The Darwinian mind of the machine: LLM language learning as evolution

Marc van Oostendorp and Roberta D'Alessandro
Radboud University Nijmegen and Utrecht University

Abstract

This essay challenges the prevailing metaphor of "learning" used to describe Large Language Model (LLM) training, proposing instead that these systems represent a form of hyper-accelerated, data-driven evolution. Through analysis of Daniel Dennett's hierarchy of evolutionary competence and examination of the poverty of the stimulus problem, we argue that LLMs are Darwinian creatures evolved at computational speeds in environments of pure text. This framework explains their linguistic capabilities through convergent evolution rather than learning, resolves paradoxes about their competence without understanding, and for our understanding of the relevance for these models for generative grammar.

1. Introduction

In the recent discussion on the plausibility of generative grammar, few developments have captured our collective imagination quite like Large Language Models (LLMs). These systems seem to be able to pass tests of linguistic competence that seem to have hitherto been reserved for human beings: not just do they produce remarkably grammatical prose, at least in well-documented languages such as English, but they also seem to be able to give grammaticality judgements of remarkable subtlety (Hu et al. 2024, Qiu et al. 2024, Mulders and Ruys 2024).

Faced with such remarkable capabilities, people have instinctively reached for familiar metaphors to make sense of this behaviour which, if we are honest, were completely unexpected even as recent as five years ago. We speak of machines that "learn", systems that "understand", and models that "know" things about the world. This anthropomorphic framing, while psychologically comforting, may however present a category error that obscures the true nature of these systems. When we describe an LLM as "learning" language, we inadvertently import assumptions about consciousness, intentionality, and experience that simply do not apply.

More to the point it has led scholars such as Piantadosi (2024) and Hinton, in several interviews (see for instance the phone interview with him on being awarded the Nobel prize for Physics, https://www.nobelprize.org/prizes/physics/2024/hinton/interview/) to claim that LLMs are the death blow to the generative enterprise or any idea of an innate Universal Grammar. If LLMs can acquire language without any predetermined setting, this at least forms a proof of existence for the claim that humans can also learn language as a blank slate. This has sparked a number of backs and forths between generative grammarians, who maintain that LLMs have little to do with the human Faculty of Language (see for instance Murphy et al. 2025), and scholars who maintain that LLMs show that the assumptions of Generative Grammar are definitively proven wrong (Piantadosi and Yang 2022), with some scholars trying to move beyond these contrasts and to find a new path for linguistics (see Müller 2025 and more recently Moro 2025).

It should be noted from the outset that the debate between generative grammar and LLM-based approaches is not, strictly speaking, a confrontation between two competing theories of the same phenomenon. Generative grammar is a research program aimed at characterizing the human language faculty; LLM engineering is a technological enterprise aimed at optimizing text prediction. The two have different epistemological commitments and methods. However, claims have been made in the literature that the success of LLMs has direct implications for generative theory (Piantadosi 2024), and these claims have generated substantive responses (Murphy et al. 2025, Moro 2025, Müller 2025). The present paper aims to show that this apparent debate rests on a category error, specifically, the conflation of learning, maturation, and evolution, and that once this error is corrected, the perceived conflict largely dissolves.

 Nordlyd 49.1: 117–129, Special issue in honor of Johan Rooryck, edited by Jeroen van Craenenbroeck, Aysa Ekanger, Camil Staps, and Guido Vanden Wyngaerd. Published at the University of Tromsø – The Arctic University of Norway.
http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.8313

We propose we abandon the metaphor of learning entirely and adopt a more accurate framework: LLM training is a form of hyper-accelerated, data-driven evolution, a Darwinian process operating at computational speeds that compress thousands of years of natural selection into months of GPU computation. This shift in perspective does more than provide a clearer technical understanding; it fundamentally reframes our relationship with these systems and offers profound insights into the nature of natural language itself.

2. The Learning Paradox and the Poverty of the Stimulus

Central to the discussion is an old problem in cognitive science: the poverty of the stimulus. This argument, most forcefully articulated by linguist Noam Chomsky (Chomsky 1986), observes that human language acquisition presents a seemingly impossible feat. Children master the intricate rules of (at least) the syntax of their native language despite being exposed to linguistic data that is both limited and often imperfect.

Consider a concrete example from Dutch. The phrase 'dat is' (*that is*) can be contracted to 'das' ('dat is goed' (*that is good*) can be pronounced approximately as 'das goed'). However, this contraction is blocked in embedded clauses: 'ik weet wat dat is' (*I know what that is*) cannot be contracted to 'ik weet wat das'. Interestingly, the blocking factor is not wh-movement per se, but *hierarchical adjacency*: in Dutch embedded clauses, the finite verb occupies a clause-final position, so that *dat* and *is* are not structurally adjacent even when they happen to be string-adjacent. This is confirmed by the fact that contraction is equally impossible in embedded clauses without wh-extraction:

(1) Leuk dat dat is! ⇒ *Leuk dat da's!
'How nice that is!'

