Do Neural Language Models Learn Like Children? Revisiting the Poverty of the Stimulus Hypothesis with PoSH-Bench

Xiulin Yang
Georgetown University
xy236@georgetown.edu

Abstract

The Poverty of the Stimulus (PoS) Hypothesis holds that children acquire complex linguistic knowledge despite receiving limited and ambiguous input, implying innate constraints. Neural language models, which lack such domain-specific priors, offer a computational testbed for this claim. Prior results have been mixed due to variation in data, model architecture, and evaluated phenomena. We present a systematic and developmentally realistic investigation with our new training-and-evaluation suite, **PoSH-Bench**. Transformer models are trained on input reflecting the quantity and quality of children's linguistic experience and evaluated on five phenomena central to acquisition research. We find that models generalize to these phenomena with as little as 10M words but remain less efficient than children. Examining three inductive biases from cognitive and training-dynamics literature, we show that although these biases improve overall linguistic competence, they do not enhance performance on PoS-related phenomena. These findings reveal a dissociation between linguistic competence and structural generalization, indicating that even cognitively grounded inductive biases in neural networks may operate differently from the mechanisms underlying human generalization. PoSH-Bench thus establishes a framework for bridging computational and cognitive approaches to language science.[1]

1 Introduction

Human language acquisition poses a central question: how do learners consistently generalize far beyond their experience? Even toddlers infer abstract structural rules from sparse evidence (Koulaguina and Shi, 2013), highlighting the efficiency of human generalization. Understanding this capacity has long motivated both linguistic theory and computational modeling. One proposal is the *Poverty of the Stimulus* (PoS) hypothesis, which holds that children's reliable acquisition from limited and ambiguous input implies the presence of language-specific innate constraints (Chomsky et al., 1976).

Recent advances in artificial neural networks (ANNs) allow these theoretical claims to be tested under controlled conditions. A growing line of work treats ANNs as cognitive models which train them on human-scale input and evaluate with human-style experimental probes to inform and refine theories of human intelligence (e.g., Frank and Goodman, 2025). This view motivates a central question in both NLP and CogSci: can human-like generalization emerge in language models trained on developmentally plausible input? If weakly biased, domain-general learners acquire key linguistic structures from input comparable in quantity and quality to that available to children, then at least some aspects of linguistic competence may arise without extensive innate constraints.

Despite recent progress, two key gaps remain. **First**, most prior work trains or evaluates models on adult-directed or artificial data (Oba et al., 2024; Patil et al., 2024; Wilcox et al., 2024; McCoy et al., 2020; Ahuja et al., 2025). Such data differ markedly in size, register, and complexity from the input available to children. Studies using child-directed speech (CDS) offer valuable insights (e.g., Huebner et al., 2021) but remain limited in size (typically under 8M words), below children's cumulative early exposure (10M-50M). **Second**, existing syntactic evaluation suites such as BLiMP (Warstadt et al., 2020), Zorro (Huebner et al., 2021), SLOG (Li et al., 2023), and COGS (Kim and Linzen, 2020) probe generalization broadly but not along the canonical PoS dimensions emphasized in acquisition research. The field therefore lacks a unified benchmark spanning multiple PoS phenomena, which would allow systematic comparison across

[1]Models: https://huggingface.co/collections/xiulinyang/posh-bench; Code: https://github.com/xiulinyang/posh_bench.

constructions and between human and model generalization.

To address these issues, we introduce POSH-BENCH (**P**overty **o**f the **S**timulus **H**ypothesis **Bench**mark), a controlled *training-and-evaluation suite* for examining PoS-related learning. The training datasets contain developmentally plausible input (child-directed and other age-appropriate sources) at scales of 10M, 30M, 50M, and 100M words, approximating the linguistic exposure of early learners. The models are evaluated on canonical PoS phenomena, including question formation, anaphoric *one*, binding, *wanna*-contraction, and island constraints. We further manipulate the availability of positive direct evidence (PDE) and incorporate cognitively motivated inductive biases (e.g., recency, hierarchical constraints) to test how these factors shape generalization.

We ask three questions: (1) To what extent can transformer models acquire PoS-related phenomena from developmentally plausible input? (2) How do input type, data scale, and positive direct evidence affect learnability? (3) Do (cognitively grounded) inductive biases facilitate or constrain such learning?

Our findings show that transformer models can partially acquire PoS-related phenomena even without positive direct evidence, achieving above-chance generalization from as little as 10M words (approximately the lower bound of linguistic input experienced by 3–5-year-old children). Increasing data size improves performance, though the gains are shallower than those observed in human learners. Models trained on simpler, speech-like input outperform those trained on more complex text, and cognitively inspired inductive biases enhance overall linguistic competence but hinder PoS-related generalizations. Together, our approach provides a framework for evaluating structural generalization under developmentally realistic conditions, advancing empirical inquiry toward a closer alignment between human and neural learners.

2 Background: The Learnability Puzzle

2.1 The Heart of the Learnability Debate

Children acquire complex linguistic knowledge from limited data that is, in principle, compatible with multiple hypotheses in the learner's hypothesis space. Yet, they consistently and rapidly converge on the correct generalization.[2] Two main theoretical approaches aim to explain this success. Following Pearl (2022), we refer to these approaches as *linguistic nativism* and *non-linguistic nativism*.[3] The debate centers on two interrelated questions: (i) whether the inductive biases that guide acquisition are domain-specific or cognitively general, and (ii) whether the input available to children is truly impoverished or sufficiently informative when processed by powerful learning mechanisms.

