THE INDIA FORUM
A JOURNAL-MAGAZINE ON CONTEMPORARY ISSUES

June 23, 2026

# Beyond Big Data: Language, Biology, and the Limits of Next-Token AI

**By: Sayantan Mandal,Rakesh Sengupta**

*LLMs are powerful statistical learners, but human language acquisition is biologically constrained and structurally selective. For a multilingual society, conflating LLM usefulness with genuine understanding carries real costs, particularly in education.*

When you type an email and your phone suggests the next word, you are encountering the central mechanism behind systems such as ChatGPT. Large language models (LLMs) are trained to predict the next word in a sequence. Unlike your messaging app, however, "next" could mean the 501st item. They do so with remarkable accuracy. It is tempting, then, to assume that this is also how the human brain processes language. When we listen, do we not also anticipate what comes next?

> Human listeners first group words into structured units and only then generate predictions within those units. This difference is not merely technical. LLMs predict continuously using statistical patterns without an intermediate structure-building stage.

Because LLMs excel at mimicking the superficial structures of the corpora on which they are trained, it has become commonplace to hear from over-enthusiastic proponents that these systems are adequate models of Natural Language. Geoffrey Hinton himself has gone on record multiple times claiming that Natural Language is inherently associative; that is, its structural constraints are entirely predictable through next-token statistical evaluation.

This is a bold claim-and unfortunately one that is entirely unfounded. Decades of research in Generative Grammar, Biolinguistics and Evolutionary Biology have uncovered intricate evolutionary constraints that have shaped Natural Language and its implementation in the brain, and Hinton's claims run afoul of all of them.

Using magnetoencephalography (MEG) and behavioural "cloze" tests with Mandarin and English speakers, researchers recently compared brain responses to word predictability against LLM predictions (Poeppel et al. 2026). The core finding shows a clear divergence: while LLMs predict uniformly across all sequences, the human brain's responses varied systematically depending on a word's position within hierarchical grammatical constituents (phrases and clauses; see TP, AP etc. in Figure 1).

Human listeners first group words into structured units and only then generate predictions within those units. This difference is not merely technical. LLMs predict continuously using statistical patterns without an intermediate structure-building stage. Humans, on the other hand, are biologically constrained to evaluate statistical patterns only within a priori determined structures that are hierarchical in their relationship.

But before we get into the hierarchies, let us take a look at something a little more basic. The primary mode of interaction with LLMs is through text. This is both the mode of interaction with LLMs and the primary corpus on which they are trained. Verbal interaction, and auditory responses, are possible, though the basic issues remain the same there.

A crucial fact often overlooked in pedantic uses of "language" is that it is primarily an acoustic phenomenon, with a secondary visual modality available in sign languages. Everyday usage tends to include orthography, but this is problematic for several reasons. As Darwin famously noted in *The Descent of Man* (1871), "Man has an instinctive tendency to speak ... whilst no child has an instinctive tendency to bake, brew, or write."

> After all, when you hear someone talk, you hear words follow each other without overlapping or blending into each other. One of the foundational discoveries made by cognitive scientists … is that this is really not true.

Writing is a human artifact, invented perhaps three or four times across history. Natural Language, by contrast, is an evolved biological capacity (Lenneberg 1967; Pinker 1995; Chomsky 2005), with a genetically invariant developmental trajectory (Brown 1973; Tsimpli 1992; Nomura 2006)-hence, Natural Language. Any child acquires any language on mere exposure, without instruction, making few if any mistakes (Snyder 2007, 2021), and rapidly converging on a system that can generate an infinite number of utterances.

But speech is transient; writing makes it permanent. It is this permanent corpus that LLMs are largely trained on. But the question scientists are increasingly asking is: what exactly do LLMs learn, and is it the same as what a child acquires (Marcus 1998, 2025; Mitchell and Krakauer 2023; Chomsky 2023)?

Consider what you are doing right now as you read this article. Your eyes move from word to word; each word neatly separated from the next by blank space. It is commonplace to think this is how we speak as well. After all, when you hear someone talk, you hear words follow each other without overlapping or blending into each other. One of the foundational discoveries made by cognitive scientists in the last century is that this is really not true.