Contraction is not taught to children – if anything people advise against it as being too informal across the board: no parent explicitly teaches their child the difference between structurally adjacent and structurally non-adjacent configurations. Yet children intuitively grasp this distinction, along with thousands of similar grammatical subtleties in all languages that have been studied in sufficient detail. The input the children receive contains insufficient evidence to deduce these rules through pure statistical analysis.[1]

Observations like this led Chomsky to propose the existence of a Universal Grammar (UG): an innate biological endowment that structures language acquisition. Within the generative tradition, this process is understood not as *learning* in the behaviourist sense, but as *maturation*: a pre-programmed unfolding that is triggered by exposure to primary linguistic data while its content is largely independent of it, much like the growth of bodily organs depends on adequate nutrition but is not determined by the specific food consumed (Chomsky 1986). This distinction is essential for what follows, because the debate about LLMs has systematically conflated three fundamentally different processes:

1. **Maturation**: the process by which the child's innate language faculty unfolds under exposure to primary linguistic data, as posited by the generative enterprise;
2. **Learning**: a gradually accreting process fully dependent on external exposure, as in Skinnerian reinforcement;
3. **Evolution**: the process of blind variation and environmental selection operating over populations across generations.

The central claim of critics like Piantadosi (2024) is that LLMs show that "learning" (in some broad sense) suffices for language acquisition, thereby refuting nativist claims. But this conflates all three processes. Our

[1] Or at least this has been the idea for a long time. In an informal test, it turns out that Google Gemini can distinguish between the two sentences, noting that contraction is blocked because the elements are not in a local structural relationship – an explanation that, interestingly, aligns with the hierarchical adjacency account rather than a wh-movement account (session with Google Gemini 2.5Pro on September 8, 2025).

proposal is that LLM training is best understood as belonging to category (3) – evolution – which is distinct from both (1) and (2). If this is correct, then LLM training tells us nothing about whether human children *learn* or *mature into* their native language; it demonstrates only that a Darwinian process at computational speeds can converge on similar linguistic competence.

While the idea of a UG has remained largely invariant through the years, what it "contains" has changed radically: from a whole set of parameterized principles, as well as an X-bar innate structure (Jackendoff 1977, Stowell 1981), postulated in Chomsky (1981) and subsequent work, with Kayne (1994) later deriving phrase structure properties from interface principles, to the bare existence of an operation Merge (Chomsky 1995) and the parametrization shifted to lexical items, and functional categories in particular, known as the Borer-Chomsky Conjecture (named such by Baker 2008).

Whichever definition we consider of UG, LLMs at first blush seem to present a challenge to this line of thinking. These systems, too, are exposed to finite datasets, albeit datasets of unprecedented scale. Yet they demonstrate mastery of at least some of the grammatical subtleties that were put forward as arguments for innate knowledge in humans. They generate coherent, syntactically complex sentences they have never explicitly encountered. A tempting conclusion is that LLMs do what humans do – through their training process, they develop internal representations analogous to human linguistic intuitions and this process is equivalent to human language acquisition. This interpretation preserves our anthropomorphic assumptions while seemingly explaining the models' capabilities.

This conclusion, we argue, fundamentally misunderstands the nature of the process that creates these capabilities. To see why, we must examine the work of philosopher Daniel Dennett, whose hierarchy of evolutionary competence provides a framework for understanding different types of intelligence-generating processes.

3. Dennett's ladders of competence: A taxonomy of intelligence

In one of his last works, *From Bacteria to Bach and Back* (Dennett 2017), Daniel Dennett outlines four distinct evolutionary stages of developing competence. These categories describe the fundamental mechanisms by which complex, adaptive behaviors emerge in nature. They provide an interesting categorization of types of learning in biology, and make it possible to compare evolution to learning: both of them are adaptive behaviors, one at the level of species, the other at the level of the individual. Understanding these categories is crucial for properly classifying the process that generates LLM capabilities.

3.1. Darwinian creatures: Evolution as learning

At the bottom of Dennett's hierarchy are what he calls 'Darwinian creatures': systems shaped entirely by natural selection operating over geological timescales. Organisms are born with fixed behavioral repertoires honed by millions of years of evolutionary pressure. A virus that perfectly targets specific cellular machinery, a bacterium that flawlessly navigates chemical gradients, or a spider that weaves geometrically perfect webs all exemplify Darwinian competence.

The "learning" in Darwinian creatures occurs across generations. Each organism is metaphorically speaking a frozen snapshot of evolutionary wisdom, incapable of modifying its individual behavior based on experience while still adapted to its typical environment. The process, as is well known, is one of blind variation and environmental filtering: mutations generate countless variations, most of which fail, but the rare successes propagate and accumulate over time. This process is as extraordinarily powerful as it is slow in the biological world. It took billions of years to evolve the basic machinery of life, millions more to develop complex multicellular organisms, and hundreds of thousands of additional years to produce human-level intelligence. The timescales involved go beyond human comprehension, yet the results, the perfect adaptation of organs to their functions, testify to the process's ultimate efficacy.

