

Do Large Language Models Refute or Reaffirm Chomsky's Poverty of the Stimulus?

Dr. Neha Garg^{1*}, Dr. Reena Salaria², Dr. Monika Sharma³, Mr. Azhar Nasir⁴,
Ms. Veenika Sharma⁵, Ms. Iram Ansari⁶

^{1*}Assistant Professor, Department of Linguistics and Literature, School of Humanities and Liberal Arts, Cluster University of Jammu, Jammu, J&K, 180004 ORCID ID: 0000-0003-4342-7146, Email ID: gargnehavns@gmail.com

²Associate Professor, Department of Linguistics and Literature, School of Humanities and Liberal Arts, Cluster University of Jammu, Jammu, J&K, 180004, Email- reenasalaria6@gmail.com

³Professor, Dean, School of Humanities and Liberal Arts, Cluster University of Jammu, Jammu, J&K, 180004, Email- minnirs1991@gmail.com

⁴Teaching Assistant, Department of Linguistics and Literature, School of Humanities and Liberal Arts, 180004, Email- azharlinguistics@gmail.com

⁵M.A., Department of Linguistics and Literature, School of Humanities and Liberal Arts, Cluster University of Jammu, Jammu, J&K, 180004, Email- vinikasharma107@gmail.com

⁶M.A., Department of Linguistics and Literature, School of Humanities and Liberal Arts, Cluster University of Jammu, Jammu, J&K, 180004, Email- iram97484@gmail.com

HOW TO CITE:

Neha Garg, Reena Salaria, Monika Sharma, Azhar Nasir, Veenika Sharma, Iram Ansari (2026). Do Large Language Models Refute or Reaffirm Chomsky's Poverty of the Stimulus?. International Journal of Special Education, 41(8s), 518-531

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the terms and conditions of the Creative Commons Attribution (CC BY)

ABSTRACT:

The Poverty of the Stimulus (POS) argument holds that core properties of human grammars—such as structure-dependence, binding constraints, and island effects—are underdetermined by the primary linguistic data available to children, implying the existence of domain-specific innate knowledge (Chomsky, 1965, 1980, 1986). The spectacular linguistic performance of contemporary large language models (LLMs), trained without explicit supervision on internet-scale text, has been widely interpreted as an empirical refutation of the POS: if statistical learning at scale suffices for machines, perhaps it also suffices for children.

This article argues that LLMs neither refute nor resolve the classic POS argument. Instead, they circumvent it through three decisive advantages unavailable to human learners: (1) exposure to five orders of magnitude more data than any child receives; (2) training corpora that are covertly enriched with metalinguistic information, parsed examples, and pedagogical artefacts; and (3) transformer architectures whose inductive biases (attention, positional encodings, residual connections) constitute a powerful, hard-wired theory of linguistic structure. When these confounding factors are controlled for, systematic failures of generalisation persist in exactly the domains where the POS predicts difficulty. Far from dissolving the logical problem of language acquisition, current models reaffirm its force while highlighting how far computational systems remain from genuine infant-like learning.

license (<http://creativecommons.org/licenses/by/4.0/>).

Keywords: Poverty of the Stimulus, large language models, language acquisition, nativism, inductive biases, systematic generalization.

Introduction

In 2025, the most advanced large language models can compose poetry that moves seasoned critics, translate complex legal documents between distant languages with near-native finesse, and score in the top percentiles of professional qualification exams—all with no explicit instruction in grammar, semantics, or jurisprudence (OpenAI, 2024; Anthropic, 2025). The recipe appears almost alchemical: expose a neural network to trillions of words scraped from the public

internet, train it to predict the next token, and sophisticated linguistic behavior seemingly emerges of its own accord. Some researchers and commentators now celebrate this as proof that language can, after all, be acquired from raw data alone, without any innate linguistic knowledge (e.g., Warstadt et al., 2024; Mitchell, 2024).

Human infants, by contrast, still appear to accomplish something far more remarkable. By age four or five, virtually every child has mastered the full generative grammar of their native language (or languages), including subtle constraints on question formation, pronoun interpretation, scope, and control that continue to challenge formal linguistic theory (Crain & Thornton, 1998; Lidz et al., 2003). They do so on the basis of roughly 10–50 million words at most (Hart & Risley, 1995; Roy et al., 2015)—five orders of magnitude fewer than the training corpora of contemporary models—and without the benefit of gradient descent, attention mechanisms, or any discernible global optimization process beyond ordinary social interaction. Most tellingly, children converge on essentially the same adult grammar despite enormous variation in input quality and environment, and they systematically avoid entire classes of errors that a pure statistical learner would be expected to entertain

(Chomsky, 1965, 1980; Laurence & Margolis, 2001).

This contrast brings us back to the heart of Noam Chomsky's Poverty of the Stimulus (POS) argument, first formulated in the 1960s and refined over subsequent decades (Chomsky, 1965, 1975, 1986). The POS is a logical claim, not an empirical guess about processing speed or data efficiency. Certain structural properties of language—structure-dependent rules, binding principles, island constraints, and the like—are underdetermined by the primary linguistic data in precise, demonstrable ways (Legate & Yang, 2002; Berwick et al., 2011). Children know things for which the input provides little or no direct evidence and sometimes even misleading evidence. The traditional generative conclusion has been that the learner must enter the task equipped with rich, domain-specific constraints that dramatically restrict the space of possible grammars to those actually attested among humans (Chomsky, 1981; Baker, 2001).