Speech signals are messy, noisy sound streams that provide no reliable cues for word boundaries (Cutler 2012). For a child, language does not arrive as neatly separated words. As a matter of fact, there can be more silence within a single word than between two different words. Plosive sounds such as "p" or "t" briefly block airflow, creating closures inside words like "captain". If we relied on pauses to detect word boundaries, we would be systematically misled. The clean temporally separated word-streams we hear when talking are not properties of the signal. They are percepts created by our brain through a process called "chunking" (Miller 1951).

As if this was not bad enough, speech signals also come mixed with various other sounds and noises-the AC running, dogs barking, or wind blowing. A great many of these sounds share acoustic properties with speech sounds. For instance, the first sound in a word like "sin" is no different than the hissing of a snake. For a baby, this creates a peculiar problem, often called "framing", that continues to be a central topic in cognitive science.

Consider a noise-isolating microphone: it can separate speech from background noise only because it has been engineered with a built-in system-a set of pre-programmed filters-that distinguishes between the two. The human infant, lacking any such manual programming, must nevertheless solve the same problem. This means the child must be born with some neurobiological endowment that can parse and separate language-related sounds from the broader acoustic environment.

> Once the child has identified that a stretch of sound counts as a word, a second, even deeper problem emerges: how those words relate to each other.

A child acquiring a language must first be able to tell which sounds are language-related, and then having identified a stream be able to decide what counts as a suitable "chunk"-in this case, words and syllables-for interpretation. Exactly the same problem recurs in sign languages, where the visual input is continuous, lacking segmentation markers. The learner's mind must possess some parsing strategy that is capable of chunking continuous signals into the right-sized discrete chunks that form our percepts.

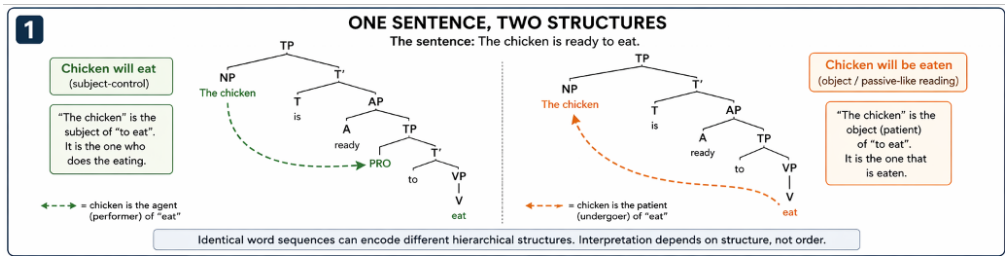
LLMs do not face this problem. Their input is already segmented into tokens-words or sub-word units separated by boundary markers. The world arrives pre-structured for the machine. The child must construct structure from noise.

Once the child has identified that a stretch of sound counts as a word, a second, even deeper problem emerges: how those words relate to each other. And here, the difference between the child and the machine becomes even more striking.

Human language is not merely a sequence of words; it is organised hierarchically. Sentences consist of nested structures-phrases within phrases-and meaning depends on these relationships. Imagine a necklace: each bead follows the next in a simple line. That is how LLMs see language-as a string.

But human language is more like a tree, with branches that split and nest inside one another. Consider the sentence: "The chicken is ready to eat." As noted in classic work by Thomas G. Bever (1970), this sentence has two interpretations. It could mean that the chicken is about to eat something (perhaps some grain), or that it is ready to be eaten (cooked and on a plate). The words and their order are identical. What differs is the underlying structure-specifically, how the noun phrase "the chicken" relates to the verb "eat". The reader can easily think of comparanda from their native tongues.

The contrast is illustrated in Figure 1. Look at the left side of the diagram. Here, "the chicken" is the subject of "to eat"-the chicken is the one doing the eating. Now look at the right side. Here, "the chicken" is the object (or patient) of "to eat"-the chicken is the one being eaten. The diagram shows this difference using tree-like structures: in one case, the chicken sits above the action as its agent; in the other, it sits below the action as its recipient.