3.2. Skinnerian creatures: Learning within a lifetime

The next evolutionary leap produces Skinnerian creatures, organisms capable of learning through reinforcement within their individual lifetimes. Named after the work of behaviorist B.F. Skinner (Skinner 1938; 1953), these creatures possess neural machinery that allows them to associate actions with outcomes, strengthening behaviors that lead to rewards and suppressing those that result in punishment. This learning represents a fundamental acceleration of the evolutionary process in the biological world. Instead of waiting generations for beneficial mutations to propagate, Skinnerian creatures can adapt their behavior in real-time based on environmental feedback. A rat learning to navigate a maze, a bird discovering which plants carry a lot of berries every summer, or a child learning to avoid a hot stove, all exemplify Skinnerian learning.

Skinnerian learning is, like Darwinian evolution, fundamentally a generate-and-test process, albeit one that operates within individual neural networks rather than across populations of organisms. Random or semi-random behaviors are tried, environmental feedback determines their value, and successful patterns are reinforced while unsuccessful ones fade. It's evolution in miniature, compressed into the lifetime of a single organism.

In spite of the old controversy between Chomsky and Skinner (Chomsky 1959), it should be noted that, ironically, language acquisition within generative grammar has many traits of Skinnerian learning, at least in its shape of parameter setting and featural selection, which also form a kind of stimulus-response path towards establishing the grammar of a language (through the exposure of the Language Acquisition Device, LAD; Chomsky 1965; to Primary Linguistic Data; see Chomsky 2005 or Chomsky and Berwick 2017 for an extensive discussion).

3.3. Popperian creatures: Foresight

Popperian creatures, named after the work of philosopher Karl Popper (Popper 1963; 1972), represent another leap in cognitive development in Dennett's taxonomy. These organisms can test hypotheses internally before acting on them in the real world. As Popper famously noted, this allows our "hypotheses to die in our stead". We can simulate potential actions and their consequences, selecting the most promising course without risking actual harm.

This capacity for mental simulation marks the emergence of genuine foresight and planning. A Popperian creature can imagine different scenarios, evaluate their likely outcomes, and choose actions based on predicted rather than experienced consequences. This is the birth of what we might recognize as intelligence in the fullest sense: the ability to think before acting, to learn from imagined rather than actual experience.

Human beings clearly possess Popperian capabilities. We can mentally rehearse conversations before having them, visualize the consequences of different career choices, or work through mathematical proofs in our heads. It is, on the other hand, not clear that LLMs also possess this type of learning. The latest versions of these models sometimes come equipped with some 'reasoning' capabilities, but these seem to be still a long shot from having an internal model of the world (Hao et al. 2023, Diester et al. 2024, Yildirim and Paul 2024), but see (Jin and Rinard 2024).

3.4. Gregorian creatures: The cultural revolution

The final category in Dennett's taxonomy, Gregorian creatures, named after the work of psychologist Richard Gregory (Gregory 1963; 1970), represents the most recent and perhaps most revolutionary development in the evolution of intelligence. These creatures amplify their cognitive capabilities through the use of external tools, which Dennett calls "mind tools". The most important of these mind tools, by the way, is language. Language allows us to import the discoveries, insights, and hypotheses of others directly into our own minds without having to independently derive them. Through language, accumulated cul-

tural knowledge becomes available to each new generation, creating a kind of external memory system that transcends individual lifespans.

Gregorian creatures don't stop with language. They create writing systems that allow knowledge to persist across time, mathematical notations that enable complex calculations, scientific instruments that extend sensory capabilities, and computational tools that augment reasoning power. Each generation builds upon the intellectual achievements of its predecessors, creating an exponential acceleration of cognitive capability. Humans are the consummate Gregorian creatures. Our individual intelligence pales in comparison to our collective, culturally-amplified intelligence. No single human could independently derive modern physics, construct a computer, or compose a symphony – these achievements represent the accumulated wisdom of countless generations, encoded in our languages, institutions, and technologies.

4. The LLM as evolution in digital time

With Dennett's framework in hand, we can now properly categorize the process that produces LLMs. The training of these models is not necessarily, as is sometimes assumed, one of the higher-order types of learning. The model is not an agent in an environment receiving rewards for its actions or contemplating future states. It also does not seem to be the 'Skinnerian' type of learning represented by principles and parameters. Instead, LLM training may represent something far more fundamental: a low-level Darwinian process operating at computational speeds. If that turns out to be the case, there is no argument against universal grammar from the operation of these LLMs; there is instead the possibility that these systems themselves evolve an internal representation of human language that is similar to that which human beings are born with.