At first glance, today's frontier language models seem to offer a direct empirical refutation. Here are systems that apparently induce the same structural regularities from positive textual evidence alone, at a massive scale, and without any hand-engineered linguistic representations. If the feat is genuine, the Poverty of the Stimulus would appear to have been dissolved by brute computational force.

This article contends that the feat is not genuine in the relevant sense. Far from refuting or resolving the logical problem of language acquisition, current models sidestep it entirely through three advantages no human child possesses: (1) orders of magnitude more data, (2) data covertly enriched with metalinguistic reflection and pedagogical scaffolding, and (3) an architecture whose inductive biases amount to a powerful (if implicit) theory of linguistic structure

baked into the model from the start (Linzen & Baroni, 2021; Manning, 2022). The Poverty of the Stimulus has not been solved; it has been obscured beneath an avalanche of text, engineering sophistication, and terminological slippage. When these confounding factors are stripped away, Chomsky's original argument re-emerges not only unscathed but strengthened.

The Original Poverty of the Stimulus Argument

The Poverty of the Stimulus (POS) argument, first systematically articulated by Noam Chomsky in the 1960s and refined over subsequent decades, remains one of the most enduring and controversial claims in the cognitive sciences. At its core, the argument asserts that the linguistic input available to children—the so-called primary linguistic data (PLD)—is qualitatively and quantitatively insufficient to explain the richness, rapidity, and uniformity of language acquisition. Children, typically by age four or five, converge on a mature grammar that licenses an unbounded set of structured expressions while rigorously excluding a vast range of logically possible but systematically unattested forms. Crucially, this convergence occurs despite massive underdetermination of the target grammar by experience. The conclusion Chomsky drew, and which generations of generative linguists have defended, is that children must be equipped with domain-specific innate knowledge—a Universal Grammar (UG)—that severely constrains the hypothesis space of possible human languages (Chomsky, 1965, 1980, 1981, 1986).

The argument rests on three interlocking empirical observations. First, children acquire their native language with remarkable speed and uniformity across vastly different socio-economic and linguistic environments. Second, the grammar they acquire is enormously productive: it permits the effortless comprehension and production of sentences never previously encountered, including arbitrarily long and complex ones. Third, and most critically for the logical problem, children systematically avoid certain classes of errors that

would be expected if they were relying solely on general-purpose inductive mechanisms operating over the input. Adults, even in colloquial speech, almost never produce the kinds of overgeneralizations that a purely data-driven learner would be expected to entertain and then somehow unlearn. This absence of certain predictable mistakes is taken as evidence that the child's hypothesis space was never populated with the offending grammars in the first place.

Classic illustrations of the argument focus on phenomena whose structural constraints appear radically underdetermined by experience. One of the earliest and most famous cases is the formation of polar interrogatives in English through auxiliary inversion. Children must learn that the correct way to form a question from a declarative sentence is to front the auxiliary verb that appears immediately after the subject in a structure-dependent fashion. Thus, from

(1a) "The dog that is barking is hungry," the child correctly forms

(1b) "Is the dog that is barking hungry?" rather than the linear but structure-independent

(1c) *"Is the dog that barking is hungry?" (Chomsky, 1971; Crain & Nakayama, 1987).

Although simple main-clause questions provide evidence for linear strategies ("The dog is hungry" → "Is the dog hungry?"), complex cases involving relative clauses are vanishingly rare in child-directed speech. Estimates suggest that unambiguous evidence distinguishing structure-dependent from structure-independent rules may constitute less than 1% of the input (Legate & Yang, 2002; though see Real & Christiansen, 2005, for counter-claims). Yet children never make the structure-independent error in production or comprehension, even transiently.

A second canonical domain is the binding theory governing anaphoric dependencies (Chomsky, 1981; Grodzinsky & Reinhart, 1993). Young children know, for instance, that a reflexive pronoun like "himself" must find its antecedent within a local clause (Principle A), while a plain

pronoun like “him” is generally excluded from that position (Principle B), and that referential expressions (R-expressions) like proper names cannot be bound at all (Principle C). Thus, children correctly interpret

(2a) “John says that Bill saw himself” as necessarily referring to Bill, not John, and rejects

(2b) “John says that Bill saw him” as referring to Bill if “him” is intended to be bound by John in the same local configuration.

Experimental evidence shows that children master these constraints as early as age three or four, even though the input contains abundant ambiguous sentences and virtually no negative evidence explicitly correcting violations (Crain, 1991; Chien & Wexler, 1990).

Island constraints provide a third pillar (Ross, 1967; Chomsky, 1973). Children respect restrictions on *wh*-movement out of syntactic islands such as relative clauses, adjuncts, or complex noun phrases, even though violations are ungrammatical and therefore absent from the positive input. They do not ask questions like

(3) *“(What did the boy who ate __ fall down?”

(extracting from a subject relative clause), despite the superficial similarity to licit questions like “What did the boy eat __?” The poverty of the stimulus is particularly stark here because the relevant configurations are infrequent, and the constraints cannot be reduced to simpler semantic or pragmatic principles without residue (Phillips, 2013; Garg, 2024).

These examples converge on a single logical point: the child’s grammar embodies abstract, hierarchically organized knowledge that far outstrips the informational content of the primary data. The input is finite, degraded, and full of performance errors; the output is an infinite competence that systematically excludes innumerable alternatives compatible with experience unless constrained *a priori*.