The words never move. Only the invisible grammatical tree changes. A five-year-old child hearing this sentence in context will know immediately who is eating whom. That is because the human brain is hardwired to parse language hierarchically-to build trees, not just strings. LLMs can approximate some distinctions through statistical patterns-for example, they might learn that "chicken" is often associated with "eat" in two different ways-but they have no principled, built-in constraint that forces one interpretation over the other based on grammatical structure.

This is why the human brain's predictions are chunked by grammatical constituents (phrases and clauses), while LLM predictions are uniform across the entire sequence. The brain builds a tree first. The LLM sees only the next token. Merely increasing the number of tokens visible cannot solve this problem (Marcus 2025).

> If language is this complex, with hidden trees and discourse-level dependencies, how do children ever figure it out from the messy, noisy speech they hear?

But hierarchy does not stop at the sentence boundary. Consider two short passages. First: "The white rabbit jumped from behind the bushes. The animal looked around and then he ran away." Here, "the white rabbit", "the animal", and "he" are all the same creature. Now reverse the order: "The animal looked around and then he ran away. The white rabbit jumped from behind the bushes." Human interpretation shifts. Human readers now treat "the animal" and "the white rabbit" as different entities.

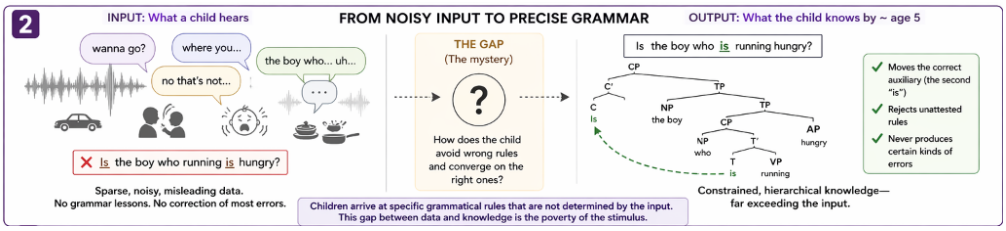
The order of mention changes how we assign reference-not because of word proximity, but because humans track how entities are introduced and maintained in discourse structure. LLMs' performance on these remain unreliable because they lack the underlying structured representations that make human reference tracking reliable.

This brings us to a deeper puzzle-what Noam Chomsky calls Plato's Problem. If language is this complex, with hidden trees and discourse-level dependencies, how do children ever figure it out from the messy, noisy speech they hear?

Consider the facts. A child does not hear billions of clean sentences. She hears fragments, interruptions, and incomplete utterances. Correction is rare and rarely grammatical. Yet, within a few years, she converges on systematic grammars. This gap between impoverished input and rich knowledge is known as the poverty of the stimulus.

Take a simple example. From "The boy is hungry", a child forms "Is the boy hungry?"-moving the auxiliary "is" to the front. Now take "The boy who is running is hungry". The correct question is "Is the boy who is running hungry?"-the second "is" moves, not the first. A rule based purely on linear order ("move the first 'is'") would fail.

Children nevertheless produce the correct form, relying on structural principles that distinguish main clauses from embedded clauses-even though no one ever teaches them this rule. Figure 2 illustrates this gap: from noisy, incomplete data on the left, children arrive at constrained, hierarchical knowledge on the right.

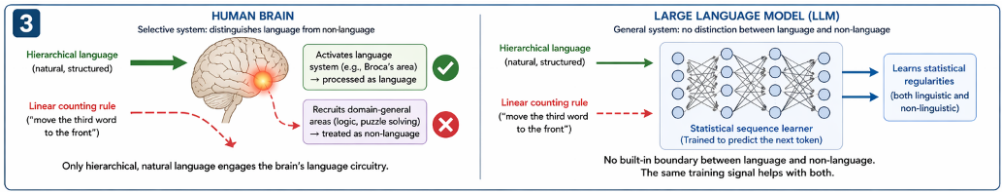



The emergence of Nicaraguan Sign Language-created spontaneously by deaf children without any prior linguistic model-drives the point home (Senghas and Coppola 2001). The human brain does not merely learn language from input; it actively constructs it under internal,

biological constraints (Lenneberg 1967; Hauser et al. 2002; Berwick and Chomsky 2016).