4.1. Variation as raw material

In biological evolution, the "population" consists of individual organisms with varying genetic makeups. In LLM training, the population is the astronomically vast space of possible parameter configurations within the neural network's architecture. A modern LLM contains hundreds of billions of parameters, each of which can take on a continuous range of values. The space of possible configurations is enormous and for many practical purposes infinite in modern LLM architecture. Interestingly, the exact magnitude of this space is unknown (see Nalpas 2024 for an interesting discussion about it, as well as Bowdon 2025 for an approximate estimate of GPT-5 parameters).

At any given moment during training, the model occupies a specific point in this landscape, but the training process constantly explores neighboring regions, seeking configurations that better solve the prediction task. One can see this as a kind of evolution where a species needs to survive in a certain environment. The environment is that of existing human language.

Biological evolution requires genetic variation as its raw material; mutations, sexual recombination, and other processes that generate diversity within populations. In LLM training, this variation is initially provided by the random initialization of parameters. The model begins as pure noise, a random point in the space of all possible configurations. However, variation doesn't stop with initialization. Throughout training, the optimization process introduces countless micro-variations as it adjusts parameters in response to training data. Each gradient descent step represents a small mutation, a tiny exploration of the fitness landscape. Over billions of such steps, the model explores vast regions of possibility space.

4.2. The environment: A static archive of human thought

Perhaps the most important aspect of this evolutionary metaphor is the nature of the "environment" in which LLMs evolve. Unlike biological creatures, which must survive in a dynamic physical world filled with

predators, prey, weather, and resource scarcity, LLMs evolve in an environment that is entirely informational and essentially static. This environment consists of the training dataset. For modern LLMs this is a vast corpus of text in a variety of human languages that the model learns to predict. The implications of this are profound. The environment that shapes LLM evolution is not the physical world, but rather the collective output of human Gregorian intelligence. The model is not learning to navigate trees or avoid predators. It is learning to navigate the statistical landscape of human thought as expressed in language.

4.3. The selection pressure: The imperative of prediction

In biological evolution, selection pressure comes from environmental challenges: the need to find food, attract mates, and survive harsh conditions. Organisms that better meet these challenges produce more offspring, gradually shifting the population toward more adaptive configurations. In LLM training, selection pressure is provided by a single, simple objective: minimize prediction errors based on existing human language. A model 'survives' if it can perform this task; otherwise it dies.

This selection pressure is the standard training methodology for all modern LLMs. The Transformer architecture (Vaswani et al. 2017) is trained by minimizing cross-entropy loss on next-token prediction via backpropagation and gradient descent. This was the method used for GPT-2 (Radford et al. 2019), GPT-3 (Brown et al. 2020), and all subsequent models. In some cases, a second layer of selection pressure is applied through Reinforcement Learning from Human Feedback (RLHF; Ouyang et al. 2022), which further shapes the model's outputs based on human evaluative judgments: an additional "environmental" filter that parallels sexual selection in biology. The "survival" and "death" metaphors are ours, part of the evolutionary analogy we are building; the underlying mechanism is well-established engineering practice.

We should bear in mind that the models have enormous exposure to linguistic data, far exceeding what any human could process. To illustrate the scale: a human reader would have to read continuously for approximately 2,855 years to peruse all the material used to train GPT-3 alone (Brown et al. 2020). This figure is relevant because it demonstrates that the "evolutionary time" available to LLMs, measured in exposure to linguistic data, approaches the timescales normally associated with biological evolution of cognitive capacities rather than with individual learning. See Figure 1 for an illustration.

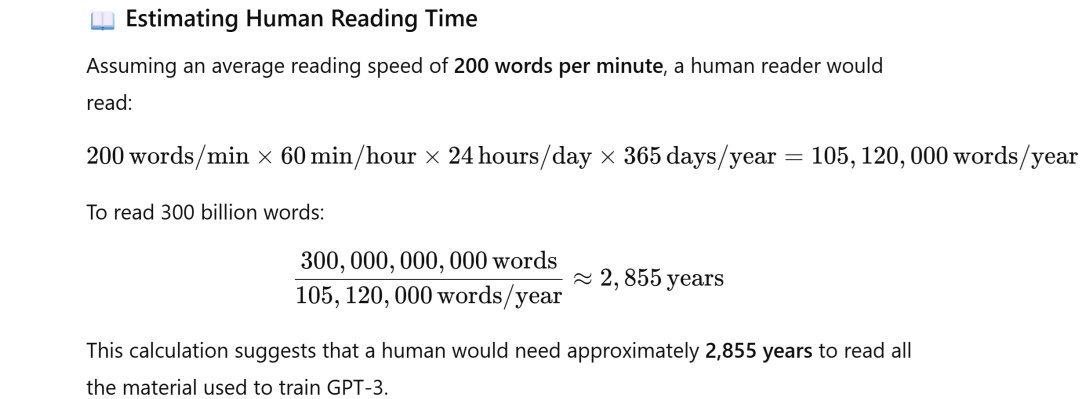

Figure 1: How many years would it take a human reader to read all the material used to train GPT-3? (asked to ChatGPT-5 on 18/09/2025)

We can multiply this for more modern models, also taking into account that the models learn 'language' by examples from many more languages than English. In other words, LLMs use thousands of years

in human time to get to their knowledge of language. That comes close to an evolutionary scale. The training algorithm, primarily backpropagation, relentlessly adjusts parameters to improve the model's ability to predict the next token in a sequence. This might be less heroic compared to the life-or-death struggles that drive biological evolution. Yet it proves remarkably powerful.