Linguistic nativism holds that these inductive biases are domain-specific. It distinguishes four kinds of evidence based on two dimensions: positive vs. negative and direct vs. indirect (Pearl, 2022). *Positive evidence* signals which forms are grammatical, while *negative evidence* signals which are not. *Direct evidence* explicitly targets the correct hypothesis, whereas *indirect evidence* requires inference from context or co-occurrence patterns. Because negative evidence (e.g., explicit correction) and positive direct cues (typically complex sentences) are often ignored (e.g., Brown, 1970; Bowerman, 1988) and rare (Legate and Yang, 2002; Lidz et al., 2003), learners rely primarily on abundant yet indirect positive input, which may be compatible with multiple hypotheses, giving rise to the so-called "poverty" of the stimulus.

Non-linguistic nativism, in contrast, argues that the input is sufficiently rich when processed by powerful, domain-general mechanisms such as distributional learning, analogy, and pragmatic inference (Ambridge and Lieven, 2011). Under this view, linguistic structure emerges from the exploitation of regularities in the environment rather than from language-specific innate constraints. This theoretical divide motivates the computational inquiry: by implementing learners with different inductive assumptions, we can test whether domain-general mechanisms suffice to recover the linguistic generalizations observed in human learners.

2.2 Formalizing the Poverty of the Stimulus Hypothesis

Following the linguistic nativist tradition, the PoS hypothesis Chomsky 1965 can be formally defined as below, drawing on Pearl (2022); Pullum and

[2] *A person might go through much or all of his life without ever having been exposed to relevant evidence, but he will nevertheless unerringly employ [the structure-dependent generalization] on the first relevant occasion.* (Piattelli-Palmarini, 1980)

[3] Often also referred to as empiricism or usage-based learning.

The Poverty of the Stimulus Argument

(1) Goal: Learners L aim to acquire the target linguistic knowledge T by identifying the correct generalization h_0 from a hypothesis space

$$H = \{h_0, h_1, \ldots, h_n\}.$$

(2) Constraint: The linguistic input available to learners is finite, noisy, and sometimes misleading or ambiguous (Pearl, 2022), providing insufficient evidence to eliminate most competing hypotheses in H.
(3) Observation: Despite this underdetermination, learners consistently and reliably acquire h_0.
(4) Inference: Therefore, language acquisition cannot proceed purely in a data-driven manner.
(5) Conclusion: Learners must possess innate biases (e.g., Universal Grammar) that guide them systematically toward h_0.

Scholz (2002); Perfors et al. (2011).

3 PoS-related Phenomena Studied

This section introduces the linguistic phenomena covered in POSH-BENCH evaluation set (Table 1) and the linguistic principles each phenomenon tests. For each case, we briefly describe its core syntactic properties, discuss why direct evidence for the relevant generalization is scarce in naturalistic input, and summarize empirical findings from both human learners and computational models.

Yes/No Question Formation Yes/no question formation is a classic PoS case concerns how children learn to form yes/no questions by moving the main auxiliary rather than the first one (Chomsky, 1965; Pullum and Scholz, 2002). (e.g., Chomsky, 1965; Lightfoot, 1991; Pullum and Scholz, 2002). To learn the correct hypothesis, children would need sentences containing multiple auxiliaries with the main auxilary following the subordinate auxilary, which are extremely rare in CDS (less than 0.1%; Legate and Yang, 2002). Nevertheless, children older than 4;7 consistently apply the hierarchical rule (Crain and Nakayama, 1987). By contrast, computational models show mixed results: simple n-gram learners capture surface patterns but not hierarchical generalization (Reali and Christiansen, 2004; Kam et al., 2008); Bayesian models succeed only in idealized hypothesis spaces (Perfors et al., 2011); and neural networks trained on CDS or artificial data often fail without strong inductive biases (McCoy et al., 2020; Yedetore et al., 2023; Ahuja et al., 2025; Murty et al., 2023; Qin et al., 2024). In the benchmark, we test three types of relative clauses being the subordinate clause: subject, object, and reduced relative clause.

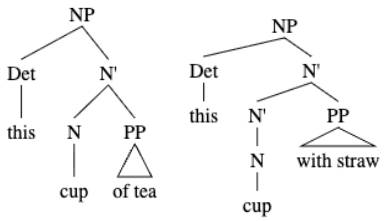

Figure 1: Syntactic trees for *this cup of tea* (left) and *this cup with straw* (right), illustrating that *one* replaces the higher N′ constituent rather than the head noun.

Island Constraints Island constraints (Ross, 1967) restrict syntactic movement, preventing constituents from being extracted out of certain structural domains such as complex noun phrases, wh-clauses, and adjunct clauses. They are classic PoS cases because their acquisition requires *negative evidence*: children must infer that certain movements are ungrammatical despite the lack of explicit correction. Empirical studies show that children acquire sensitivity to island constraints by around age 3 (De Villiers et al., 1990; De Villiers and Roeper, 1995; Goodluck et al., 1992; Hirzel, 2022), demonstrating early mastery of hierarchical dependencies. From a computational perspective, island constraints remain challenging. Pretrained language models capture some basic island effects (Chowdhury and Zamparelli, 2018; Wilcox et al., 2024; Howitt et al., 2024) but fail in more complex configurations such as parasitic gaps and across-the-board movement (Lan et al., 2024). Recent BabyLM-scale experiments (Chang et al., 2025) further show that models trained on 10–100M words of developmentally plausible data fail to acquire wh- and adjunct islands. In our benchmark, we include three major subcategories, Complex NP, Wh-, and Adjunct Islands, to assess whether models can generalize hierarchical movement restrictions under realistic input conditions.

