, Venus, 2nd fl., 101 Reykjavík, Iceland
goncalo@iiim.is ORCID: 0000-0001-8776-4852
[2] Center for Analysis & Design of Intelligent Agents, Reykjavik University
thorisson@ru.is ORCID: 0000-0003-3842-0564

Abstract. If it looks like a duck, swims like a duck, and quacks like a duck, then your LLM's priors are likely to predict the next tokens to amount to the word "duck" based on its learned data distribution—but has it reasoned about its data to deduce the duck? Large language models (LLMs) produce remarkably fluent text, enough to result in widespread claims of their ability to "understand" and "reason". However, a dissection of the key architectural features of LLMs, and more generally ANNs, in particular their reliance on probabilistic pattern-matching, exposes their absence of critical structures analogous to the neuro-biological substrates known to be involved in human reasoning, goal-directed behavior, and cumulative learning. Furthermore, LLMs lack mechanisms to perform explicit goal-driven cause-effect guided use of deduction, induction, abduction and analogy; if a context requires an unseen and unlikely output (x^*) not supported in the training-data manifold $\mathcal{M}$ (i.e. outside the convex hull of what was seen during training), the model has no basis for producing an answer corresponding to the physical world, being instead limited to interpolate on $\mathcal{M}$, from which next-token predictions are drawn via weighted sums over attention heads. Our formalism suggests that token-level statistical interpolation already suffices for the observed behavior; explicit internal reasoning modules are therefore not required to explain output. Consequently, we argue that claims attributing human-like cognition to contemporary LLMs are empirically unsupported, confusing surface fluency with cognitive processes in what essentially are two levels of the same misattribution: (i) Artificial Dualism: researchers project hidden reasoning modules into purely statistical models; (ii) Turing Trap: observers project agency from fluent dialogue alone.

1 Introduction

Algorithms for artificial neural networks (ANNs) have been proposed as early as in 1943 by McCullogh and Pitts [51], with one of the first implemented ANN systems being Minsky's SNARC (1952) [54]. These ideas have since resulted in a large number of variants, one of the newer ones being large language models (LLMs)—large-scale transformer ANNs, built from layers of multi-headed

self-attention and feed-forward sub-layers, that process text and generate seemingly coherent, contextually appropriate output in an autoregressive fashion [10]. More recent versions of the algorithm utilize the self-attention mechanism to weigh the importance of different tokens in a sentence relative to each other [93, 10]. Input text is tokenized, converted into vectors using embeddings, and processed through transformer layers that calculate attention scores to dictate focus on relevant tokens [93, 10, 26]. The model then selects the next token based on learned distributions, iteratively generating a (arbitrarily) long sequence of text [93, 10, 26]. Whether it is a single enormous model with hundreds of billions of parameters or a mixture of experts (MoE) where many expert models coexist controlled by a task manager model [40], these neural networks are capable of modelling complex linguistic abstractions, capturing patterns in syntax, semantics, pragmatics, and even elements of style and tone [10, 11, 62]. Although some believe the architecture to be definitive and capable of increasingly complex emergent properties, others believe that despite their size, these models are simply parroting training data [99, 98, 100, 5, 72, 49, 8, 23, 35, 107].

However, large-scale generative machine learning pipelines have been extremely useful in applied domains such as drug discovery, materials science, and chemistry in proposing vast libraries of candidate molecules, materials, or designs, which are then systematically filtered by external rule-based criteria or simulations to ensure viability [15, 80]. In pharmaceutical discovery, for example, deep generative models can enumerate new compounds, before pruning using medicinal chemistry rules and ontologies to eliminate implausible or problematic molecules [48, 4]. Similar hybrid strategies appear in materials science, where ANNs generate candidate crystal structures or molecules that are subsequently screened by physics-based simulations and logic constraints before experimental consideration [80, 58]. These examples of deterministic filters and knowledge-driven checks compensate for the fact that current algorithms lack functional reasoning capabilities in the classical sense.

By integrating external rule sets, ontologies, and physics-based evaluations as post-generation filters, researchers create a feedback loop that enforces domain knowledge and logical consistency on candidates to profit from the automation capabilities and enormous volumes of output data [39, 71]. In essence, these filters perform a kind of surrogate reasoning: they rigorously apply the deductive rules of chemistry and physics and the inductive generalizations gleaned by human experts, thus guiding the generative model's outputs toward feasible and meaningful solutions that adhere to real-world constraints [29, 15].

This reliance on external filters and domain-specific heuristics underscores a fundamental limitation: LLMs, while powerful at pattern classification and text generation, inherently lack mechanisms to validate truth or to perform explicit reasoning processes over causal-chains. There exist formal approaches to causal inference, such as Pearl's Do-calculus, that provide a rigorous framework for representing cause–effect relations using structural causal models and directed acyclic graphs, enabling estimation of intervention effects and counterfactuals beyond mere correlation [67]. Similarly, in Reinforcement Learning (RL), world-

model-based planning methods learn latent dynamics of the environment and perform virtual roll-outs to plan actions, illustrating how explicit predictive or causal models can guide decision-making [83].

Classically, RL as a field has focused on solving decision making by explicitly modeling sequential decisions involving agents that interact with their environment while learning optimal policies to achieve reward-encoded goals [83]. In modernity, the field has incorporated deep learning in what is now called Deep Reinforcement Learning (DRL) [12]. Numerous DRL models have emerged to tackle increasingly complex tasks that require not just planning, but learning of deeply complex strategies in multi-agent scenarios: Deep Q-Networks (DQNs), a classic approach of DRL, has achieved human-level control on Atari games [55], curiosity-driven intrinsic motivation modules have fostered exploration in sparse-reward environments [65], league-based multi-agent training has produced Grandmaster-level play in StarCraft II with AlphaStar [94], self-play without domain priors has taken AlphaZero to superhuman performance in chess, shogi, and Go [79], unified learning and planning resulted in MuZero's mastery across Atari, board, and planning benchmarks [74], and finally, similarly to AlphaStar, large-scale self-play systems like OpenAI Five have dominated against the best human teams in the world in real-time play in an incredibly complex and high dimensional state-space game called Dota 2 [60].