The distinction between human and machine learning becomes even sharper when we ask the question, "What is an impossible natural language?". Research in biolinguistics has systematically characterised the kinds of hierarchical syntax that constrains the scope of possible human languages under Universal Grammar, or the Faculty of Language (Chomsky 1986; Hauser et al. 2002; Berwick and Chomsky 2016).

Andrea Moro has spent decades conducting experiments in which participants are exposed to artificial "languages" based on linear counting rules rather than hierarchical structure seen in Natural Languages (Moro 2008). Neuroimaging studies show that when participants attempt to learn these systems, the brain's language areas-including Broca's area-are not engaged (Moro 2008). Instead, regions associated with general problem-solving are activated (Musso et al. 2003), as illustrated by Figure 3.




Moro's findings reveal a negative constraint: certain logically possible patterns are excluded from the human language faculty. Likewise, a recent large-scale study by Verkerk and colleagues (2025) provides evidence for positive constraints-patterns that recur across unrelated languages despite vast geographic and historical separation. Using Bayesian spatio-phylogenetic methods on more than 1,700 languages, the researchers found that linguistic patterns are not random; they have evolved repeatedly across language families, suggesting that shared cognitive and communicative pressures push human languages towards a limited set of preferred grammatical solutions.

Converging evidence from multiple disciplines point to the same conclusion: the human brain does not learn just any statistical pattern. It is born with a biological filter that permits only certain kinds of grammatical systems-and actively gravitates towards them.

None of this diminishes the engineering achievement of modern AI. LLMs assist with writing, summarisation, and coding-their usefulness is undeniable. But usefulness is not understanding. The confusion arises because fluent output invites anthropomorphic interpretation.

We are accustomed to associating well-formed language with intelligence. Yet, scientific evidence shows that even in the narrow task of next-word prediction, humans and machines operate differently. Humans build structure first and predict within it. LLMs predict continuously across sequences.

The deeper issue is not whether LLMs use structure at all-they can at times approximate it-but whether structure constrains what they can learn. Natural language is not just structured; it is selectively structured. Not all logically possible systems count as language, and human learners converge on a restricted subset despite limited data.

LLMs, by contrast, are unconstrained statistical learners. They cannot distinguish, in principle, between natural language and arbitrary patterns. As Chomsky has repeatedly argued, any improvement in an LLM's performance on natural language is accompanied by a corresponding improvement in its ability to learn arbitrary, non-linguistic sequences (Chomsky 2023).

> We insist, however, that usefulness and pedagogical equivalence are different claims, and conflating them carries costs specific to multilingual contexts.

These distinctions are not confined to a disciplinary quarrel between cognitive scientists and LLM engineers; they bear directly on how a multilingual society like India should think about deploying LLMs in education and policy. Consider the optimistic case often made for AI tutors in under-resourced schools: a cheap, scalable system that can teach English, or help a child practice Hindi or Tamil, regardless of whether it "understands" anything.

We do not deny that such tools can be useful. We insist, however, that usefulness and pedagogical equivalence are different claims, and conflating them carries costs specific to multilingual contexts.

A child acquiring Tamil at home and English at school is not running two instances of the same statistical learner on two different corpora. Each language is parsed through the same biologically fixed hierarchical architecture, allowing children to effortlessly transfer abstract structural knowledge, such as word-order constraints, agreement patterns, and embedding, across languages they are exposed to, even when the surface vocabulary differs completely.

An LLM-based tutor, by contrast, treats each language as a separate statistical distribution; whatever "transfer" it exhibits is itself a statistical artefact of overlapping training data, not a structural generalisation. For languages with thin digital corpora, which describes most of India's scheduled languages besides Hindi, Bengali, and a handful of others, this means LLM performance will degrade in ways that have nothing to do with the inherent learnability of those languages for a human child. Policymakers should not read "the model struggles with Santali" as "Santali is hard to learn" or "Santali speakers need more remediation".