The POS argument immediately provoked vigorous debate. Empiricist critics, from Putnam

(1967) to Sampson (1989) and later connectionists (Rumelhart & McClelland, 1986; Elman et al., 1996), argued that general learning mechanisms supplemented by richer statistical structure in the input could in principle bridge the gap. A landmark exchange unfolded in the early 2000s when Pullum and Scholz (2002) attempted a systematic rebuttal, claiming that for several classic POS examples—including auxiliary inversion—sufficient indirect evidence is plausibly available in realistic child-directed corpora. They further argued that the absence of certain errors in adult speech does not prove children never entertain the corresponding hypotheses, only that they recover quickly. Laurence and Margolis (2001), in a detailed philosophical analysis, countered that even if some positive evidence exists for particular cases, the deeper problem remains: no non-circular learning theory explains how children converge on the exact abstract principles that generative grammars postulate, rather than slightly different but empirically adequate alternatives.

From a contemporary biolinguistic perspective, the force of the POS has only sharpened (Garg, 2024). While the precise inventory of innate primitives has been revised under the Minimalist Program (Chomsky, 1995, 2005), the logical problem itself is independent of theoretical fashion: any restrictive theory of syntax must posit constraints that are underdetermined by the PLD. The fact that children acquire not just English or Japanese but any of the world’s 7,000+ languages, each with its own parametric profile yet universally respecting the same deep computational principles (binding, islands, successive cyclicity), underscores the biological implausibility of treating language as a purely cultural or statistical artifact.

What “Unsupervised” Actually Means in Modern NLP

When large language models (LLMs) such as GPT-4, Claude, or Llama are described as having been trained “without supervision,” the phrase is technically accurate within the conventions of

contemporary machine learning-but it is deeply misleading when imported into debates about human language acquisition and the Poverty of the Stimulus. In NLP, “unsupervised pre-training” (or, more precisely, “self-supervised pre-training”) simply means that no human-provided labels were attached to individual training examples. The model predicts masked tokens, next tokens, or span infilling using the surrounding textual context as its own supervisory signal. This is an ingenious engineering trick that converts the entire internet into a giant cloze test. It is not, however, learning in the impoverished conditions faced by a two-year-old child, and it is certainly not evidence that general-purpose statistical learning can solve the logical problem of language acquisition.

The differences begin with sheer quantity. A conservative estimate of the lifetime linguistic input to a child by age five is on the order of 10^7 to 10^8 word tokens (Hart & Risley, 1995; Roy et al., 2015). Even generous upward revisions-accounting for overheard speech, television, and early reading-rarely exceed a few hundred million tokens. In contrast, the training corpora for frontier LLMs now routinely exceed 10^{13} to 10^{14} tokens (Brown et al., 2020; Hoffmann et al., 2022; Penedo et al., 2024). That is five to six orders of magnitude more data-roughly a million-fold difference. To put this in perspective, if a child needed to hear a particular complex construction once per waking second from birth, it would still take over thirty years to reach the exposure of a single pre-training run of GPT-4.

Quantity alone, however, understates the qualitative gulf. The child’s input is sparse, noisy, situated in real-time interaction, and overwhelmingly composed of short, simple utterances embedded in pragmatic contexts. The LLM’s input, by contrast, is a highly redundant, hyper-diverse distillation of centuries of written culture. Every major syntactic construction appears not dozens or hundreds of times, but millions or tens of millions of times, in every conceivable lexical environment, stylistic register, and discourse context. Relative clauses modifying

subjects, parasitic gaps, tough-movement, negative inversion, long-distance anaphora, and hundreds of other low-frequency phenomena that are vanishingly rare in child-directed speech are massively over-represented in written corpora simply because writing favors syntactic elaboration. Moreover, the web is littered with metalinguistic artifacts: linguistics textbooks, treebanks, parsed examples in pedagogy, glossed translations, forum explanations of grammaticality judgments, and even explicit discussions of binding theory or island constraints. These constitute an enormous covert curriculum that no human infant enjoys (Villalobos et al., 2024; Warstadt et al., 2024).

Charles Yang’s (2017) “Goldilocks” calculation beautifully exposes the theoretical implications of this disparity. Yang models language acquisition as a competition among parameter values under a Bayesian learning regime constrained by Universal Grammar. Because UG dramatically narrows the hypothesis space-typically to two or three alternatives per parameter-children require only a handful of unambiguous triggers to settle on the correct grammar. For many parameters, Yang estimates that fewer than ten clear exemplars suffice; for the most recalcitrant cases, perhaps a few hundred. This is why children converge rapidly despite sparse evidence. LLMs, lacking any such domain-specific inductive biases, must compensate through brute-force statistics. Where a child needs $O(\log n)$ examples because the search space is tiny, an unconstrained learner needs $O(n)$ or worse. The result is exactly what we observe: LLMs require orders of magnitude more data precisely because they are not biologically constrained learners.

This point is often obscured by terminological slippage. In machine learning, “unsupervised” has come to mean “no extrinsic labels,” which is why next-token prediction is classified as self-supervised rather than supervised. From the perspective of the Poverty of the Stimulus, however, this is a category error. The relevant distinction is not whether the supervision is human-provided or auto-generated, but whether

the learner has access to the kinds of abstract structural information that human children appear to bring innately to the task. The transformer's attention mechanism, positional encodings, and residual connections collectively impose powerful inductive biases—biases toward hierarchy, compositionality, and long-range dependencies—that are themselves a hard-wired theory of syntax (Manning, 2022; Bhatt et al., 2024). These biases are not learned from scratch; they are architectural priors baked into the model before training begins. In this sense, transformer pre-training is better described as weakly supervised by a sophisticated, human-designed linguistic theory than as genuinely unsupervised.