Despite their impressive successes, these algorithms do not generalize to out-of-distribution (OOD) data. To do so, they must be retrained on the game they are expected to play [12]. LLMs have also been benchmarked in non-linguistic tasks following the OOD tradition—Liga and Pasetto used Tic-Tac-Toe in ASCII form, pitting LLMs against the minimax algorithm to explore emergent features, previously suggested to resemble consciousness, but LLMs were much more likely to achieve draws or to lose than to win [47]. Topsakal and Harper [91] found GPT-4 to win more often than GPT-3.5 at Tic-Tac-Toe, but still neither model played optimally. Carvalho and Pollice extended these findings without finetuning in a zero-shot setting through their ChildPlay benchmark, which includes three simple classic board games and three novel games encoded in previously unseen ASCII formats. They observed that win rate did not necessarily improve with more recent models even for simple games, and overall performance, evaluated by optimal play criteria such as avoiding illegal moves, blocking opponents' winning moves, or executing winning moves, remained mediocre [13]. Most recently, Apple's study *The Illusion of Thinking* corroborates these findings, showing model collapse once puzzle complexity crosses a modest threshold; reasoning effort *declines* precisely when it is needed most. They believe to have shown the inadequacy of today's static test sets: they reward surface heuristics and data leakage, not causal inference [77]. Other dynamic OOD evaluation frameworks exist—such as ThinkBench [37], DeepSeek-R1 [21], and the Information Bottleneck LM objective [106]—and further reveal that LLMs remain brittle under novel distributional shifts, often requiring explicit retraining or architectural modifications to maintain performance.

Recent AGI-oriented surveys reach similar high-level conclusions. Goertzel et al. argues that today's LLMs lack the architectural components needed for grounded problem-solving [27]; Bennett highlights that syntax-only training falls short of genuine meaning representation [6]. Schneider and Bołtuć warn that such systems may look "natural-like" yet remain alien in motivation [73]. In practice, AI researchers and users often fall prey to anthropomorphism: one recent survey found that nearly half of LLM-focused articles use anthropomorphic language [5]. In present work, we show these intuitions to be misguided. By treating token prediction as probabilistic inference over a training manifold $\mathcal{M}$, we try to demonstrate mathematically why any claim of emergent cognition is unfounded and further analyze three attribution fallacies—the Black Box fallacy, the Artificial Dualism Problem (ADP), and the Turing Trap. With these, we attempt to show why people seem so quick to attribute higher-order cognitive characteristics to these algorithms.

2 Key Proposition

Human reasoning, in its various forms, underpins our ability to understand the world, solve novel problems, and generate new knowledge [88]. Key modes of reasoning include:

- **Deduction**: Inferring specific conclusions that are logically guaranteed if the premises are true (e.g., from "All A are B" and "x is A", deduce "x is B"). This often involves the application of established rules of inference to given information.
- **Induction**: Generalizing from specific observations to broader hypotheses or rules whose truth is probable but not guaranteed (e.g., observing many white swans and inferring "All swans are white"). This is fundamental to learning from experience and forming new concepts.
- **Abduction** ("inference to the best explanation"): Formulating the most plausible hypothesis to explain a given set of observations (e.g., observing wet ground, and given the knowledge that rain makes the ground wet, abducing that it likely rained). This involves generating and evaluating potential explanations.
- **Analogy**: Goal-directed systematic comparison where two or more things are compared, to highlight or uncover attributes of interest; useful for comparing that which is known, and can help a learning agent deal with unfamiliar tasks and environments.

The central question we are concerned with regarding LLMs is whether sophisticated linguistic outputs are evidence of their capacity to explicitly apply such logical processes systematically or if such cases are merely a shadow of the textual patterns of reasoning already found in their training data.

Valid human reasoning in the physical world[3] has, as its primary requirement, *knowledge of what is and is not possible*. In any particular given situation, what

[3] By 'valid reasoning' we mean reasoning whose outcome can, at least in theory, be verified by observation or experiment.

is possible, and not possible, is in turn based on how the brain models how the physical world works. Since the set of possible real-world situations is often untractable, pragmatic considerations prevent such knowledge to be directly stored and indexed, and thus the necessary knowledge for reasoning in any situation must be produced by applying the above methods to derive usable knowledge. The most efficient representation of physical events is as cause-effect relations [33]; these enable not only the prediction of what may happen but also the production of plans for making things happen [86, 87].

In humans, language emerges as a by-product of reasoning over causal diagrams and learned world models rather than as the primary driver of thought [81]. It provides the representational medium and cognitive scaffold for reasoning: inner speech regulates and manipulates thoughts via left inferior frontal gyrus circuits [95, 56]. Human reasoning combines fast, intuitive judgments (System 1) driven by unconscious heuristics and pattern recognition, and slow, deliberate analysis (System 2) that consciously manipulates mental representations via working memory and executive control [41, 25]. It spans a diversity of inference types, each recruiting overlapping but task-specific neural circuits, notably the fronto-parietal network [70]. Metacognitive oversight, implemented by prefrontal inhibitory control mechanisms, monitors, justifies, and sometimes inhibits automatic responses to maintain logical coherence [9]. Together, these processes let us draw new conclusions, form general rules, generate explanations, and map structures across domains. In contrast, LLMs produce output directly from language. Statistical patterns are encoded during training as high-dimensional probability distributions over token sequences, mapping input tokens to output tokens. No reasoning is happening from first principles. This is not to say that if enough strings of text representing the use of an inference rule have been seen during training, some of these probability distributions will not have encoded it - but we are missing important mechanisms that would enable any reliable generalization to OOD data. Overfitting the largest possible model to all available data (in essence, the standard practice in developing commercial LLMs) does not magically solve this issue. The fact is that architecturally, transformer LLMs have the causal arrow reversed: textual correlations drive token prediction, and any appearance of reasoning is then a by-product of statistical pattern completion. This makes LLMs bad, if not invalid, reasoners.

An LLM defines a conditional probability distribution $p_\theta(x_{t+1} \mid c) = jat\text{softmax}(s_\theta(c, x_{t+1}))$, where $s_\theta(c, x)$ is the model's score for token x following context $c = (x_1, \ldots, x_t)$. The parameters θ are optimized on a vast training dataset $\mathcal{M}_{data} = \{(c', x') : p_{\text{train}}(x' \mid c') > 0\}$, essentially learning to predict probable continuations based on statistical co-occurrences.

The transformer architecture, with its layers of multi-head self-attention and feed-forward networks (FFNs), computes these scores. While attention heads $o_h^{(l)}(c)$ produce convex combinations of value vectors $\{W_h^{V,(l)} z_j^{(l-1)}\}$, subsequent operations (linear projections $W^{O,(l)}$, FFNs, residual connections, and layer normalizations) transform these representations through a complex, non-linear function $F_\theta : c \mapsto z_{t+1} = z_t^{(L)}$. This final hidden state z_{t+1} determines $s_\theta(c, x)$.

Critically, F_θ is trained to map input text patterns to output text patterns. It is, in essence, an extremely high-dimensional conditional probability table refined by billions of parameters. There is no explicit mechanism or module within this architecture designed to implement logical rules for deduction, formulate and test general hypotheses for induction, or generate and evaluate causal explanations for abduction. Instead, any semblance of such reasoning in the output is a reflection of patterns absorbed from $\mathcal{M}_{data}$, where text generated by humans using these reasoning processes, be it correctly or spuriously, is abundant.