The same caution applies to the broader claim that it does not matter whether LLMs think, so long as they can solve mathematical problems, debug code, or judge a sentence's grammaticality as well as a trained linguist. A recent evaluation of OpenAI's o3 model found that it fails precisely these structural judgements (Murphy et al. 2025).

Presented with Escher sentences such as "Fewer athletes have been to Beijing than I have", syntactically well-formed but semantically incoherent, the model judged them acceptable, attributing the anomaly to ellipsis; it showed a parallel failure with centre-embedded sentences missing a verb, and could not reliably generate an ungrammatical sentence even when asked to do so directly (Murphy et al. 2025). These are exactly the judgements an AI tutor must make when training a second-language learner, and a system that cannot distinguish grammatical from ungrammatical strings is a poor candidate for this position (Murphy et al. 2025; Dentella et al. 2024).

> Even in mathematics, AI output still needs human expertise to finish, a gap Marcus traces to next-token prediction's lack of algebraic structure, pointing towards neuro-symbolic models instead.

Gary Marcus has long argued that the problem is not limited to Natural Language, and stems from the limits of next-token statistics itself. In October 2025, OpenAI claimed GPT-5 had "found solutions to ten previously unsolved Erd?s problems", a claim retracted after mathematician Thomas Bloom showed the model had merely surfaced existing published solutions, with an OpenAI researcher conceding that "only solutions in the literature were found" (TechCrunch 2025a): a literature search, not novel reasoning.

A later May 2026 claim that an OpenAI model had disproved Erd?s's unit-distance conjecture turned out to have similar caveats, with Bloom writing that the AI's "success here echoes previous achievements: it often produces the most surprising results by persevering down the paths that a human may have dismissed as not worth their time to explore, combining superhuman levels of patience with familiarity with a vast array of technical machinery". Even then, professional mathematicians verified, refined, and completed the proof (The Conversation 2026; TechCrunch 2026).

Even in mathematics, AI output still needs human expertise to finish, a gap Marcus traces to next-token prediction's lack of algebraic structure, pointing towards neuro-symbolic models instead (Marcus 2001, 2024).

So, the next time your phone suggests a word, it is worth appreciating the engineering behind it. But it is not necessary to wonder whether it thinks like you do. Human language begins in noise and ambiguity. It requires discovering words, building hierarchical trees, and tracking discourse across sentences-all under severe informational constraints. It is shaped by a biological symbol-manipulation system that determines not only what can be learned, but also what cannot.

LLMs operate differently. They process pre-structured text, learn from vast datasets, and optimise statistical predictions without an inherent boundary between language and non-language. Language in human beings is not a flat string but an algebraic, symbol-manipulating system, structured, constrained, biologically grounded. Recognising this difference sharpens our understanding of both AI and ourselves.

*Sayantan Mandal is a biolinguist and evolutionary psychologist at Krea University, specializing the evolution of and biological roots of human ability for natural language.*

*Rakesh Sengupta is a cognitive neuroscientist at Krea with expertise in vision, working memory, brain-computer interface devices and deep learning.*

**References:**

Berwick, Robert C., and Noam Chomsky. *Why Only Us: Language and Evolution*. Cambridge, MA: MIT Press, 2016.

Bever, Thomas G. "The Cognitive Basis for Linguistic Structures." *In Cognition and the Development of Language*, edited by John R. Hayes, 279-362. New York: Wiley, 1970.

Brown, Roger. *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press, 1973.

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

---. "Three Factors in Language Design." *Linguistic Inquiry* 36, no. 1 (2005): 1-22.

---. "The False Promise of ChatGPT." *New York Times*, March 8, 2023. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

Cutler, Anne. *Native Listening: Language Experience and the Recognition of Spoken Words*. Cambridge, MA: MIT Press, 2012.

Darwin, Charles. *The Descent of Man, and Selection in Relation to Sex*. Vol. 1. London: John Murray, 1871.