Anaphoric *One* Anaphoric *one* (Baker, 1978) refers to an anaphor that substitutes for an N′ constituent rather than merely the head noun. For example, *one* can replace a modified noun phrase (*this cup with a straw* → *this one with a straw*) but not a complement (*this cup of tea* → **this one of tea*) (see Figure 1). Although direct evidence for this distinction is vanishingly rare in CDS (less than 0.002%), 18-month-old infants show sensi-

Phenomenon	#Pos.DE	Age of Acquisition	Cats.	Minimal Pairs
Question Formation	<0.2% (Legate and Yang, 2002)	3;2–5;11 (Crain and Nakayama, 1987); 3–5 (Nakayama, 1987)	SubjR	Will the man who *did* read the book __ leave? *Did the man who __ read the table *will* leave?
			ObjR	Will the man who the boy *did* see __ leave? *Did the man who the boy __ see *will* leave?
			Reduced ObjR	Can the man the boy *did* see __ explain? *Did the man the boy __ see *can* explain?
Island Constraints	NA	3 (De Villiers et al., 1990; Hirzel, 2022; Goodluck et al., 1992); 3;1–6;1 (De Villiers and Roeper, 1995); 4–5 (Fetters and Lidz, 2017); after 6 (Dąbrowska et al., 2009)	Complex NP	What did she [$_{VP}$ claim that he likes __]? ✓ *What did she make [$_{NP}$ the claim that he likes __]? ✗
			Wh	What did she think [$_{THAT}$ that he likes __]? ✓ *What did she wonder [$_{WH}$ whether he likes __]? ✗
			Adjunct	What did she think [$_{CP}$ that he likes __]? ✓ *What did she worry [$_{ADV}$ if he likes __]? ✗
Anaphoric *one*	0.02% (Lidz et al., 2003)	1;6 (Lidz et al., 2003)	Syntactic	I see a piece of pizza. Did you see another one? *I see a piece of pizza. Did you see another one of pizza?
Binding	0.067%	3–5 (Crain and McKee, 1985); 4–6;6 (Chien and Wexler, 1990)	Principle-A-L	The boy said the girl likes herself. *The boy said the girl likes himself.
			Principle-A-C	The boy said the girl's dad likes himself. *The boy said the girl's dad likes herself.
Wanna	NA	3;11 (Hwang, 2024)	Wanna	When do you wanna go? *Who do you wanna go?

Table 1: Phenomena included in the POSH-BENCH. # Pos.DE refers to the stats for the amount of positive direct evidence. Age of Acquisition lists exisiting studies that confirm children's acquisition within a age range. The benchmark covers 5 phenomena with 10 subcategories in total. We show one example for each subcategory.

tivity to the hierarchical rule (Lidz et al., 2003). Bayesian models reproduce this behavior under idealized assumptions (Regier and Gahl, 2004; Foraker et al., 2009; Pearl and Mis, 2011, 2016), making anaphoric *one* a paradigm case where abstract structure emerges from impoverished data.

Binding Under Principle A of the Binding Theory (Chomsky, 1993), an anaphor must be *c-commanded* and co-indexed by its antecedent within a local domain. To identify this rule, children need positive direct evidence that contains a reflexive pronoun and has more than one preceding antecedent. Such configurations are extremely rare in CDS (approximately 0.07%; see Appendix B). Behavioral studies show that by age 6, children achieve near-adult accuracy (Chien and Wexler, 1990), with even earlier mastery in other languages (McKee, 1992; Emond and Shi, 2025). Although binding is often included as part of broader syntactic evaluation benchmarks, there has been little work testing whether models can generalize this dependency directly. We include two subtypes of Principle A (locality and c-command) to assess sensitivity to structural binding constraints.

***Wanna*-Contraction** In English, *want to* can contract to *wanna*, except when the subject of the infinitival clause has been extracted. This subtle restriction highlights hierarchical dependencies between movement and contraction. Children show adult-like acceptability by 3;11 (Hwang, 2024), though production errors persist (Getz, 2019; Zukowski and Larsen, 2011). To our knowledge, no computational studies have directly modeled this alternation. We include the *wanna* construction in our benchmark as a minimal-pair test of structural abstraction in contraction.

Summary Across these five categories, children demonstrate early mastery despite minimal/absent positive evidence, whereas current computational models show limited or inconsistent generalization.

4 PoSH-Bench: A Training and Evaluation Suite

4.1 Training Data

POSH-BENCH provides developmentally realistic training input that approximates both the quantity and diversity of linguistic experience available to children. Empirical estimates of children's cumu-

Data Source	10M	30M	50M	100M
Speech Transcriptions				
CHILDES	4M	7M	9M	9M
OpenSubtitles	2M	9M	20M	20M
BNC	1M	5M	8M	8M
Switchboard	0.5M	1M	1M	1M
Subtotal	7.5M	22M	38M	38M
Written Texts				
TinyStories	1M	3M	4M	22M
Project Gutenberg	1M	3M	4M	26M
Simple English Wikipedia	0.5M	2M	4M	14M
Subtotal	2.5M	8M	12M	62M

Table 2: Word counts (in millions) by data source and total target size for each training dataset.