Let $\mathcal{Z}^{(L)}_{train} = \{z^{(L)}_t(c') : (c', \cdot) \text{ is consistent with } \mathcal{M}_{data}\}$ be the set of all final hidden states generated by F_θ from training-representative contexts. We posit the model primarily interpolates the convex hull of these observed training states, meaning for any context c, $z_{t+1}(c)$ effectively lies within $\text{Conv}(\mathcal{Z}^{(L)}_{train})$. This operational constraint, shaped entirely by $\mathcal{M}_{data}$, limits the model's ability to reliably reach for OOD data. A fitted convex-manifold is then insufficient for reasoning beyond "blind" extrapolation—granted, aided by the extremely large learned distributions. Asking such an ANN to solve a novel problem or generate novel reasoning forces it into an undefined space, making the output unlikely to be relevant or meaningful if not deployed at massive scales (i.e., outputting millions of candidate completions).

A. *Learned Function and Operational Space:* The transformer $F_\theta : c \mapsto z_{t+1}$ has its parameters θ optimized such that for contexts c' representative of $\mathcal{M}_{data}$, $F_\theta(c')$ yields $z^{(L)}_t(c') \in \mathcal{Z}^{(L)}_{train}$ which, via the softmax layer, correctly predicts x' from $\mathcal{M}_{data}$. As a function learned from examples, F_θ typically acts as an interpolator. Thus, for any input c, the resulting $z_{t+1}(c)$ is expected to be a point within (or near) $\text{Conv}(\mathcal{Z}^{(L)}_{train})$. This acknowledges that while F_θ is complex and involves operations like FFNs and residual connections that break simple convexity propagation from initial value vectors, its final output states are constrained by the manifold of such states seen during training.

B. *OOD Reasoning vs. Interpolation:* Consider a task requiring genuine OOD reasoning (deductive, inductive, abductive, or by analogy) to arrive at a conclusion x^* from context c^*, e.g. to solve the millenium problem of P vs NP. Such reasoning implies understanding the problem, constructing a novel understanding, and applying a rule in a new way, which would correspond to an ideal final hidden state z^*_{ideal}. If $z^*_{ideal} \notin \text{Conv}(\mathcal{Z}^{(L)}_{train})$, the model, being confined to this interpolative space, is unlikely to produce z^*_{ideal} without interaction or inductive bias [6]. Instead, it generates $z_{t+1}(c^*) \in \text{Conv}(\mathcal{Z}^{(L)}_{train})$. This $z_{t+1}(c^*)$ will reflect familiar patterns from $\mathcal{M}_{data}$ rather than the specific novel inference required for x^*. Consequently, $p_\theta(x^* \mid c^*)$ will be low.

C. *Implications for Emergence in Long Sequences:* Autoregressive generation of a sequence $X = (x_1, \ldots, x_K)$ involves a trajectory of states $z_{t+k+1}(c_k)$, where each $z_{t+k+1}(c_k) \in \text{Conv}(\mathcal{Z}^{(L)}_{train})$. The claim that robust, OOD reasoning might

"emerge" over long sequences implies that this trajectory could spontaneously implement a globally coherent novel logical argument. However, if each step is limited to $\text{Conv}(\mathcal{Z}_{train}^{(L)})$ and selected based on learned statistical likelihoods rather than logical validity or explanatory power for novel situations, the sequence remains an elaborate form of pattern completion bound to $\mathcal{M}_{data}$. The model lacks the internal mechanisms to discover or apply novel abstract rules of deduction, induction, abduction, or analogy in a thoughtful manner, and to generate goals explicitly—properties that would allow it to navigate to a $z_{ideal}^{*} \notin \text{Conv}(\mathcal{Z}_{train}^{(L)})$ in a principled way. Thus, any "emergent" properties must be explained as recombinations of learned patterns as lucky "shoots in the dark" rather than genuine OOD reasoning. ■

Black Boxes are Not Pitch-Black: A rapidly growing body of *mechanistic interpretability* work demonstrates how individual attention heads implement induction, copy-and-paste, or simple arithmetic [57, 59, 28, 105, 36]. To the author's knowledge, no study has revealed circuitry for goal formation or complex causal simulation.

Missing biological counterparts: The brain is our only ground truth for what we know to be possible in terms of cognition, and thus should not be ignored. Regions such as the prefrontal cortex (executive control) [53, 43], thalamus (multimodal integration) [97, 31, 32] and hippocampus (memory consolidation) [82, 75] form dense, recurrent, neuromodulated loops long suspected to underpin consciousness and abstract reasoning [90, 17, 22]. A transformer stack, by contrast, is a strictly feed-forward computation [93, 24]. It performs conditional probability lookup, not the continuous iterative, self-modifying processes required for cumulative learning [89]. A normal adult cortex contains $\sim 8.6 \times 10^9$ neurons [3] and $\sim 3 \times 10^{14}$ synapses [84, 63]. Chemical signalling exploits dozens of transmitters, yielding rich temporal codes and plasticity cascades [42], while ANNs have only abstracted firing rates of neurons or action potentials in the case of spiking neural networks [66]. GPT-3.5 stores 1.75×10^{11} static weights [10]; all adaptation ends once gradient descent stops [30]. Even speech-critical Broca's area contains $\mathcal{O}(10^8)$ neurons in recurrent microcolumns [1, 78], whereas a single 96-head attention block uses only $\sim 10^4$ learned parameters and no internal state [93]. Detailed reconstructions of cortical microcircuits show dendritic nonlinearities and state-dependent reconfiguration far beyond present transformers [50]. Mere parameter count, then, is a poor proxy for the qualitative machinery that supports genuine brain function, and subsequently, reasoning.

Scale & compute are not substitutes for understanding: Wei *et al.* and Yao *et al.* have shown that LLMs can be coaxed into levels of abstraction without external methods through techniques such as chain-of-thought and tree-of-thought prompting (i.e. having an LLM prompting itself or branching off into multiple scenarios and then picking the most likely one) [101, 104], but this abstraction may be illusory, because the underlying process is still next-token prediction by the same model.

Regardless, reasoning-specific scaling in LLMs exhibits severe plateaus. Chen et al.'s survey finds that simply increasing context length, chain-of-thought depth, or the number of collaborating agents often yields no improvement, showing instead degraded performance once a critical threshold is passed, due to redundancy and error compounding [14]. Wang et al. formalize this test-time scaling plateau with their TTSPM model, deriving saturation budgets beyond which additional candidate generations or reasoning rounds afford negligible gains, and empirically validate these bounds on AIME, MATH-500, and GPQA benchmarks [96].