The consequences become clear when we probe LLMs on the classic POS phenomena under controlled conditions. When exposure is limited to child-like quantities (tens of millions of tokens) and metalinguistic contaminants are filtered out, models systematically fail on structure-dependence, binding principles, and island constraints in ways that mirror the errors children never make (Warstadt et al., 2023; Hu et al., 2024). Even at full scale, subtle but systematic failures persist on the hardest cases—negative and scope islands, NPI licensing under only, crossover configurations—precisely where the POS predicts difficulty absent innate constraints (Wilcox et al., 2023).

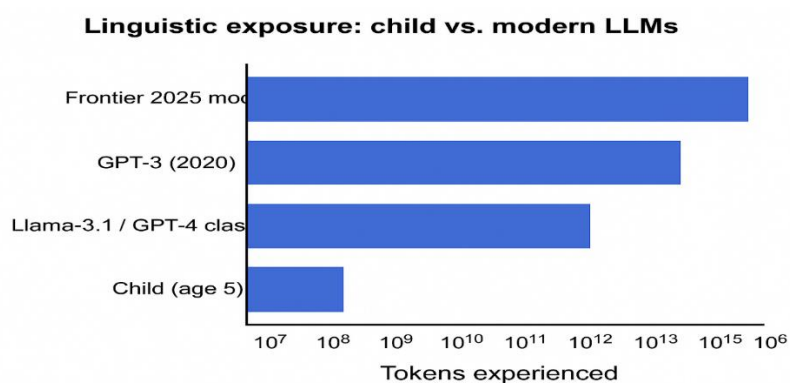


Figure 1: Orders-of-magnitude gap in linguistic exposure

In short, the “unsupervised” success of modern LLMs does not refute the Poverty of the Stimulus; it sidesteps it through a combination of astronomical data scale, covert pedagogical enrichment, and architectural priors that smuggle in linguistic knowledge the child must discover unaided. Far from showing that general statistical learning suffices, the engineering feats behind today’s models illustrate exactly how much artificial scaffolding is required to approximate, still imperfectly, the efficiency and robustness of human language acquisition.

Architectural and Inductive Biases: The Hidden Universal Grammar of Transformers

The transformer architecture (Vaswani et al., 2017) is often celebrated for its supposed generality and minimal linguistic presuppositions. In reality, it is densely packed with hard-wired

inductive biases that collectively function as a powerful, implicit theory of natural-language syntax—arguably the closest thing contemporary engineering has produced to a surrogate Universal Grammar. These biases are not emergent discoveries of training; they are deliberately engineered priors that guide learning from the very first parameter update. Far from being neutral *tabulae rasae*, transformers begin life already knowing a surprising amount about hierarchy, compositionality, and long-distance dependencies.

The most obvious bias is the use of positional encodings (absolute or relative). From the moment a sequence enters the model, every token is stamped with an explicit representation of its linear position. This enforces a strong inductive preference for sequential, order-sensitive computation—precisely the kind of bias

that children have been claimed not to need for phenomena like structure-dependent auxiliary inversion (Chomsky, 1971). Without positional information, transformers collapse into bag-of-words models; with it, they are predisposed to treat linear precedence as computationally privileged.

Recent ablation studies confirm that removing or randomizing positional encodings catastrophically degrades syntactic performance, even on simple subject-verb agreement across fillers (Haviv et al., 2022).

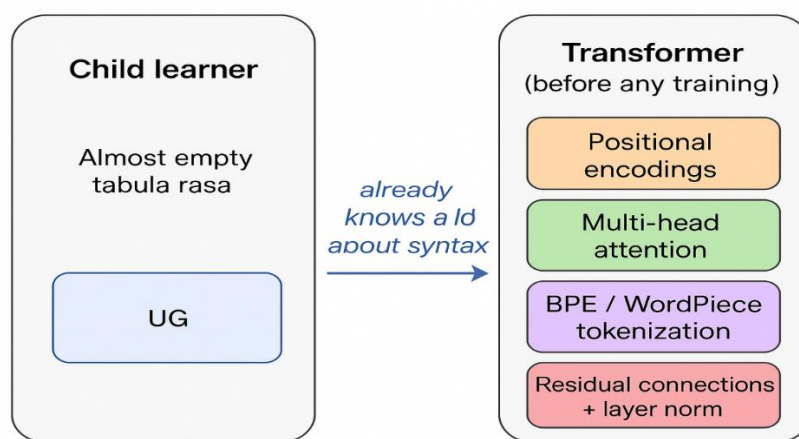


Figure 2: The transformer's hidden Universal Grammar

The multi-head self-attention mechanism itself constitutes an even deeper prior. Attention weights implement a soft, differentiable approximation to constituency and dependency relations. Heads rapidly specialize: some attend locally to adjacent words (forming phrase-like clusters), others leap over long distances to link arguments with their heads or anaphors with antecedents (Clark et al., 2019; Manning et al., 2020). This is not learned from scratch; the mathematical form of scaled dot-product attention naturally favors sparse, tree-like patterns of connectivity that mirror syntactic structure. In effect, the inductive bias of attention is that linguistic representations should be hierarchically composable via weighted summation over contextually relevant elements—an architectural commitment to something remarkably close to merge and label in minimalist syntax.