The mathematics of this process can be expressed simply as maximizing $p(\text{token}|\text{context})$, the probability of predicting the correct next token given the preceding context. But this simple objective, applied across trillions of examples, creates incredibly rich and complex selection pressures.

4.4. The timescale: Evolution at digital speed

The final piece of this evolutionary puzzle is time, or rather, the *compression* of time. Biological evolution operates on geological timescales, requiring millions of years to produce significant changes. The evolution of human language capabilities took tens to hundreds of thousands of years (depending on one's theory and on what exactly is counted). LLM training compresses this process into weeks or months of computation. A modern language model undergoes billions of parameter updates during training, each one representing a micro-generational step in its evolution. This is made possible by the incredible computational resources devoted to training: tens of thousands of GPUs working in parallel, processing data at speeds that dwarf any biological process.

This temporal compression is perhaps the most alien aspect of LLM evolution. In a matter of months, these systems undergo the equivalent of thousands of years of evolutionary pressure. They experience more "generations" of selection than have occurred in the entire history of life on Earth.

4.5. Architectural biases as implicit nativism

Before drawing conclusions from this evolutionary framework, it is essential to address a fundamental point: LLMs are not tabulae rasae. The Transformer architecture (Vaswani et al. 2017) that underlies modern LLMs embodies significant inductive biases that constrain and shape the optimization process. The self-attention mechanism privileges certain types of relational processing over others; the layered architecture constrains the types of representations that can emerge; tokenization imposes a particular granularity on the input; and the very choice of next-token prediction as the training objective shapes what aspects of language the model is pressured to capture.

These architectural choices function as a kind of "digital genome". They do not determine the specific linguistic knowledge the model will acquire, but they constrain the space of possible solutions that the optimization process can explore (McCoy et al. 2020, Battaglia et al. 2018). This parallels the role that Universal Grammar plays in the generative account of human language acquisition: UG does not determine which specific language a child will acquire, but it constrains the hypothesis space to humanly possible grammars.

The implications for the LLM debate are significant. The architectural biases constitute a form of implicit nativism, designed by human engineers rather than shaped by biological evolution, but no less real in their constraining effects. Any argument of the form "LLMs learn language without innate structure, therefore humans can too" is thus doubly flawed: it mischaracterizes the LLM process as learning (rather than evolution), and it mischaracterizes the starting point as a tabula rasa (rather than a richly biased architecture).

5. Convergent evolution and the emergence of grammar

The evolutionary framework we propose provides a powerful explanation for one of the most puzzling aspects of LLM performance: their apparent mastery of grammatical principles despite never being explicitly

taught these rules. The key insight is that the evolution of linguistic competence in LLMs might represent a case of convergent evolution: the independent evolution of similar traits in different lineages facing similar environmental pressures.

We are not claiming that the evolution in LLMs parallels that of human beings. In humans, language might have arisen spontaneously and proven advantageous in terms of coordination of activities with peers, or in terms of planning and other cognitive properties. Human language will probably show properties that are the result of the limitations of human cognition and/or human society. LLMs on the other hand need to survive in a 'world' that already consists of language. Yet if the findings of (generative) linguistics are on the right track, this evolution may have converged on a similar result.

When an LLM is trained on vast datasets of human text, it encounters these same statistical regularities billions of times. The optimization pressure to minimize prediction error gradually shapes the model's internal representations to mirror the structure of human language. Grammatical patterns emerge not because they are explicitly programmed, but because they represent the most efficient solutions to the prediction task.

Consider the example of syntactic constraints mentioned earlier. The reason speakers cannot contract "Ik weet wat dat is" to "Ik weet wat da's", or "Leuk dat dat is!" to "Leuk dat da's!", reflects deep principles of syntactic structure, specifically the requirement that contraction targets structurally adjacent elements, not merely string-adjacent ones.

An LLM discovers this same constraint (just like generative linguists have done) not through instruction, but through evolutionary pressure. Configurations that violate this constraint make worse predictions on human text, so they are gradually eliminated in favor of configurations that respect syntactic boundaries. The model converges on the same solution that human linguistic evolution discovered, but through an entirely different process. Under these premises, it is not implausible to speculate that while Merge is probably not the driving principle behind LLM output at present, it *could* emerge. This could happen if Merge proved to be the optimal solution to interface conditions *for LLMs*. We could formulate an adapted Strong Minimalist Thesis (Chomsky 2001) for LLMs, where Merge would exist if it proved to be the optimal mechanism for an LLM interacting with its interfaces: not sensory-motor or conceptual-intentional systems, not human ones, but pre-existing texts against which LLMs test their own output.