Shojaee et al. conclude in a similar vein that frontier reasoning models undergo a complete accuracy collapse beyond moderate task complexity and that their reasoning effort paradoxically declines even when token budgets remain adequate—evidence that scale alone cannot sustain structured, robust inference [77].

In practice, these limits impose steep costs for marginal benefits. This seems to indicate that textual data alone cannot provide the inductive biases required for general, correctly applied causal abstraction, hierarchical memory, or explicit goal-directed planning. We believe that breaking through these plateaus will require new architectural primitives, and for that we must not ignore the brain and what it has to teach us about information processing, namely that language and the patterns of thought found therein are an outcome of structured, embodied neural computations—dynamic causal inference, hierarchical working memory, and goal-directed control — language is **not** the source of reasoning, but the outcome of underlying brain mechanics.

The Artificial Dualism Problem: We believe that interpreting LLMs' outputs as an expression of reasoning, rather than as the output of an arbitrary probability function, is a mistake. We call this topic the artificial dualism problem (ADP): when experts reify latent vectors as if they were explicit rules, goals, or beliefs. Unlike computational dualism in embedded-agency work—which studies how a policy is embedded in, or separated from, its physical substrate [61, 45, 7]—ADP is purely observer-side: it is a misattribution error. Nor is ADP related to classical mind–body dualism; we make no claim about non-reductive physicalism or immaterial minds. ADP is an ontological category mistake: it projects mechanisms capable of rule learning, goal creation, or beliefs into the model's latent vectors. By contrast, the Turing Trap is an evidential inference error: it projects those same mental states from surface behaviour. Our approach addresses the bulk of statements and propositions that attribute higher-order cognitive functions to LLMs.

When the internal mechanics of a generative model are opaque to a user, the simplest folk-psychology move is to insert an imagined reasoner behind its output. The move is bolstered by surface features—grammar, coherence, apparent insight—that humans evolved to interpret as markers of agency. This is a fallacy in that mechanisms for which there is no evidence, apart from the fluency of text, are necessarily posited. The Blake Lemoine/LaMDA episode [46, 85] is an example of this: the engineer filled explanatory gaps with talk of sentience de-

spite a complete lack of supporting evidence. Mechanistic-interpretability studies repeatedly reveal specialized pattern-matching circuits, not world-model-driven reasoning. ADP thus resembles a "God-of-the-gaps" fallacy: explanatory voids are patched with an unwarranted cognitive capacity. Crucially, ADP is falsifiable through the research efforts of mechanistic-interpretability—through experiment, the effects of different circuity and nodes may be understood.

The Turing Trap: In 1950 Alan Turing [92] proposed a working definition of intelligence that he called the "imitation game", wherein a machine would chat with a human judge through a text terminal; if the machine could converse in a manner indistinguishable from a human, it should be considered intelligent. Later, this idea, which Turing originally proposed as a temporary stop-gap for a proper definition of intelligence, was turned on its head and made into a goal that the field of AI should strive for. Labeled the 'Turing Test', this idea has been thoroughly criticized for failing to account for the underlying mechanisms that produce such responses both in machines and in humans [76, 34]. We use the name 'Turing trap' to describe this fallacy of anthropomorphizing cognitive capacities in ANNs simply because they can generate human-like responses to "pass the Turing Test". The Turing Trap then, we argue, occurs when observers mistake surface-level performance of ANNs for genuine cognitive capacities. This mistake seems to us to be driven by anthropomorphism, the human tendency to attribute human-like qualities to non-human entities [2]. We think that one of the primary factors is the high level of fluency and apparent coherence in the text generated by LLMs which can create a powerful illusion of depth and intentionality.

3 Discussion & Conclusion

The central thesis of this paper challenges the narrative endorsed by figures such as Nobel Laureate Geoffrey Hinton [44], namely that Large Language Models (LLMs), and more generally ANN-based architectures, exhibit emergent reasoning or understanding in a cognitive sense. While LLMs often produce coherent chains of text, these arise from large-scale interpolation over familiar data rather than genuine reasoning or goal pursuit; without integration of explicit causal or world-model components, 'reasoning' remains effectively in-domain and falters under rigorous OOD evaluations [13, 77, 47, 91]. Other AGI proponents have echoed this theory, like Goertzel et al. who argued that today's LLMs "lack the basic cognitive architectures" needed for genuine problem–solving and therefore should not be viewed as incremental steps toward human-level AGI [27]. Bennett et al. likewise believes that, absent grounded interaction, a language model's facility with syntax is insufficient for the computation of meaning [6].

The reader might be interested to produce the following experiment by themselves which highlights our main thesis: Try to convince an LLM to solve an unsolved scientific problem, such as a Millennium problem, like P vs NP. This is the question of whether every problem for which a solution can be verified quickly

(in polynomial time, denoted as P) can also be solved quickly [16]. If we begin by asking an LLM to list known strides and advances in the theory and then work from those, we will see that the model quickly hits a ceiling. We argue here that that's because the manifold space that concerns information related to the P vs NP problem "ran out", leaving the model with a brute-forced extrapolation, without the ability to generate well-justified or truly novel insights, given the lack of mechanisms to effectively make use of well-supported cause-effect relations. One can picture producing a trillion trees of hundreds of chains-of-thought each and then checking if any solves the problem, but this is not what anyone means by reasoning. This limitation shows the fundamental difference between data-driven models and human creativity and intelligent logical reasoning, which can, even if slowly and constrained by limited experience, enable exploration of uncharted territory regardless of existing data. Neuroscience is yet to discover all the ingredients for building general intelligence, but it does offer many detailed, mechanistic accounts of brain circuits often overlooked in AI (see Damasio et al. [19, 18, 64, 20]).

The inability of LLMs to internally validate truths and reason might stem not only from data constraints but also from fundamental theoretical limits inherent in their design. Many of these limitations reflect known epistemological and computational boundaries of formal systems. Russell's theory of types, which stratified language to avoid self-referential paradoxes, emphasized that certain truths necessitate stepping outside a given system for proper resolution [102]. Wittgenstein's picture theory posited that while language can represent facts through a shared logical form, it cannot explicitly articulate or verify its own logical foundations — such foundations can only be "shown", not stated [103]. Roger Penrose has expanded these ideas, suggesting that aspects of human insight and understanding might be a prerequisite for meaning, regardless of algorithmic capabilities [68, 69]. Although we do not necessarily agree with Penrose's computational incompleteness, the parallel is notable: without a human observer, the symbols manipulated by LLMs remain inert and without meaning.