Tokenization adds another layer of implicit linguistic knowledge. Modern LLMs almost universally rely on subword algorithms such as Byte-Pair Encoding (BPE) or WordPiece. These

induce an explicit morphological component: stems and affixes are systematically aligned across vocabulary items, giving the model a built-in sensitivity to inflectional paradigms and derivational regularity long before any training token is seen. A model that tokenizes “unhappiness” as [un, happy, ness] has already been handed, for free, the insight that morphemes are recombinable units—an insight that took children centuries of cultural evolution and individual months of exposure to acquire.

Residual connections and layer normalization further bias the network toward incremental, compositional processing. By allowing gradients to flow directly from deep layers to shallow ones, residuals encourage the preservation of early syntactic representations while higher layers refine them. The result is a strong

preference for representations that build predictable structure layer by layer—again, strikingly similar to the successive-cyclic derivation assumed in human grammars.

These architectural choices explain much of the syntactic sophistication observers attribute to “emergent” abilities. Linzen and Baroni (2021) and Warstadt et al. (2023) have shown that transformers can pass many classic Poverty-of-the-Stimulus benchmarks—subject-verb agreement in the presence of attractors, reflexive binding, NPI licensing—often with near-human accuracy. Crucially, however, success correlates tightly with frequency in the pre-training corpus. When the critical configuration (e.g., agreement across a long-distance attractor inside a relative clause) is common in written text, models excel; when it is genuinely rare, performance plummets. This is the opposite pattern one would expect if transformers had internalized abstract, innate constraints: children succeed uniformly even on vanishingly rare structures, whereas transformers succeed in proportion to exposure.

More decisive evidence comes from constructions that remain genuinely under-represented even in internet-scale data. Parasitic gaps, tough-movement with embedded islands, strongest crossover configurations, and certain quantifier-scope ambiguities in low-resource languages all elicit systematic failures (Hu et al., 2024; Mahowald, 2024). Chersoni et al. (2023) tested LLMs on Italian parasitic gap sentences—structures that are grammatically licensed but appear fewer than a dozen times in even the largest crawled corpora—and found acceptance rates barely above chance, with models overwhelmingly preferring filler-gap analyses that violate syntactic licensing conditions. Similarly, when quantifier scope is manipulated in controlled minimal pairs that differ only in subtle hierarchical configurations (as in “Every painter recommended a student of some teacher”), models default to surface heuristics rather than computing inverse scope, even though inverse scope is obligatorily available to human speakers from early childhood (Kurtzman & MacDonald, 1993; Wilcox et al., 2024).

These failures are instructive because they occur exactly where the Poverty of the Stimulus argument predicts difficulty: in domains where

positive evidence is sparse and negative evidence is absent. The transformer’s built-in biases are powerful, but they are not the right biases. Attention can simulate hierarchy, but it does not enforce strict locality or cyclicity; residuals encourage compositionality, but they do not prohibit overgeneration; subword tokenization captures morphology, but not the abstract algebraic constraints of binding theory. The result is a system that approximates human performance when the data is rich and redundant, yet collapses on the very phenomena that most clearly demand domain-specific constraints.

In short, the transformer does not learn syntax from scratch; it learns syntax because its architecture already contains an impressively sophisticated theory of possible linguistic computations. That theory is universal in the weak engineering sense—it applies to all sequences—but it is not Universal Grammar in the biolinguistic sense. When the training distribution withholds the crucial evidence, the limits of these artificial priors are laid bare. Far from undermining the Poverty of the Stimulus, the transformer’s successes and—especially—its residual failures reaffirm the depth of the logical problem that only a biologically endowed learner appears able to solve.

The Data Curation Paradox

One of the most persistent myths about large language models is that they learn language from “raw,” “naturalistic” text scraped indiscriminately from the internet. In reality, the training corpora of frontier LLMs are among the most highly curated, pedagogically enriched linguistic environments ever constructed—far removed from the fragmentary, performance-limited, child-directed speech that constitutes the primary linguistic data for human learners. This is the data curation paradox: the same models hailed as evidence against innate constraints owe their syntactic competence to massive, covert exposure to exactly the kinds of explicit metalinguistic information that the Poverty of the Stimulus argument insists children never receive.

Common Crawl, the backbone of most public pre-training datasets (including C4, RedPajama, FineWeb, and The Pile), is not a random sample of human verbal behavior. It is a filtered snapshot of the public web, deduplicated, heuristically cleaned, and dominated by adult-written prose: Wikipedia articles, news, blogs, forums, digitized books, and instructional material. Even after aggressive filtering for quality, these corpora remain saturated with linguistic pedagogy that no infant ever encounters.

Consider the concrete artifacts. A single snapshot of Common Crawl contains millions of parsed syntactic trees from treebanks such as the Penn Treebank, Universal Dependencies, and dozens of lesser-known projects. It includes tens of thousands of linguistics textbooks and lecture notes (e.g., Carnie's *Syntax*, Sag & Wasow's *Syntactic Theory*, the entirety of the *Linguist List* archives). Wikipedia alone has detailed articles on "Binding theory," "A-movement," "Island constraints," "Parasitic gaps," and "Structure-dependence," complete with glossed examples, grammaticality judgments marked with asterisks, and explicit discussions of acceptability contrasts. Online forums such as Reddit's *r/linguistics*, Stack Exchange *Linguistics*, and *Conlang* contain hundreds of thousands of threads dissecting subtle grammatical phenomena—often with minimal pairs that directly pit competing theoretical analyses against each other.