Phenomenon	Existing Benchmarks
Question Formation	McCoy et al. (2020); Yedetore et al. (2023)
Anaphoric *one*	None
Island Constraints	Warstadt et al. (2020); Huebner et al. (2021); Wilcox et al. (2018); McCoy and Griffiths (2025)
Binding	Warstadt et al. (2020); Huebner et al. (2021); McCoy and Griffiths (2025)
Wanna	None

Table 3: Existing benchmarks that contain the phenomena of interest

lative exposure range from 2–60M words by age five (Hart and Risley, 1992; Gilkerson et al., 2017; Frank, 2023). Balancing these findings with computational feasibility, we construct three developmentally motivated scales - 10M (lower bound), 30M (midrange), and 50M (upper bound) - plus a 100M-word extension representing early adolescence (MacWhinney, 2000).

Children's linguistic input is not limited to CDS: they also overhear conversations (Casillas et al., 2020; Cristia et al., 2019; Floor and Akhtar, 2006; Akhtar, 2005), television programs (Linebarger and Walker, 2005), and shared book reading (Montag et al., 2015). Therefore, we sample a broad range of speech transcriptions rather than restricting to strictly child-directed utterances. Speech data makes up roughly 76% of the 50M dataset (BABY-50M), a proportion kept constant for the 10M and 30M splits for fair comparisons. The remaining portion draws from simplified written sources (e.g., Simple English and TinyStories (Eldan and Li, 2023)) to approximate shared reading and media exposure.[4]

Building on this base corpus (see Table 2), we design three controlled variants to probe the effects of input register and the availability of positive direct evidence. These variants allow us to separate the influence of data source (WIKI) and the absence of direct evidence (BABY-F).

WIKI A Wikipedia-based version of comparable size provides an adult-oriented baseline with denser syntax and more abstract content. Detailed statistics are provided in Appendix A.

[4]The 100M split is based on the BabyLM 2025 corpus (Charpentier et al., 2025), excluding children's own utterances in CHILDES and replacing them with TinyStories.

BABY-F A filtered version of BABY that removes all PDE sentences, simulating a learner exposed only to indirect evidence. The 100M corpus remains unfiltered since exposure at this scale surpasses early developmental input ranges. The filtering details can be found in Appendix B.

BABY Similar to BABY-F, but with a small proportion of PDE sentences (approximately 0.2% complex questions and 0.07% binding) reintroduced by randomly replacing existing sentences. This reflects the presence of rare yet potentially informative examples in children's natural linguistic input.

4.2 Evaluation Suite

For the target phenomena, several existing benchmarks partially overlap with our focus (Table 3). However, POSH-BENCH differs in explicitly targeting syntactic constructions central to the PoS debate, providing a theoretically motivated and cognitively grounded evaluation of language models. The benchmark comprises five major categories and ten subcategories drawn from the acquisition literature. Each subcategory includes 500 manually verified minimal pairs designed to ensure both syntactic contrast and semantic plausibility. To enable fair comparison across input conditions (WIKI vs. BABY), all lexical items are sampled from the intersection of the top 5K most frequent words shared between the WIKI-100M and BABY-100M corpora.

5 Method: Training and Evaluation

5.1 Models and Training

All models follow the GPT-2 architecture (Radford et al., 2019). For each dataset split, we train a separate tokenizer to account for register and vocabulary differences between WIKI and BABY. Although the WIKI-100M and BABY-100M corpora contain a comparable number of words, the result-

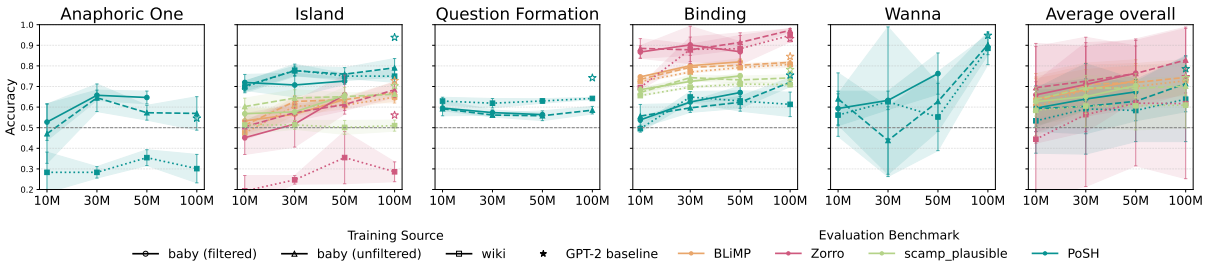

Figure 2: Poverty of the Stimulus Category-wise average for each benchmark. The long dashed line represent chance level performance and the star represent the performance of pretrained GPT2-small. The shadow/error bar represent SD across 3 random seeds.

ing token counts vary substantially with different vocabulary sizes. We experiment with five common vocabulary sizes and select 32,768, which minimizes the discrepancy in compression rate across datasets (See Table 10 in Appendix C).