In conclusion, we believe the field should aim to end where we begin—by re-grounding AI in the human mind, our only true model of real-time, embodied reasoning. Studying neural circuits for working memory, hierarchical control, embodied interaction, and neuromodulation should be inspiring architectures that sustain goals, continuously model the world, and reason within the flow of time—capabilities that current transformers fundamentally lack but other technologies are pursuing, such as neuromorphic computing [52, 38]. With these functional gaps, claims of LLM sentience are mistaking scale for substrate. A larger model is still executing the same next-token training objective. Combined with the stochastic-parrot insight that LLMs do not understand the meaning of the language they process [5], our analysis underscores that any appearance of reasoning is a statistical mirage. LLMs are a sophisticated and dynamic mirror of human knowledge, reflecting, not transcending, the ingested data. This is not to say that LLMs are incapable of interpolating truly novel data—the quality and meaning of this data is what leaves a lot to be desired.

Disclosure of Interests. The authors have no competing interests.

References

1. Amunts, K., Schleicher, A., Bürgel, U., Mohlberg, H., Uylings, H.B., Zilles, K.: Broca's region revisited: cytoarchitecture and intersubject variability. Journal of Comparative Neurology **412**(2), 319–341 (9 1999). https://doi.org/10.1002/(SICI)1096-9861(19990920)412:2<319::aid-cne10>3.0.CO;2-7
2. Arleen Salles, K.E., Farisco, M.: Anthropomorphism in ai. AJOB Neuroscience **11**(2), 88–95 (2020). https://doi.org/10.1080/21507740.2020.1740350, pMID: 32228388
3. Azevedo, F.A.C., Carvalho, L.R.B., Grinberg, L.T., Farfel, J.M., Ferretti, R.E.L., Leite, R.E.P., Jacob Filho, W., Lent, R., Herculano-Houzel, S.: Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. Journal of Comparative Neurology **513**(5), 532–541 (4 2009). https://doi.org/10.1002/cne.21974
4. Baell, J.B., Holloway, G.A.: New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. J. Med. Chem. **53**(7), 2719–2740 (2010). https://doi.org/10.1021/jm901137j
5. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. p. 610–623. FAccT '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442188.3445922, `https://doi.org/10.1145/3442188.3445922`
6. Bennett, M.T.: On the Computation of Meaning, Language Models and Incomprehensible Horrors, p. 32–41. Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-33469-6_4, `http://dx.doi.org/10.1007/978-3-031-33469-6\_4`
7. Bennett, M.T.: Computational dualism and objective superintelligence. In: Thórisson, K.R., Isaev, P., Sheikhlar, A. (eds.) Artificial General Intelligence. pp. 22–32. Springer Nature Switzerland, Cham (2024)
8. Borji, A.: Stochastic parrots or intelligent systems? a perspective on true depth of understanding in llms. SSRN Electronic Journal (01 2023). https://doi.org/10.2139/ssrn.4507038
9. Borst, G., Houdé, O.: Training in logic inhibits selection of perceptual principles: A study with the wason selection task. Developmental Science **17**(5), 741–749 (2014)
10. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020), `https://proceedings.neurips.cc/paper\_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`

11. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J.A., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y.F., Lundberg, S.M., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y.: Sparks of artificial general intelligence: Early experiments with GPT-4. ArXiv **abs/2303.12712** (2023), https://api.semanticscholar.org/CorpusID:257663729
12. de Carvalho, G.H., Vos, T.: Game-solving drl: An introductory literature survey of the last decade and a critical methodological review. osf.io preprint (8 2024). https://doi.org/10.31219/osf.io/7zmx2_v1, https://doi.org/10.31219/osf.io/7zmx2_v1
13. de Carvalho, G.H., Knap, O., Pollice, R.: Show, don't tell: Evaluating large language models beyond textual understanding with ChildPlay (2024), https://arxiv.org/abs/2407.11068
14. Chen, Z., Wang, S., Tan, Z., Fu, X., Lei, Z., Wang, P., Liu, H., Shen, C., Li, J.: A survey of scaling in large language model reasoning (2025), https://arxiv.org/abs/2504.02181
15. Chenthamarakshan, V., Hoffman, S.C., Owen, C.D., Lukacik, P., Strain-Damerell, C., Fearon, D., Malla, T.R., Tumber, A., Schofield, C.J., Duyvesteyn, H.M.E., Dejnirattisai, W., Carrique, L., Walter, T.S., Screaton, G.R., Matviiuk, T., Mojsilovic, A., Crain, J., Walsh, M.A., Stuart, D.I., Das, P.: Accelerating drug target inhibitor discovery with a deep generative foundation model. Sci. Adv. **9**(25), eadg7865 (2023). https://doi.org/10.1126/sciadv.adg7865
16. Clay Mathematics Institute: P vs NP (2024), https://www.claymath.org/wp-content/uploads/2022/06/pvsnp.pdf, accessed: 2024-08-19
17. Craik, K.: The Nature of Explanation. Cambridge University Press (1943), https://books.google.ch/books?id=EN0TrgEACAAJ
18. Damasio, A.R.: Investigating the biology of consciousness. Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences **353**(1377), 1879–1882 (11 1998). https://doi.org/10.1098/rstb.1998.0339
19. Damasio, A.R.: How the brain creates the mind. Scientific American **281**(6), 112–117 (12 1999). https://doi.org/10.1038/scientificamerican1299-112
20. Damasio, A.R., Grabowski, T.J., Bechara, A., Damasio, H., Ponto, L.L.B., Parvizi, J., Hichwa, R.D.: Subcortical and cortical brain activity during the feeling of self-generated emotions. Nature Neuroscience **3**(10), 1049–1056 (10 2000). https://doi.org/10.1038/79871
21. DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J.L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R.J., Jin, R.L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S.S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W.L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X.Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang,

X., Shan, X., Li, Y.K., Wang, Y.Q., Wei, Y.X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y.X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., Zhang, Z.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025), https://arxiv.org/abs/2501.12948