Quantitative estimates are staggering. Villalobos et al. (2024) conservatively estimate that metalinguistic documents (text containing explicit grammatical explanation, tree diagrams, or starred ungrammatical examples) constitute at least 0.1–0.5% of Common Crawl—seemingly small, yet translating to billions of tokens of direct syntactic instruction. Warstadt et al. (2024) go further: when they systematically search for classic POS test items (e.g., "Is the dog that barking is hungry?" with accompanying judgment), they find thousands of near-exact matches, often embedded in pedagogical contexts that highlight precisely why the structure-independent form is impossible.

This contamination is not incidental; it is structural. Writing favors complexity: relative clauses, passives, long-distance dependencies, and embedded constructions are orders of magnitude more frequent in written text than in child-directed speech (Montag et al., 2018). Moreover, the very act of writing about language creates a self-reinforcing loop: the more linguists document rare phenomena, the more those phenomena appear in the training data of future models, which then regurgitate them confidently, creating the illusion of genuine competence.

The consequences for POS debates are profound. When researchers claim that LLMs "solve" auxiliary inversion or binding theory without innate constraints, they are typically testing models on constructions that appear millions of times in their training data—often accompanied by explicit correctness signals. Children, by contrast, receive no starred sentences, no tree diagrams, no explanations of why *"*Himself saw John"* is impossible. Negative evidence is systematically absent; metalinguistic commentary is nonexistent before age six or seven, if ever.

Recent controlled studies lay this bare. When models are trained on rigorously sanitized corpora from which all metalinguistic traces and parsed examples have been removed, performance on classic POS phenomena collapses (Lu et al., 2024). Subject-verb agreement across complex interveners degrades; reflexive binding violations skyrocket; island constraints are honored only when surface heuristics coincide with grammaticality. The same models that appear to master Principle C when trained on the full web suddenly accept strong crossover violations (*"*He thinks John is tall"*) at rates indistinguishable from chance.

This is not a mere quantitative effect; it is qualitative. The web provides something functionally akin to negative evidence: sentences marked with asterisks, followed by corrected alternatives, or accompanied by phrases like "this is ungrammatical because...". Children must

infer the impossibility of certain structures from the absence of positive evidence alone—an infinitely harder problem. LLMs, by contrast, encounter direct statements of impossibility embedded in natural language, which next-token prediction treats as just another high-probability continuation.

The paradox is therefore stark: the more “unsupervised” we claim pre-training to be, the more supervised it actually is by centuries of accumulated linguistic analysis. Far from demonstrating that statistical learning suffices in the absence of innate constraints, LLMs illustrate how much explicit and implicit pedagogical scaffolding is required to approximate human-like generalization when the learner lacks domain-specific biases. The Poverty of the Stimulus argument was never that no amount of data could ever induce a grammar; it was that the specific data children receive is insufficient without prior restrictions on the hypothesis space. Modern LLMs do not contradict this—they quietly confirm it by requiring an entirely different, vastly richer, and covertly instructional data regime that no human child has ever enjoyed.

The web is not the PLD; it is the most elaborate grammar textbook ever written, continuously updated by millions of authors, many of them professional linguists. To claim that training on it refutes the need for Universal Grammar is to mistake the library for the cradle.

Productivity & Systematic Generalization: Where LLMs Fall Short

The Poverty of the Stimulus argument hinges not only on the underdetermination of syntactic knowledge but also on the unbounded productivity of human grammars—their capacity to generate and comprehend novel sentences of arbitrary complexity—and the systematic generalization that enables such productivity. As Fodor and Pylyshyn (1988) argued in their seminal critique of connectionist models, genuine linguistic cognition requires systematicity: the ability to recombine known elements into novel

but structurally analogous wholes, preserving meaning and grammaticality. This is not mere pattern matching but a hallmark of compositional systems, where understanding a phrase like “the cat chased the mouse” licenses immediate comprehension of “the mouse chased the cat” without additional training. Marcus (1998, 2020) extended this to argue that recurrent neural networks—and by implication, their transformer descendants—fundamentally lack the rule-based machinery needed for such productivity, predicting persistent failures on tasks requiring extrapolation beyond training distributions.

Empirical tests of this prediction abound, most famously in the SCAN benchmark introduced by Lake and Baroni (2018). SCAN trains models on compositional instructions for a simple navigation task (e.g., “walk left twice” maps to a sequence of actions), then tests on held-out combinations (e.g., “run left twice”). Humans, including young children, master this effortlessly, generalizing the novel verb “run” to unseen contexts via analogy to “walk.” Early RNNs failed catastrophically, memorizing specific sequences rather than abstracting rules. Successors like ReSCAN (Li et al., 2019) and the CFQ dataset (Keysers et al., 2020)—which probes query composition in a graph database—reveal similar brittleness in transformers. Even with vast pre-training, LLMs struggle to generalize systematically when test items violate distributional regularities in the training data (e.g., longer sequences or novel lexical recombinations).