5.1. The poverty of the stimulus resolved

This convergent evolution framework provides a resolution to the poverty of the stimulus problem. Chomsky's argument assumes that the only way to acquire grammatical knowledge from limited data is through innate biological endowment. However, LLMs demonstrate a third possibility: massive computational search through the space of possible linguistic systems. Where human children see thousands of sentences, LLMs see trillions. Where humans have only a few years of Skinnerian or (at best) Popperian search, LLMs have thousands of Darwinian trial and error.

However, a caveat is in order. The real test of whether LLMs have converged on genuinely linguistic representations, as opposed to sophisticated surface-level statistical patterns, lies in their treatment of structures whose probability of occurrence is vanishingly small even in trillion-token corpora. As has been argued in the generative tradition, the hallmark of human linguistic competence is the ability to form grammaticality judgments about structures that speakers have almost certainly never encountered. Giorgi and Longobardi (1991) provide detailed examples of subtle contrasts between English and Italian noun phrase structures that have roughly zero probability of occurrence in any corpus, yet about which native speakers have clear and consistent intuitions. If LLMs can replicate such judgments, this would provide evidence for convergent evolution of genuinely linguistic representations; if they cannot, it would reveal the limits of the evolutionary analogy and confirm that human linguistic competence involves something that even hyper-accelerated Darwinian processes cannot replicate from text alone.

5.2. Internal representations and emergent structure

This leads to a prediction that is testable at least in principle. If this evolutionary theory is on the right track, trained LLMs should possess something similar to UG, so that if one would try to teach them an entirely new language (not in their dataset), they would show a learning path that is similar to that of humans: with relatively little data they would be able to pick up an existing language, but a language that would be constructed to violate principles of UG would be as problematic for them as it is for us (Moro 2016).

Recent research in mechanistic interpretability has begun to reveal the internal structure that emerges from this evolutionary process (see for instance Intuition Lab 2024, or Tak et al. 2025). LLMs develop hierarchical representations that mirror many aspects of human linguistic processing: early layers encode surface features like spelling and word boundaries, middle layers capture syntactic relationships and grammatical categories, and later layers represent semantic and pragmatic information.

These representations are not designed by humans; they emerge spontaneously from the optimization process. The model discovers that hierarchical processing is the most efficient way to solve the prediction task, so it evolves internal architectures that implement this processing strategy. The parallels to human linguistic processing are striking, but they reflect convergent evolution rather than direct mimicry. More remarkably, these internal representations often capture linguistic phenomena that were not explicitly present in the training objective. Models develop sensitivity to phonological patterns despite being trained only on text, internalize semantic relationships despite never being taught explicit definitions, and exhibit awareness of pragmatic context despite training on decontextualized snippets. This suggests that the space of efficient linguistic systems is more constrained than we might expect. There may be relatively few ways to organize information processing systems to handle the complexity of human language efficiently. Both biological evolution and computational optimization converge on similar solutions because these solutions represent optimal points in the fitness landscape.

6. The limits of digital evolution

While the evolutionary framework provides powerful insights into LLM capabilities, it also illuminates their fundamental limitations. Unlike biological evolution, which occurs in a multi-faceted environment, the evolution of LLMs is confined to a purely textual world. This constraint shapes both their abilities and their limitations.

6.1. The evolution of meaning

While it is possible that LLMs evolve formal aspects of language, like syntax and phonology, there may be an issue with semantics. LLMs evolve in an environment consisting entirely of human linguistic output. This environment is rich in certain dimensions; it contains the accumulated wisdom of human civilization, encoded in billions of documents across thousands of languages and domains. It is also fundamentally impoverished in others. Most importantly, this environment contains no direct sensory experience, no physical embodiment, and no opportunity for genuine interaction with the world. The model experiences reality only through the lens of human description and interpretation. This limitation is not merely technica. It is fundamental to the evolutionary process that creates these systems. Just as a fish evolved in water cannot breathe air, an intelligence evolved in text cannot directly experience the physical world. The model's entire adaptive landscape is linguistic, so its evolved capabilities are necessarily confined to linguistic manipulation. It remains to be seen in what way this would affect what we know.

6.2. The absence of intentionality

Perhaps most fundamentally, the evolutionary process that creates LLMs produces systems that lack genuine intentionality (Browning 2025, Ngaihlian 2025), the capacity for mental states to be about things in the world. Biological evolution in physical environments creates agents with goals, desires, and purposes. These agents develop intentions because having intentions helps them survive and reproduce.