22. Dehaene, S., Naccache, L.: Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. Cognition **79**(1), 1–37 (2001). https://doi.org/https://doi.org/10.1016/S0010-0277(00)00123-2, the Cognitive Neuroscience of Consciousness
23. Duan, H., Dziedzic, A., Papernot, N., Boenisch, F.: Flocks of stochastic parrots: Differentially private prompt learning for large language models. ArXiv **abs/2305.15594** (2023), https://api.semanticscholar.org/CorpusID:258887717
24. Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., Olah, C.: A mathematical framework for transformer circuits. Transformer Circuits Thread (2021), https://transformer-circuits.pub/2021/framework/index.html
25. Evans, J.S.B.T., Stanovich, K.E.: Dual-process theories of higher cognition: Advancing the debate. Perspectives on Psychological Science **8**(3), 223–241 (2013)
26. Fields, J., Chovanec, K., Madiraju, P.: A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? IEEE Access **12**, 6518–6531 (2024), https://api.semanticscholar.org/CorpusID:266824505
27. Goertzel, B.: Generative ai vs. agi: The cognitive strengths and weaknesses of modern llms (2023), https://arxiv.org/abs/2309.10371
28. Golgoon, A., Filom, K., Kannan, A.R.: Mechanistic interpretability of large language models with applications to the financial services industry (2024), https://arxiv.org/abs/2407.11215
29. Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P., Aspuru-Guzik, A.: Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent. Sci. **4**(2), 268–276 (2018). https://doi.org/10.1021/acscentsci.7b00572
30. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), http://www.deeplearningbook.org
31. Guillery, R.W., Sherman, S.M.: The thalamus as a monitor of motor outputs. Philosophical Transactions of the Royal Society B: Biological Sciences **357**(1428), 1809–1821 (Dec 2002). https://doi.org/10.1098/rstb.2002.1171, https://doi.org/10.1098/rstb.2002.1171
32. Halassa, M.M., Kastner, S.: Thalamic functions in distributed cognitive control. Nature Neuroscience **20**(12), 1669–1679 (dec 2017). https://doi.org/10.1038/s41593-017-0020-1, https://doi.org/10.1038/s41593-017-0020-1

33. Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach — part 1: Causes. CoRR **abs/1301.2275** (2013), http://arxiv.org/abs/1301.2275
34. Harnad, S.: The turing test is not a trick: Turing indistinguishability is a scientific criterion. SIGART Bulletin **3**(4), 9–10 (10 1992). https://doi.org/10.1145/141420.141422, https://doi.org/10.1145/141420.141422
35. Henrique, D.S.G., Kucharavy, A., Guerraoui, R.: Stochastic parrots looking for stochastic parrots: Llms are easy to fine-tune and hard to detect with other llms (2023), https://arxiv.org/abs/2304.08968
36. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015), https://arxiv.org/abs/1503.02531
37. Huang, S., Yang, L., Song, Y., Chen, S., Cui, L., Wan, Z., Zeng, Q., Wen, Y., Shao, K., Zhang, W., Wang, J., Zhang, Y.: Thinkbench: Dynamic out-of-distribution evaluation for robust llm reasoning (2025), https://arxiv.org/abs/2502.16268
38. Indiveri, G., Liu, S.C.: Memory and information processing in neuromorphic systems. Proceedings of the IEEE **103**(8), 1379–1397 (2015). https://doi.org/10.1109/JPROC.2015.2444094, https://doi.org/10.1109/JPROC.2015.2444094
39. Ivanenkov, Y.A., Polykovskiy, D., Bezrukov, D.S., Zagribelnyy, B.A., Aladinskiy, V.A., Kamya, P.O., Aliper, A., Ren, F., Zhavoronkov, A.: Chemistry42: An ai-driven platform for molecular design and optimization. J. Chem. Inf. Model. **63**(3), 695–701 (2023). https://doi.org/10.1021/acs.jcim.2c01191
40. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Computation **3**(1), 79–87 (1991). https://doi.org/10.1162/neco.1991.3.1.79
41. Kahneman, D.: Thinking, Fast and Slow. Farrar, Straus and Giroux, New York, NY (2011)
42. Kandel, E.R., Schwartz, J.H., Jessell, T.M., Siegelbaum, S.A., Hudspeth, A.J., Mack, S. (eds.): Principles of Neural Science. McGraw-Hill Education / Medical, New York, 5 edn. (2013), https://accessbiomedicalscience.mhmedical.com/content.aspx?bookid=1049\&sectionid=59138139
43. Kouneiher, F., Charron, S., Koechlin, E.: Motivation and cognitive control in the human prefrontal cortex. Nature neuroscience **12**, 939–45 (08 2009). https://doi.org/10.1038/nn.2321
44. Kruppa, M., Seetharaman, D.: A godfather of ai just won a nobel. he has been warning the machines could take over the world. The Wall Street Journal (2024), https://www.wsj.com/tech/ai/a-godfather-of-ai-just-won-a-nobel-he-has-been-warning-the-machines-could-take-over-the-world-b127da71, accessed: April 09, 2025
45. Leike, J., Hutter, M.: Bad universal priors and notions of optimality. In: Grünwald, P., Hazan, E., Kale, S. (eds.) Proceedings of The 28th Conference on Learning Theory. Proceedings of Machine Learning Research, vol. 40, pp. 1244–1259. PMLR, Paris, France (7 2015), https://proceedings.mlr.press/v40/Leike15.html
46. Lemoine, B.: Is LaMDA sentient? Letter (08 2024), https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917, accessed: 2024-08-19
47. Liga, D., Pasetto, L.: Testing spatial reasoning of large language models: the case of tic-tac-toe (2023), https://ceur-ws.org/Vol-3563/paper_14.pdf

48. Lipinski, C.A.: Lead- and drug-like compounds: the rule-of-five revolution. Drug Discov. Today Technol. **1**(4), 337–341 (2004). https://doi.org/10.1016/j.ddtec.2004.11.007
49. Lu, S., Bigoulaeva, I., Sachdeva, R., Tayyar Madabushi, H., Gurevych, I.: Are emergent abilities in large language models just in-context learning? Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Long Papers) pp. 5098–5139 (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.279, https://aclanthology.org/2024.acl-long.279/
50. Markram, H., Muller, E., Ramaswamy, S., Reimann, M.W., Abdellah, M., Sanchez, C.A., Ailamaki, A., Alonso-Nanclares, L., Antille, N., Arsever, S., Kahou, G.A., Berger, T.K., Bilgili, A., Buncic, N., Chalimourda, A., Chindemi, G., Courcol, J., Delalondre, F., Delattre, V., Druckmann, S., Dumusc, R., Dynes, J., Eilemann, S., Gal, E., Gevaert, M.E., Ghobril, J., Gidon, A., Graham, J.W., Gupta, A., Haenel, V., Hay, E., Heinis, T., Hernando, J.B., Hines, M., Kanari, L., Keller, D., Kenyon, J., Khazen, G., Kim, Y., King, J.G., Kisvarday, Z., Kumbhar, P., Lasserre, S., Le Bé, J., Magalhães, B.R.C., Merchán-Pérez, A., Meystre, J., Morrice, B.R., Muller, J., Muñoz-Céspedes, A., Muralidhar, S., Muthurasa, K., Nachbaur, D., Newton, T.H., Nolte, M., Ovcharenko, A., Palacios, J., Pastor, L., Perin, R., Ranjan, R., Riachi, I., Rodríguez, J., Riquelme, J.L., Rössert, C., Sfyrakis, K., Shi, Y., Shillcock, J.C., Silberberg, G., Silva, R., Tauheed, F., Telefont, M., Toledo-Rodriguez, M., Tränkler, T., Van Geit, W., Díaz, J.V., Walker, R., Wang, Y., Zaninetta, S.M., DeFelipe, J., Hill, S.L., Segev, I., Schürmann, F.: Reconstruction and simulation of neocortical microcircuitry. Cell **163**(2), 456–492 (10 2015). https://doi.org/10.1016/j.cell.2015.09.029, https://doi.org/10.1016/j.cell.2015.09.029
51. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics **5**(4), 115–133 (12 1943). https://doi.org/10.1007/BF02478259, https://doi.org/10.1007/BF02478259
52. Mead, C.: Neuromorphic electronic systems. Proceedings of the IEEE **78**(10), 1629–1636 (1990). https://doi.org/10.1109/5.58356, https://doi.org/10.1109/5.58356
53. Miller, E.K., Cohen, J.D.: An integrative theory of prefrontal cortex function. Annual Review of Neuroscience **24**(Volume 24, 2001), 167–202 (2001). https://doi.org/https://doi.org/10.1146/annurev.neuro.24.1.167, https://www.annualreviews.org/content/journals/10.1146/annurev.neuro.24.1.167
54. Minsky, M.: A neural-analogue calculator based upon a probability model of reinforcement. Tech. rep., Harvard University Psychological Laboratories, Cambridge, MA (1 1952), https://www.bibsonomy.org/bibtex/2d2b3af4935de200ea20c5c191c8c4d67/machinelearning
55. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M.A., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015)
56. Monti, M.M., Osherson, D.N., Martinez, M.J., Parsons, L.M.: The boundaries of language and thought in deductive inference. Proceedings of the National Academy of Sciences **106**(30), 12554–12559 (2009)
57. Nanda, N.: A comprehensive mechanistic interpretability explainer & glossary (Dec 2022), https://neelnanda.io/glossary