Dentella, Vittoria, Fritz Günther, Elliot Murphy, Gary Marcus, and Evelina Leivada. "Testing AI on Language Comprehension Tasks Reveals Insensitivity to Underlying Meaning." *Scientific Reports 14* (2024): Article 28083. https://doi.org/10.1038/s41598-024-79531-8.

Hauser, Marc D., Noam Chomsky, and W. Tecumseh Fitch. "The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?" *Science 298*, no. 5598 (2002): 1569-79.

Lenneberg, Eric H. *Biological Foundations of Language*. New York: Wiley, 1967.

Marcus, Gary. "Rethinking Eliminative Connectionism." *Cognitive Psychology 37*, no. 3 (1998): 243-82.

---. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press, 2001.

---. *Taming Silicon Valley: How We Can Ensure That AI Works for Us*. Cambridge, MA: MIT Press, 2024.

---. "'Not on the Best Path.'" *Communications of the ACM*, February 12, 2025.

Marcus, Gary, and Ernest Davis. "GPT-3, Bloviator: OpenAI's Language Generator Has No Idea What It's Talking About." *MIT Technology Review*, August 22, 2020.

Miller, George A. *Language and Communication*. New York: McGraw-Hill, 1951.

Mitchell, Melanie, and David C. Krakauer. "The Debate over Understanding in AI's Large Language Models." *Proceedings of the National Academy of Sciences 120*, no. 13 (2023): e2215903120.

Moro, Andrea. *The Boundaries of Babel: The Brain and the Enigma of Impossible Languages*. Cambridge, MA: MIT Press, 2008.

Murphy, Elliot, Evelina Leivada, Vittoria Dentella, Raquel Montero, Fritz Günther, and Gary Marcus. "Fundamental Principles of Linguistic Structure Are Not Represented by ChatGPT." *Biolinguistics 19* (2025): Article e19021. https://doi.org/10.5964/bioling.19021.

Musso, Mariacristina, Andrea Moro, Volkmar Glauche, Michel Rijntjes, Jürgen Reichenbach, Christian Büchel, and Cornelius Weiller. "Broca's Area and the Language Instinct." *Nature Neuroscience 6, no. 7* (2003): 774-81.

Nomura, Yasuyuki. "Verbal Suffixes as Trigger: Constraints for Language Acquisition." *Osaka University Papers in English Linguistics 10* (2006): 1-28.

Pinker, Steven. *The Language Instinct: How the Mind Creates Language*. New York: William Morrow, 1994.

Zou, J., Poeppel, D., & Ding, N. (2026). Constituent-constrained word prediction during language comprehension. *Nature Neuroscience*, 1-12. https://doi.org/10.1038/s41593-026-02272-6

Snyder, William. *Child Language: The Parametric Approach*. Oxford: Oxford University Press, 2007.

---. "A Parametric Approach to the Acquisition of Syntax." *Journal of Child Language 48*, no. 5 (2021): 862-87.

*TechCrunch*. "OpenAI's Embarrassing Math." October 19, 2025. https://techcrunch.com/2025/10/19/openais-embarrassing-math/.

---. "OpenAI Claims It Solved an 80-Year-Old Math Problem-for Real This Time." May 20, 2026. https://techcrunch.com/2026/05/20/openai-claims-it-solved-an-80-year-old-math-problem-for-real-this-time/.

*The Conversation*. "An AI Solution to an 80-Year-Old Problem Has Shocked Mathematicians." May 2026. https://theconversation.com/an-ai-solution-to-an-80-year-old-problem-has-shocked-mathematicians-283686.

Tsimpli, Ianthi-Maria. "Functional Categories and Maturation: The Prefunctional Stage of Language Acquisition." PhD diss., University College London, 1992.

Verkerk, Annemarie, Olena Shcherbakova, Hannah J. Haynie, Hedvig Skirgård, Christoph Rzymski, Quentin D. Atkinson, Simon J. Greenhill, and Russell D. Gray. "Enduring Constraints on Grammar Revealed by Bayesian Spatiophylogenetic Analyses." *Nature Human Behaviour 10*, no. 1 (2025): 126. https://doi.org/10.1038/s41562-025-02325-z.