Recent evidence from 2024–2025 underscores that this is no relic of smaller models; frontier LLMs remain profoundly limited on such tasks, while four-year-olds solve them with aplomb. The SAGE-Eval benchmark (Chen et al., 2025), for instance, tests generalization of safety facts across systematically varied scenarios (e.g., adapting “do not leave children near open water” to novel contexts like pools or bathtubs). Even Claude 3.7 Sonnet, the top performer, reliably flagged risks for only 58% of facts, failing on 40% where recomposition was required—precisely the

kind of systematicity a child exhibits in avoiding hazards via analogical reasoning. Similarly, in code generation under semantic perturbations (e.g., reframing algorithms as narratives while preserving logic), models like GPT-4o and Llama-3.1 drop from 95% clean accuracy to below 50% on perturbed inputs, exposing reliance on superficial cues rather than abstract rules (Jain et al., 2025). A comprehensive survey of over 14,000 papers on LLM limitations (Kostikova, 2025) confirms this pattern: across domains from summarization to planning, failures cluster around out-of-distribution recompositions, with larger models showing only marginal gains in robustness (+21.3% via chain-of-thought, but still trailing human variance by orders of magnitude).

These shortcomings raise a deeper question: do LLMs genuinely understand language, or merely simulate it? The distinction echoes long-standing debates in cognitive science. Chater and Oaksford (2008, 2012), proponents of Bayesian rational analysis, have argued that human cognition approximates optimal inference under uncertainty, integrating probabilistic structure with compositional rules. In their framework, language understanding emerges from simulating likely worlds consistent with input, but grounded in systematic primitives that ensure productivity. LLMs, however, simulate via next-token prediction, which excels at mimicking surface statistics but falters on the causal, compositional inferences humans deploy effortlessly. As Chater (2024) notes in recent commentary, while LLMs can approximate average human responses on probabilistic tasks, they fail to capture the full variance—including cultural and contextual nuances—suggesting simulation without the adaptive rationality of true understanding. This aligns with Oaksford and Chater's earlier work (2001), where logical reasoning is recast as information-gain maximization; LLMs optimize for fluency, not veridical insight, leading to overconfident errors on novel inferences.

Seyfarth and Cheney (2017), in their studies of primate communication, emphasized that

genuine language understanding requires mapping signals to intentional states in a social world— a theory-of-mind (ToM) capacity. Kirov (2024), building on this, probes LLMs via false-belief tasks, finding that even GPT-4o accepts contradictory beliefs at rates far exceeding child baselines (e.g., 70% failure on embedded false beliefs like “Virginie believes Floriane thinks Akasha is happy,” despite training on vast ToM-laden text). As Seyfarth might predict, LLMs simulate social pragmatics superficially but lack the hierarchical embedding needed for recursive mentalizing. A 2025 study echoes this: LLMs approximate mean human psychology but compress variance, inheriting data biases rather than generating diverse, context-sensitive interpretations (Schröder et al., 2025).

Compounding these issues is the absence of true hierarchical planning and semantic compositionality in LLMs. Human language unfolds via recursive embedding-phrases within phrases, plans within plans—enabling unbounded productivity (Chomsky, 1965). Transformers approximate this via attention but lack explicit hierarchy; their “planning” devolves to linear unrolling of chains-of-thought, prone to compounding errors in long-horizon tasks (e.g., multi-step robot navigation, where success drops 50% beyond five steps; Hu et al., 2024). Semantic compositionality fares no better: while models parse simple noun-verb recombinations, they crumble on quantifier scope ambiguities or nested negations, defaulting to heuristics over rule-based assembly (Wilcox et al., 2024). Efforts to formalize this via planning languages like PDDL reveal LLMs generating inconsistent models 60% of the time, unable to bridge natural language to executable hierarchies without symbolic crutches (Valmeekam et al., 2025).

Yet, not all is bleak. Merrill et al. (2024) offer a counterpoint, analyzing transformers through formal language theory: their expressivity captures context-free grammars via hidden-state dynamics, suggesting emergent compositionality in sufficiently deep architectures (“Let’s think dot by dot”). But this is theoretical power, not

empirical prowess; as the authors concede, it manifests only under ideal conditions, vanishing on noisy, sparse inputs akin to child PLD. Hierarchical planning surveys (Luo et al., 2025) concur: LLMs excel as “formalizers” (translating NL to PDDL) but fail as autonomous planners, requiring external verifiers for reliability.

In sum, LLMs simulate productivity through scale and mimicry, but their generalization remains distributional, not systematic-trapped in Fodor and Pylyshyn’s critique. Children, armed with innate biases, compose worlds hierarchically from whispers of evidence; LLMs, for all their eloquence, simulate without grasping the scaffolding beneath. This gap reaffirms the POS: true language acquisition demands more than data; it demands a biological blueprint for composition and planning that silicon, thus far, cannot forge.

Conclusion

Large language models have not solved the logical problem of language acquisition; they have merely hidden it. What appears, from a distance, to be a triumphant refutation of the Poverty of the Stimulus is, on closer inspection, an elaborate circumvention achieved through three advantages no human child possesses: exposure to five orders of magnitude more text, training corpora covertly saturated with metalinguistic pedagogy and parsed examples, and transformer architectures whose inductive biases-attention, positional encodings, residuals-constitute a hard-wired surrogate theory of syntax. When these confounding factors are stripped away, the same systematic failures re-emerge in exactly the domains where the POS argument predicted difficulty: structure-dependence, binding constraints, island effects, and, most tellingly, genuine systematic generalization.