58. Nguyen, B., Schulz, H., Lewis, S., Huang, C.W., Lu, Z., Zhou, Y., Yang, H., Hao, H., Li, J., Yang, C., Li, W., Tomioka, R., Xie, T.: A generative model for inorganic materials design. Nature **639**, 624–632 (2025). https://doi.org/10.1038/s41586-025-08628-5
59. Olah, C., Nanda, N.: A framework for understanding neural network models. `https://transformer-circuits.pub/2021/framework/index.html` (2021), accessed: 2024-08-19
60. OpenAI, :, Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., Józefowicz, R., Gray, S., Olsson, C., Pachocki, J., Petrov, M., d. O. Pinto, H.P., Raiman, J., Salimans, T., Schlatter, J., Schneider, J., Sidor, S., Sutskever, I., Tang, J., Wolski, F., Zhang, S.: Dota 2 with large scale deep reinforcement learning. Unknown p. Unknown (2019)
61. Orseau, L., Ring, M.B.: Space–time embedded intelligence. In: Bach, J., Goertzel, B., Iklé, M. (eds.) Artificial General Intelligence. 5th International Conference, AGI 2012, Oxford, UK, December 8–11, 2012. Proceedings, Lecture Notes in Computer Science, vol. 7716, pp. 209–218. Springer, Cham (Dec 2012). https://doi.org/10.1007/978-3-642-35506-6_22, `https://doi.org/10.1007/978-3-642-35506-6\_22`
62. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback (2022), `https://arxiv.org/abs/2203.02155`
63. Pakkenberg, B., Gundersen, H.J.G.: Neocortical neuron number in humans: effect of sex and age. Journal of Comparative Neurology **384**(2), 312–320 (7 1997). https://doi.org/10.1002/(SICI)1096-9861(19970728)384:2<312::AID-CNE10>3.0.CO;2-K, `https://doi.org/10.1002/(SICI)1096-9861(19970728)384:2<312::AID-CNE10>3.0.CO;2-K`
64. Parvizi, J., Damasio, A.: Consciousness and the brainstem. Cognition **79**(1-2), 135–160 (4 2001). https://doi.org/10.1016/S0010-0277(00)00127-X
65. Pathak, D., Agrawal, P., Efros, A.A., Darrell, T.: Curiosity-driven exploration by self-supervised prediction. Unknown p. Unknown (2017)
66. Paugam-Moisy, H., Bohte, S.: Computing with spiking neuron networks. Handbook of Natural Computing pp. 335–376 (2012). https://doi.org/10.1007/978-3-540-92910-9_10, `https://doi.org/10.1007/978-3-540-92910-9\_10`
67. Pearl, J.: Causal diagrams for empirical research. Biometrika **82**(4), 669–688 (1995), `http://www.jstor.org/stable/2337329`
68. Penrose, R.: The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics. Oxford University Press, Oxford, UK (1989), first edition, hardcover
69. Penrose, R.: Shadows of the Mind: A Search for the Missing Science of Consciousness. Oxford University Press, Oxford, UK (1994), first edition, hardcover
70. Prado, J., Mutreja, R., Booth, J.R.: The brain network for deductive reasoning: A quantitative meta-analysis of 28 neuroimaging studies. Journal of Cognitive Neuroscience **23**(10), 3483–3498 (2011)
71. Putin, E., Asadulaev, A., Vanhaelen, Q., Ivanenkov, Y.A., Aladinskaya, A.V., Aliper, A., Zhavoronkov, A.: Adversarial threshold neural computer for molecular de novo design. Mol. Pharm. **15**(10), 4386–4397 (2018). https://doi.org/10.1021/acs.molpharmaceut.7b01137