A pilot study comparing model capacities shows that `gpt2-mini` performs best on the 10M-scale data, while `gpt2-small` achieves lower loss for larger datasets. Accordingly, we adopt `gpt2-mini` for 10M experiments and `gpt2-small` for 30M, 50M, and 100M. Each model is trained with three random seeds for robustness.

Training proceeds for up to 100k steps with early stopping (patience = 6k) and a 4k-step linear warmup following Padovani et al. (2025) to mitigate early overfitting. Optimization uses AdamW with a learning rate of 1×10^{-4} and weight decay 0.01. Full hyperparameter and model architecture details are provided in Appendix D.

5.2 Evaluation

Each phenomenon is tested using minimal pairs constructed for the benchmark. Following standard syntactic evaluation protocols (Warstadt et al., 2020; Huebner et al., 2021), for every pair, we compute the average token-level perplexity of both sentences and count an item as correct if the grammatical sentence receives lower perplexity. Accuracy (proportion of correct preferences) serves as the primary evaluation metric, averaged across three seeds.

6 Experiment 1: Structural Generalization without Positive Evidence

This experiment tests whether transformer models can acquire the target PoS-related phenomena from input that lacks all PDE. We train models of different sizes on the BABY-F datasets and evaluate them on PoSH-BENCH. All reported results are averaged over three random seeds. Category-level scores are shown in Table 4, with fine-grained subcategory results provided in Table 8.

Size	Cat.	AnaOne	Island	QF	Binding	Wanna
10M	BABY	47.1	70.3	59.2	55.6	64.0
	BABY-F	52.7	72.0	59.5	53.8	59.3
	WIKI	28.3	69.7	62.9	49.5	56.1
30M	BABY	64.5	77.6	56.1	59.6	43.8
	BABY-F	65.7	70.7	57.3	62.5	63.3
	WIKI	28.3	77.9	61.9	64.9	62.6
50M	BABY	57.3	76.0	55.8	62.3	62.7
	BABY-F	64.7	72.6	56.3	67.1	76.3
	WIKI	35.5	75.1	63.0	63.0	55.2
100M	BABY	56.9	79.1	58.5	72.1	90.3
	WIKI	26.4	73.8	64.1	57.9	88.9

Table 4: Result overview with color-coded accuracy (red: low, green: high).

Models trained on BABY-F consistently exceeded chance-level accuracy even with only 10M words of training data. Performance varied across phenomena: *Islands* show the strongest generalization, whereas *Anaphoric one* remains the most difficult (52.7% accuracy). These results suggest that, despite the absence of PDE, transformer models can extract some hierarchical regularities from purely indirect evidence, though for some phenomena, the accuracy is still relatively low.

7 Experiment 2: Effects of Input Type, Data Scale, and Evidence Availability

The second experiment examines how input quantity, register, and PDE availability jointly affect structural generalization. We train additional models on the BABY (unfiltered) and WIKI corpora to address three questions: (1) Does increasing input size improve generalization and how much is the improvement? (2) Does exposure to more complex, adult-oriented text (WIKI) yield better syntactic ab-

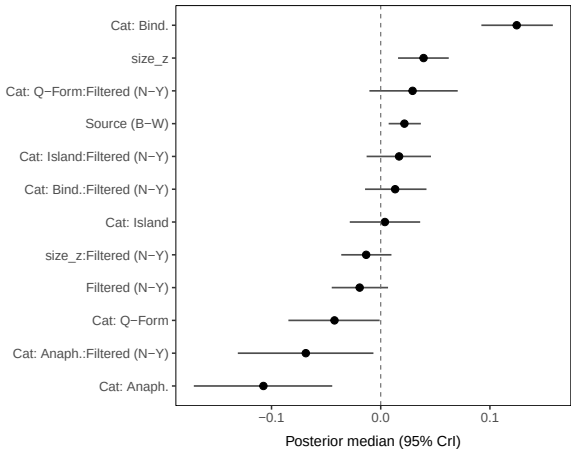

Figure 3: Posterior medians and 95% credible intervals from the Bayesian mixed model. N:not-filtered; Y:filtered; B:baby, W: wiki

straction? (3) Does introducing a small proportion of PDE ($<0.3\%$) enhance generalization?

To ensure that our conclusions are reliable and not benchmark-specific, we evaluate each model not only on POSH-BENCH but also on overlapping phenomena from existing benchmarks—BLiMP (Warstadt et al., 2020), Zorro (Huebner et al., 2021), and SCaMP (McCoy and Griffiths, 2025). This cross-benchmark evaluation provides a consistency check across independently constructed test suites, reducing the likelihood that observed effects reflect idiosyncrasies of a single dataset.

We analyze results using a Bayesian linear mixed-effects model: `acc` $\sim$ (`size` + `category`) $*$ `filtered` + `training_source` + (1|`benchmark`) All predictors are sum-coded. Sampling used four chains of 4,000 iterations each (1,000 warmup), yielding 12,000 posterior draws.[5] Posterior medians and 95% credible intervals (CrI) are visualized in Figure 3, and detailed estimates appear in Table 7.

Main effects Larger training size reliably improved overall performance (median = 0.04, 95% CrI [0.02, 0.06]). The main effect of filtering overlapped zero (–0.02 [–0.04, 0.01]), indicating no general benefit from the inclusion of PDE. Models trained on BABY input performed slightly better than those trained on WIKI (0.02 [0.01, 0.04]). Performance varied across linguistic categories: *Binding* yielded the highest relative accuracy, whereas *Anaphoric one* and *Question formation* were lower.