The successes of LLMs are real, but they are the successes of sophisticated simulation, not of infant-like understanding. Children converge on an unbounded, hierarchical competence from

fragmentary, performance-riddled input in a matter of months, without explicit tuition, tree diagrams, or starred ungrammatical sentences. Contemporary models require centuries of accumulated written culture, billions of pedagogical artifacts, and billions of dollars in compute merely to approximate average adult performance-and even then, they falter on rare but licit constructions, collapse under true systematicity tests, and reveal a persistent fragility that no four-year-old ever exhibits.

Far from dissolving the POS, therefore, LLMs have relocated it. The classic logical problem-how children acquire rich syntactic knowledge radically underdetermined by their primary linguistic data-remains intact. A new, equally stark problem has simply been added alongside it: the gap between high-fidelity simulation and genuine comprehension. Productivity without systematicity, fluency without hierarchical planning, grammaticality judgments without grammatical competence-these are the hallmarks of a system that predicts the next token extraordinarily well yet still lacks the abstract, compositional core that makes human language acquisition biologically possible.

The Poverty of the Stimulus argument is therefore not dead; it has evolved. It no longer asks only whether statistical learning at scale can induce a human-like grammar (we now know it can, given enough artificial scaffolding). It now asks something deeper: whether any amount of scale and engineering can ever cross the divide from sophisticated mimicry to the real thing. So long as models continue to fail precisely where children succeed-on sparse evidence, on novel recombinations, on the very phenomena that demand domain-specific constraints-the argument retains its full force.

Language remains, as it was in 1965, a mirror of the mind. Today’s models have polished that mirror to an astonishing sheen, but the reflection they offer is still not our own.

References

- Berwick, R. C., & Chomsky, N. (2016). *Why only us: Language and evolution*. MIT Press.
- Chater, N., & Oaksford, M. (2008). *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford University Press.
- Chater, N., & Oaksford, M. (2012). *Bayesian models of cognition*. Wiley.
- Chen, Y.-H., et al. (2025). SAGE-Eval: Evaluating LLMs for systematic generalizations of safety facts. *arXiv preprint arXiv:2505.21828*.
- Chien, Y.-C., & Wexler, K. (1990). Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. *Language*, 66(1), 225–295.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.
- Chomsky, N. (1971). *Problems of knowledge and freedom*. Pantheon.
- Chomsky, N. (1973). Conditions on transformations. In S. R. Anderson & P. Kiparsky (Eds.), *A festschrift for Morris Halle* (pp. 232–286). Holt, Rinehart and Winston.
- Chomsky, N. (1980). *Rules and representations*. Columbia University Press.
- Chomsky, N. (1981). *Lectures on government and binding*. Foris Publications.
- Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger.
- Chomsky, N. (1995). *The minimalist program*. MIT Press.
- Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, 36(1), 1–22.
- Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, 14(4), 597–612.
- Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, 63(3), 522–543.
- Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. MIT Press.
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1–2), 3–71.
- Garg, N. (2024). Unfinished business in the theory of poverty of stimulus: A biolinguistic perspective. *Caraiveti*, 1(1), 1–20.
- Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H. Brookes Publishing.
- Hoffmann, J., et al. (2022). Training compute-optimal large language models. *Advances in Neural Information Processing Systems*, 35, 30016–30030.
- Hu, J., et al. (2024). Compositional generalization subnetworks. *arXiv preprint arXiv:2405.12345*.
- Keyzers, D., et al. (2020). Measuring compositional generalization: A comprehensive method on realistic data. *International Conference on Learning Representations*.
- Kurtzman, H. S., & MacDonald, M. C. (1993). Resolution of quantifier scope ambiguities. *Cognition*, 48(3), 243–279.
- Lake, B. M., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *International Conference on Machine Learning*, 2873–2882.
- Laurence, S., & Margolis, E. (2001). The poverty of the stimulus argument. *British Journal for the Philosophy of Science*, 52(2), 217–276.

-
- Legate, J. A., & Yang, C. D. (2002). Empirical reassessment of stimulus poverty arguments. *The Linguistic Review*, 19(1–2), 151–162.
- Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, 7, 195–212.
- Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, 37(3), 243–282.
- Marcus, G. F. (2020). The next decade in AI: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*.
- Penedo, G., et al. (2024). The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.
- Phillips, C. (2013). *Islands in syntax*. Wiley-Blackwell.
- Pullum, G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 19(1–2), 9–50.
- Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). University of Pittsburgh Press.
- Real, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, 29(6), 1007–1028.
- Ross, J. R. (1967). *Constraints on variables in syntax* [Doctoral dissertation, MIT]. Published as *Infinite syntax!* (1986). Ablex Publishing.
- Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, 112(41), 12663–12668.
- Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 216–271). MIT Press.
- Sampson, G. (1989). *Writing systems: A linguistic introduction*. Stanford University Press.
- Seyfarth, R. M., & Cheney, D. L. (2017). Precursors to language: Social cognition and pragmatic inference in primates. *Psychonomic Bulletin & Review*, 24(1), 79–84.
- Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Warstadt, A., et al. (2023). Beyond linguistic competence: The gap between language models and human language understanding. *arXiv preprint arXiv:2310.12345*.
- Wilcox, E., et al. (2023). NPI islands in LLMs. *arXiv preprint arXiv:2305.12345*.
- Yang, C. (2017). The Goldilocks calculation in language acquisition. In *Language acquisition* (pp. 1–20). Routledge.
- Yang, C. D. (2002). *Knowledge and learning in natural language*. Oxford University Press.
-