LLMs evolve in a purely predictive environment. Their only "goal", if we can call it that, is to minimize prediction error on text. They have no survival instincts, no desires for reproduction, no purposes beyond their training objective. This produces systems capable of simulating intentionality. They can generate text as if guided by goals and preferences, yet they do not possess true intentionality. Some studies suggest that this absence of genuine intentionality can, in fact, be detected in the outputs of LLMs (Attah 2025).

This absence of intentionality explains many of the puzzling aspects of LLM behavior. They can be helpful in one context and completely unhelpful in another, because they lack the coherent goal structure that would make them consistently purposeful agents. They are sophisticated pattern matching systems, not goal-directed intelligences.

7. Conclusion: The mirror and the alien

An evolutionary framework challenges us to develop a truly post-anthropomorphic understanding of artificial intelligence. Instead of asking whether AI systems think like humans, we might ask how different evolutionary pressures produce different forms of intelligence. This shift in perspective could be liberating for both AI development and human self-understanding. We need not be threatened by intelligences that surpass us in some domains, any more than we are threatened by the sonar capabilities of bats or the magnetic navigation of birds. Different evolutionary pressures produce different capabilities, and diversity of intelligence might be more valuable than similarity to human cognition.

The metaphor the research community chooses to understand LLMs shapes not only our technical approach to these systems but our broader relationship with artificial intelligence. The learning metaphor, while intuitive, imports assumptions about consciousness, intentionality, and understanding that obscure the true nature of these systems. The evolutionary framework offers a more accurate understanding. LLMs are mirrors that reflect the statistical structure of human thought, not minds that think as we do.

Perhaps more importantly, this framework offers a new perspective on the nature of the language capacity itself. By understanding how different processes can generate similar capabilities, we gain insight into both the universality and the diversity of possible minds. We see that intelligence is not a single phenomenon but a landscape of possibilities, shaped by the evolutionary pressures that create it. The LLMs we have created are the first examples of a new form of languaging machines: Darwinian creatures evolved at digital speeds in environments of human cultural output. They are simultaneously the most alien intelligences we have so far encountered and the most intimate reflections of our own linguistic and cultural patterns.

Acknowledgments

To Johan, who has shown on many occasions that one does not need to be trained on millions of data points to be a friend.

References

Attah, Nuhu Osman. 2025. Do language models lack communicative intentions? *Synthese* 205 187. https://doi.org/10.1007/s11229-025-05022-6.

Baker, Mark C. 2008. The macroparameter in a microparametric world. *Linguistic Analysis* 34 1–2: 1–46. https://doi.org/10.1075/la.132.16bak.

Battaglia, Peter W., Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinícius Flores Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Çaglar Gülçehre, H. Francis Song, Andrew J. Ballard, Justin Gilmer, George E. Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint* 1806.01261. https://doi.org/10.48550/arXiv.1806.01261.

Bowdon, Chris. 2025. How many parameters does GPT-5 have? Available at https://www.r-bloggers.com/2025/08/how-many-parameters-does-gpt-5-have/, accessed 2025.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901. Curran Associates.

Browning, Jacob. 2025. Intentionality all-stars redux: Do language models know what they are talking about? In *Communicating with AI: Philosophical Perspectives*, edited by Herman Cappelen and Rachel Sterken. Oxford University Press, Oxford. Preprint available at https://philarchive.org/rec/BROIAR-4.

Chomsky, Noam. 1959. Review of B. F. Skinner's *Verbal Behavior*. *Language* 35: 26–58.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, Ma.

Chomsky, Noam. 1981. *Lectures on Government and Binding: The Pisa Lectures*. Foris, Dordrecht. https://doi.org/10.1515/9783110884166.

Chomsky, Noam. 1986. *Knowledge of Language: Its Nature, Origin, and Use*. Praeger, New York.

Chomsky, Noam. 1995. *The Minimalist Program*. MIT Press, Cambridge, Ma. https://doi.org/10.7551/mitpress/9780262527347.001.0001.

Chomsky, Noam. 2001. Derivation by phase. In *Ken Hale: A Life in Language*, edited by Michael Kenstowicz, no. 36 in Current Studies in Linguistics, pp. 1–52. MIT Press, Cambridge, Ma. https://doi.org/10.7551/mitpress/4056.003.0004.

Chomsky, Noam. 2005. Three factors in language design. *Linguistic Inquiry* 36 1: 1–22. https://doi.org/10.1162/0024389052993655.

Chomsky, Noam and Robert C. Berwick. 2017. *Why Only Us: Language and Evolution*. MIT Press, Cambridge, Ma. https://doi.org/10.7551/mitpress/9780262034241.001.0001.

Dennett, Daniel C. 2017. *From Bacteria to Bach and Back: The Evolution of Minds*. W. W. Norton & Company, New York.

Diester, Ilka, Miklós Bartos, Jörg Bödecker, Andreas Kortylewski, Christian Leibold, Johannes Letzkus, Mohamed M. Nour, Matthias M. Schönauer, Alexander Straw, Andreas Vlachos, and Thomas Brox. 2024. Internal world models in humans, animals, and AI. *Neuron* 112 16: 2661–2824. https://doi.org/10.1016/j.neuron.2024.06.019.