Interaction effects. A credible interaction was found between Anaphoric one and filtering (–0.07 [–0.13, –0.01]), indicating that models trained *without* positive direct evidence (i.e., in the filtered condition) related to question formation and binding performs better on this phenomenon, which is surprising. For Question Formation and Binding, the interaction terms showed positive posterior medians, which suggests an opposite trend, but their 95% credible intervals included zero, indicating that these effects were not credible.

Summary Taken together, these findings suggest that increasing data size and training on simpler, child-oriented text enhance structural generalization.[6] The models may gain a small benefit from limited positive direct evidence, though this effect is not statistically credible. Future work could further investigate how much and what kinds of positive direct evidence are necessary for models to achieve more human-like learning efficiency (cf. Oba et al., 2024).

8 Experiment 3: Inductive Biases – Helpful or Not?

The final experiment investigates whether inductive biases, either implicit in training dynamics or explicitly motivated by cognitive theories, enhance structural generalization under impoverished input. We test three types of biases: one implicit bias arising from extended training (*grokking*), and two cognitively plausible biases reflecting hierarchical abstraction and limited working memory. All experiments use the BABY-F 10M dataset unless otherwise noted.

3.1 Implicit Bias from Prolonged Training

Murty et al. (2023); Power et al. (2022) report that moderately deep transformers (4–6 layers) can exhibit *grokking* which is a late-emerging generalization phase after overfitting when trained for very long durations. This behavior has been interpreted as an implicit bias resulting from the dynamics of optimization. While Chang et al. (2025) found little benefit with `gpt2-small` when testing on island constraints, we follow Murty et al. (2023) test a shallower 6-layer, 4-head configuration (`gpt2-xxs`) trained for 300k steps, compared to an early-stopped baseline at 50k.

[5] Convergence diagnostics were satisfactory ($\hat{R} = 1.00$), with large effective sample sizes and only two isolated divergent transitions.

[6] However, it is also possible that the observed benefit arises from the diversity of genres, as suggested by Feng et al. (2024).

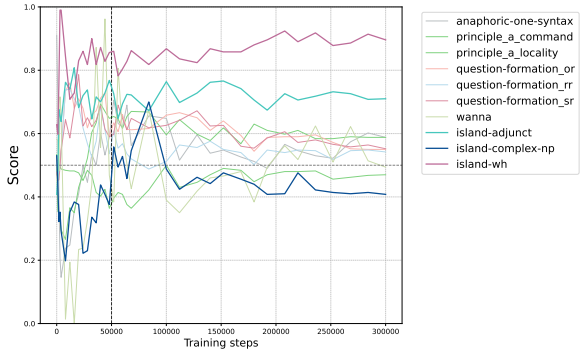

Figure 4: Learning Trajectories of Different Phenomena. Step 50k: the early stop checkpoint; Step 300k: the checkpoint long after early stop

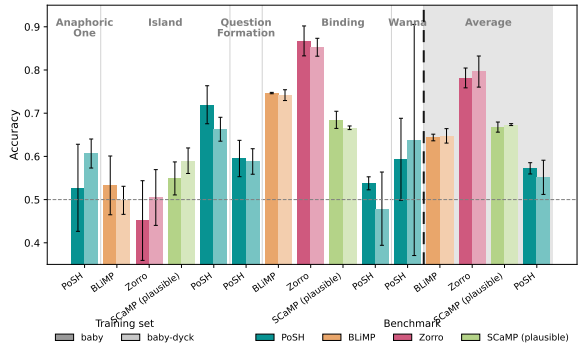

Figure 5: Results of pre-pretraining vs pretraining settings; Error bars indicates the SD

As shown in Figure 4, extended training generally reduced performance rather than improving it. Unlike most PoSH-Bench phenomena, whose accuracies declined stably after 50k steps, island constraints were less affected. A modest improvement was observed for Complex-NP islands ($+20\%$ between 50k and 80k), and wh-islands also showed slight increases even after overfitting. However, as this experiment was conducted with a single seed, these results should be interpreted with caution.

3.2 Hierarchical Bias via Pre-Pretraining. To test whether an explicitly hierarchical inductive bias improves syntactic abstraction, we follow Hu et al. (2025) and pre-pretrain `gpt2-mini` on a shuffled k-Dyck language for 2k steps before training on BABY-F. This manipulation provides models with prior experience in recursive hierarchical structure, an inductive bias that is cognitively plausible and supported by PoS hypothesis and syntactic bootstrapping theories (Fisher et al., 2010).

Results in Figure 5 show that, while pre-pretraining improves average accuracy on general benchmarks (BLiMP, Zorro, SCaMP), it does not transfer to PoSH-Bench. A closer analysis reveals that only *Anaphoric one* benefits modestly and *Wanna* does too but with relativly less reliable improvement, whereas *Island*, *Binding* and *Question formation* decline. In short, although hierarchical pre-pretraining enhances linguistic competence overall, it fails to promote human-like structural generalization and may even hinder it in some cases.