72. Schaeffer, R., Miranda, B., Koyejo, S.: Are emergent abilities of large language models a mirage? In: Advances in Neural Information Processing Systems 37. pp. 209–218. Curran Associates, Inc. (2023). https://doi.org/10.48550/arXiv.2304.15004, https://doi.org/10.48550/arXiv.2304.15004
73. Schneider, H., Bołtuć, P.: Alien versus natural-like artificial general intelligences. In: Hammer, P., Alirezaie, M., Strannegård, C. (eds.) Artificial General Intelligence. pp. 233–243. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-33469-6_24
74. Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., Silver, D.: Mastering atari, go, chess and shogi by planning with a learned model. Nature **588**(7839), 604–609 (Dec 2020). https://doi.org/10.1038/s41586-020-03051-4
75. Scoville, W.B., Milner, B.: Loss of recent memory after bilateral hippocampal lesions. Journal of Neurology, Neurosurgery & Psychiatry **20**(1), 11–21 (2 1957). https://doi.org/10.1136/jnnp.20.1.11, https://doi.org/10.1136/jnnp.20.1.11
76. Searle, J.R.: Minds, brains, and programs. Behavioral and Brain Sciences **3**(3), 417–424 (1980). https://doi.org/10.1017/S0140525X00005756
77. Shojaee*†, P., Mirzadeh*, I., Alizadeh, K., Horton, M., Bengio, S., Farajtabar, M.: The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity (2025), https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf
78. Silbereis, J.C., Pochareddy, S., Zhu, Y., Li, M., Sestan, N.: The cellular and molecular landscapes of the developing human central nervous system. Neuron **89**(2), 248–268 (1 2016). https://doi.org/10.1016/j.neuron.2015.12.008, https://doi.org/10.1016/j.neuron.2015.12.008
79. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., Hassabis, D.: Mastering chess and shogi by self-play with a general reinforcement learning algorithm. Unknown p. Unknown (2017)
80. Siriwardane, E.M.D., Zhao, Y., Perera, I., Hu, J.: Generative design of stable semiconductor materials using deep learning and density functional theory. npj Comput. Mater. **8**, 164 (2022). https://doi.org/10.1038/s41524-022-00850-3
81. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. The MIT Press, Cambridge, MA, 2 edn. (2001). https://doi.org/10.7551/mitpress/1754.001.0001, https://doi.org/10.7551/mitpress/1754.001.0001
82. Squire, L.R., Zola-Morgan, S.: The neuropsychology of memory: New links between humans and experimental animals. Annals of the New York Academy of Sciences **444**, 137–149 (1985). https://doi.org/10.1111/j.1749-6632.1985.tb37585.x, https://doi.org/10.1111/j.1749-6632.1985.tb37585.x
83. Sutton, R., Barto, A.: Reinforcement learning: An introduction (1998). https://doi.org/10.1109/TNN.1998.712192
84. Tang, Y., Nyengaard, J.R., De Groot, D.M.G., Gundersen, H.J.G.: Total regional and global number of synapses in the human brain neocortex. Synapse **41**(3), 258–273 (2001). https://doi.org/10.1002/syn.1083
85. Thoppilan, R., Freitas, D.D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H.S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhou, Y., Chang, C., Krivokon, I., Rusch,

W., Pickett, M., Meier-Hellstern, K.S., Morris, M.R., Doshi, T., Santos, R.D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., y Arcas, B.A., Cui, C., Croak, M., Chi, E.H., Le, Q.: Lamda: Language models for dialog applications. CoRR **abs/2201.08239** (2022), https://arxiv.org/abs/2201.08239

86. Thórisson, K.R.: Seed-programmed autonomous general learning. Proceedings of Machine Learning Research **131**, 32–70 (2020)
87. Thórisson, K.R.: The explanation hypothesis in general self-supervised learning. In: Proceedings of Machine Learning Research, 159. pp. 5–27 (2021)
88. Thórisson, K.R., Kremelberg, D., Steunebrink, B.R., Nivel, E.: About understanding. In: Proceedings of the International Conference on Artificial General Intelligence. pp. 106–117. Springer-Verlag, New York, NY, USA (2016)
89. Thórisson, K.R., Bieger, J., Li, X., Wang, P.: Cumulative learning. In: Hammer, P., Agrawal, P., Goertzel, B., Iklé, M. (eds.) Artificial General Intelligence. AGI 2019. Lecture Notes in Computer Science, vol. 11654. pp. 198–208. Springer, Cham (7 2019). https://doi.org/10.1007/978-3-030-27005-6_20, https://doi.org/10.1007/978-3-030-27005-6_20
90. Tononi, G., Edelman, G.M.: Consciousness and complexity. Science **282**(5395), 1846–1851 (Dec 1998). https://doi.org/10.1126/science.282.5395.1846, https://doi.org/10.1126/science.282.5395.1846
91. Topsakal, O., Harper, J.: Benchmarking large language model (llm) performance for game playing via tic-tac-toe. Electronics **13**, 1532 (04 2024). https://doi.org/10.3390/electronics13081532
92. TURING, A.M.: I.—computing machinery and intelligence. Mind **LIX**(236), 433–460 (10 1950). https://doi.org/10.1093/mind/LIX.236.433, https://doi.org/10.1093/mind/LIX.236.433
93. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30, pp. 5998–6008 (2017)
94. Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J.P., Jaderberg, M., Vezhnevets, A.S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T.L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T.P., Kavukcuoglu, K., Hassabis, D., Apps, C., Silver, D.: Grandmaster level in starcraft ii using multi-agent reinforcement learning. Nature **575**, 350 – 354 (2019)
95. Vygotsky, L.S.: Thought and Language. MIT Press, Cambridge, MA (1986)
96. Wang, J., Zhu, B., Leong, C.T., Li, Y., Li, W.: Scaling over scaling: Exploring test-time scaling plateau in large reasoning models (2025), https://arxiv.org/abs/2505.20522
97. Ward, L.M.: The thalamus: Gateway to the mind. Wiley Interdisciplinary Reviews: Cognitive Science **4**(6), 609–622 (Nov 2013). https://doi.org/10.1002/wcs.1256, https://doi.org/10.1002/wcs.1256
98. Webb, T., Holyoak, K.J., Lu, H.: Emergent analogical reasoning in large language models. Nature Human Behaviour **7**(9), 1526–1541 (9 2023). https://doi.org/10.1038/s41562-023-01659-w, https://doi.org/10.1038/s41562-023-01659-w

99. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W.: Emergent abilities of large language models. Transactions on Machine Learning Research **TMLR**, 1–50 (2022). https://doi.org/10.48550/arXiv.2206.07682, https://doi.org/10.48550/arXiv.2206.07682
100. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W.: Emergent abilities of large language models. Transactions on Machine Learning Research **5**, 1–50 (8 2022). https://doi.org/10.48550/arXiv.2206.07682, https://doi.org/10.48550/arXiv.2206.07682
101. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023), https://arxiv.org/abs/2201.11903
102. Whitehead, A.N., Russell, B.: Principia Mathematica, vol. I. Cambridge University Press, Cambridge, UK (1910)
103. Wittgenstein, L.J.J.: Logisch-philosophische abhandlung. Annalen der Naturphilosophie **XIV, Hefte 3/4**, 185–262 (1921)
104. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., Narasimhan, K.: Tree of thoughts: Deliberate problem solving with large language models (2023), https://arxiv.org/abs/2305.10601
105. Yosinski, J., Clune, J., Nguyen, A.M., Fuchs, T.J., Lipson, H.: Understanding neural networks through deep visualization. CoRR **abs/1506.06579** (2015), http://arxiv.org/abs/1506.06579
106. Yu, F.: Memorization-compression cycles improve generalization (2025), https://arxiv.org/abs/2505.08727
107. Zečević, M., Willig, M., Dhami, D.S., Kersting, K.: Causal parrots: Large language models may talk causality but are not causal (2023), https://arxiv.org/abs/2308.13067