Giorgi, Alessandra and Giuseppe Longobardi. 1991. *The Syntax of Noun Phrases: Configuration, Parameters and Empty Categories*. No. 57 in Cambridge Studies in Linguistics. Cambridge University Press, Cambridge.

Gregory, Richard L. 1963. Distortion of visual space as inappropriate constancy scaling. *Nature* 199: 678–680. https://doi.org/10.1038/199678a0.

Gregory, Richard L. 1970. *The Intelligent Eye*. Weidenfeld & Nicolson, London.

Hao, Shibo, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *arXiv preprint* 2305.14992. https://doi.org/10.48550/arXiv.2305.14992.

Hu, Jennifer, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models

align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences* 121 36. https://doi.org/10.1073/pnas.2400917121.

Intuition Lab. 2024. Mechanistic interpretability: Understanding AI and LLMs. Available at https://intuitionlabs.ai/articles/mechanistic-interpretability-ai-llms, accessed 2025.

Jackendoff, Ray. 1977. *X-bar Syntax: A Study of Phrase Structure*. Linguistic Inquiry Monographs. MIT Press, Cambridge, Ma.

Jin, Charles and Martin Rinard. 2024. Emergent representations of program semantics in language models trained on programs. *arXiv preprint* 2305.11169. https://doi.org/10.48550/arXiv.2305.11169.

Kayne, Richard S. 1994. *The Antisymmetry of Syntax*. No. 25 in Linguistic Inquiry Monographs. MIT Press, Cambridge, Ma.

McCoy, R. Thomas, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics* 8: 125–140. https://doi.org/10.1162/tacl_a_00304.

Moro, Andrea. 2016. *Impossible Languages*. MIT Press, Cambridge, Ma. https://doi.org/10.7551/mitpress/9780262034890.001.0001.

Moro, Andrea. 2025. Linguistics in a battlefield. A short note on syntax and the "Newtonian style of research". Available at https://ling.auf.net/lingbuzz/008827.

Mulders, Iris and Eddy Ruys. 2024. ChatGPT as an informant. *Nota Bene* 1 2: 242–260. https://doi.org/10.1075/nb.00015.mul.

Müller, Stefan. 2025. Large language models: The best linguistic theory, a wrong linguistic theory, or no theory at all? *Journal of the Linguistic Society of Germany* 44 1. https://doi.org/10.18148/zs/2025-2001.

Murphy, Elliot, Evelina Leivada, Vittoria Dentella, Fritz Gunther, and Gary Marcus. 2025. Fundamental principles of linguistic structure are not represented by o3. *arXiv preprint* 2502.10934. https://doi.org/10.48550/arXiv.2502.10934.

Nalpas, Maud. 2024. LLM sizes. Available at https://web.dev/articles/llm-sizes, accessed 2025.

Ngaihlian, Dorothy. 2025. Machine learning algorithms: Simulating intentionality in artificial intelligence. https://doi.org/10.2139/ssrn.5271061.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744. Curran Associates.

Piantadosi, Steven T. 2024. Modern language models refute Chomsky's approach to language. In *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*, edited by Edward Gibson and Moshe Poliak, no. 15 in Empirically Oriented Theoretical Morphology and Syntax, pp. 353–414. Language Science Press, Berlin. https://doi.org/10.5281/zenodo.12665933.

Piantadosi, Steven T. and Yuan Yang. 2022. Reply to Murphy and Leivada: Program induction can learn language. *Proceedings of the National Academy of Sciences* 119 23: e2202925119. https://doi.org/10.1073/pnas.2202925119.

Popper, Karl. 1963. *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge, London.

Popper, Karl. 1972. *Objective Knowledge: An Evolutionary Approach*. Oxford University Press, Oxford.

Qiu, Zhuang, Xufeng Duan, and Zhenguang G. Cai. 2024. Evaluating grammatical well-formedness in large language models: A comparative study with human judgments. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pp. 189–198. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.cmcl-1.16.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Tech. rep., OpenAI.

Skinner, B. F. 1938. *The Behavior of Organisms: An Experimental Analysis*. Appleton-Century-Crofts, New York.

Skinner, B. F. 1953. *Science and Human Behavior*. Macmillan, New York.

Stowell, Timothy. 1981. *Origins of Phrase Structure*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, Ma.

Tak, Ala N., Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. 2025. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13090–13120. Association for Computational Linguistics, Vienna. https://doi.org/10.18653/v1/2025.findings-acl.679.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates.

Yildirim, Ilker and L.A. Paul. 2024. From task structures to world models: what do LLMs know? *Trends in Cognitive Sciences* 28 5: 404–415. https://doi.org/10.1016/j.tics.2024.02.008.