3.3 Dynamic Recency Bias from Limited Working Memory. Finally, we test a cognitively motivated bias derived from the *Less-is-More* Hypothesis (Newport, 1990), which posits that children's limited working memory facilitates abstraction during early learning. Following Mita et al. (2025), we implement a dynamic attention bias based on ALiBi (Press et al., 2021), where the recency weight decays over training epochs:

$$\text{Attention Score} = \text{softmax}\Big(q_i K^\top + r^t \cdot B\Big), \quad (1)$$

$$B = \big[-(i-1) \quad -(i-2) \quad \cdots \quad 0\big]. \quad (2)$$

where $q_i K^\top$ represents the conventional attention score and $r^t \cdot B$ represents a dynamic update of a recency bias as the number of epoch t increases. The update rate is determined by the decay rate r.

We train `gpt2-mini` on BABY-F 10M for 20 epochs, setting $r = 0.6$ and the recency bias vanishes around epoch 10.

As shown in Figure 6 (bottom), models with dynamic recency bias achieve higher average accuracy than those without, particularly on general benchmarks after epoch 5. However, this advantage does not extend to PoS-related phenomena: on PoSH-Bench, models trained without dynamic recency perform better. After zooming into specific PoS-related phenomena (Figure 6, top), we find that despite overall gain in these benchmarks, adding dynamic recency biases hurts on these PoS-related phenomena across all benchmarks.

9 Discussion

Returning to our central question: *to what extent can artificial neural networks reproduce human-like structural generalization when trained on developmentally plausible input?*, our findings suggest a nuanced answer. With approximately the linguistic input available to a 3–5-year-old child (around 10M words), transformer models achieve

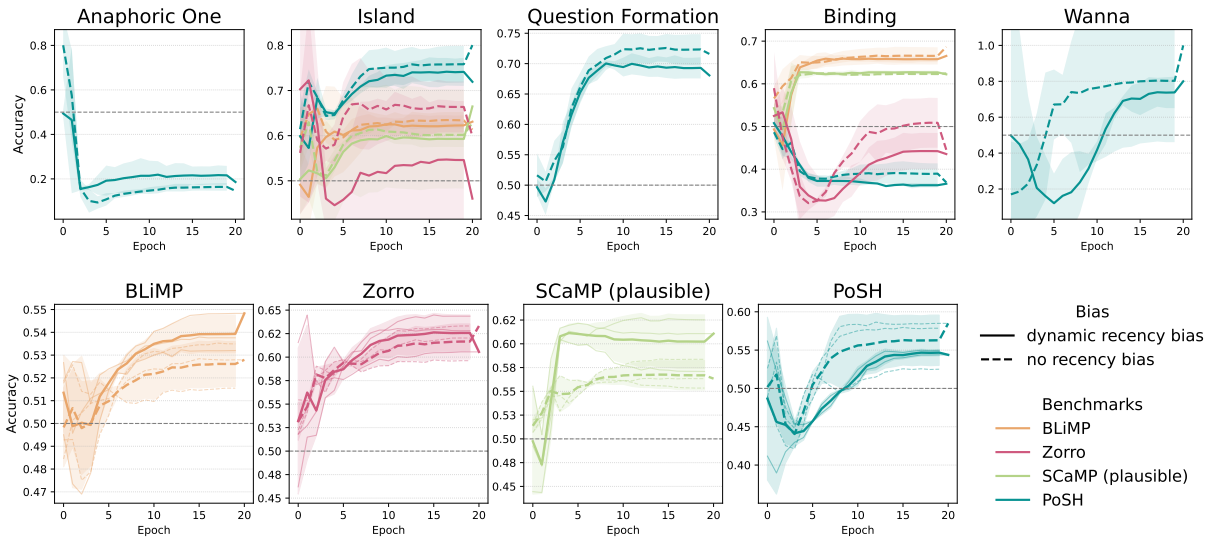

Figure 6: Results of GPT-2 mini with and without recency bias. Top: per-phenomenon performance across different benchmarks. Bottom: average accuracy across categories for each benchmark. Shaded areas denote ± 1 SD across random seeds.

non-trivial but limited learning of PoS-related phenomena. In some cases, such as Binding, their accuracy roughly matches that of children around age four: Chien and Wexler (1990) report that children begin to exceed chance on reflexive interpretation only after age four, similar to our models at 10M words. However, unlike human learners, who rapidly approach ceiling performance by age six, models show much slower progress: even with 100M words, which is roughly twice the estimated input available by age six, their accuracy plateaus around 0.7.

As Samet (2008) noted, *although [innateness] is an easy doctrine to attack, it is a hard one to kill.* Our goal is not to refute the PoS Hypothesis but to refine its scope. Specifically, our results challenge the assumption that the input available to children is as impoverished as traditionally claimed. Developmentally realistic data alone enables above-chance structural generalization in neural learners. At the same time, the persistent gap between model and human learning efficiency indicates that richer inductive biases, or broader multimodal experience, remain necessary for robust generalization.

Experiment 3 further demonstrates that introducing inductive biases improves models' overall linguistic competence but not their structural generalization on PoS-related phenomena. This dissociation raises three possibilities. First, cognitively grounded biases may aid human learners but fail to translate to current ANNs, suggesting that (i) data-driven architectures like transformers might have already reached their ceiling in exploiting distributional cues; or (ii) the learning mechanisms are so different in ANNs from humans that such biases are not testable. Second, such biases might benefit both humans and models, but only under specific architectural or developmental constraints (e.g., memory capacity, data size, data content, curriculum structure) not yet explored here. Third, some hypothesized biases may be mis-specified or incomplete, underscoring the need for more precise computational formulations of human learning constraints. Distinguishing among these possibilities will be an important direction for future research.

10 Conclusion and Future Work

Language acquisition involves the interaction between the learner, its inductive biases, and the input environment (Warstadt and Bowman, 2022). In this work, we revisited the Poverty of the Stimulus Hypothesis through the lens of modern neural language models and a developmentally grounded benchmark, PoSH-Bench. Across three experiments, we find that transformers trained on child-scale data partially acquire core linguistic generalizations but remain less data-efficient than human learners; that additional exposure to direct evidence or adult-oriented text does not substantially improve performance; and that cognitively motivated inductive biases enhance surface competence but not structural generalization.

Together, these findings suggest that develop-

mentally plausible input is more informative than previously assumed, yet current inductive biases, whether implicit or cognitively inspired, do not fully capture the mechanisms underlying human-like generalization. Future work will extend POSH-BENCH to multilingual and multimodal settings, integrate psycholinguistic developmental data, and explore more cognitively plausible hypotheses. Such work will bring us closer to understanding not only how much data matters, but what kind of learner makes that data meaningful. A further promising direction is to examine *why* ANN learners fail to generalize and how such failures can inform more human-like learning mechanisms.

Limitations

We acknowledge several limitations of the present study. First, our investigation is confined to syntactic phenomena. The Poverty of the Stimulus hypothesis, however, has been explored across multiple linguistic domains—including phonology (Wilson, 2006), syntax (e.g., Lidz et al., 2003; Crain and Nakayama, 1987; De Villiers et al., 1990; Crain, 1991; Yedetore et al., 2023; Perfors et al., 2011), lexical learning (Braine et al., 1990; Scott and Fisher, 2009), and semantics (Crain et al., 2000; Falmagne, 2013; Papafragou and Musolino, 2003). We focused on syntactic constructions that are both theoretically central and empirically well-attested, leaving other domains, such as raising and passive constructions (Becker, 2006; Hirsch et al., 2007; Choe and Deen, 2016; Armon-Lotem et al., 2016), for future work due to ongoing debates over their experimental reliability.

Second, our models receive linguistic input solely in textual form, whereas human learners experience multimodal input rich in prosody, gesture, and social interaction. Due to the current limitations of multimodal language models, we defer a systematic exploration of these channels to future work.

Third, the quality of our filtered corpus depends on the accuracy of automatic syntactic parsing. We employ the Stanza combined parser, but residual parsing errors may introduce noise in identifying and removing positive direct evidence, potentially attenuating the effects we report.

Additionally, our experiments are limited to English and GPT-2–style architectures; whether these findings extend to languages with different typological properties or to architectures with explicit memory mechanisms remains to be tested.

Finally, the inductive biases tested here represent only a subset of the possible design space. Architectural, training-dynamic, and attentional biases (e.g., Sartran et al., 2022; Oba et al., 2023; Yamakoshi et al., 2025) could influence structural generalization in ways not captured by our current manipulations. Exploring these broader classes of biases remains an important direction for future research.

Acknowledgments

References

Kabir Ahuja, Vidhisha Balachandran, Madhur Panwar, Tianxing He, Noah A. Smith, Navin Goyal, and Yulia Tsvetkov. 2025. Learning syntax without planting trees: Understanding hierarchical generalization in transformers. *Transactions of the Association for Computational Linguistics*, 13:121–141.

Nameera Akhtar. 2005. The robustness of learning through overhearing. *Developmental Science*, 8(2):199–209.

Ben Ambridge and Elena VM Lieven. 2011. *Child language acquisition: Contrasting theoretical approaches*. Cambridge University Press.

Sharon Armon-Lotem, Ewa Haman, Kristine Jensen de López, Magdalena Smoczynska, Kazuko Yatsushiro, Marcin Szczerbinski, Angeliek Van Hout, Ineta Dabašinskienė, Anna Gavarró, Erin Hobbs, and 1 others. 2016. A large-scale cross-linguistic investigation of the acquisition of passive. *Language acquisition*, 23(1):27–56.

Carl Lee Baker. 1978. *Introduction to Generative Transformational Syntax*. Prentice-Hall, Englewood Cliffs, New Jersey.

Misha Becker. 2006. There began to be a learnability puzzle. *Linguistic Inquiry*, 37(3):441–456.

Melissa Bowerman. 1988. The'no negative evidence'problem: How do children avoid constructing an overly general grammar? In *Explaining language universals*, pages 73–101. Basil Blackwell.

Martin DS Braine, Ruth E Brody, Shalom M Fisch, Mara J Weisberger, and Monica Blum. 1990. Can children use a verb without exposure to its argument structure? *Journal of Child language*, 17(2):313–342.

Roger Brown. 1970. Derivational complexity and order of acquisition. *Cognition and Development of Language*.

Marisa Casillas, Penelope Brown, and Stephen C Levinson. 2020. Early language experience in a tseltal mayan village. *Child Development*, 91(5):1819–1835.

Chi-Yun Chang, Xueyang Huang, Humaira Nasir, Shane Storks, Olawale Akingbade, and Huteng Dai. 2025. Mind the gap: How BabyLMs learn filler-gap dependencies. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15060–15076, Suzhou, China. Association for Computational Linguistics.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, and 1 others. 2025. Babylm turns 3: Call for papers for the 2025 babylm workshop. *arXiv preprint arXiv:2502.10645*.

Yu-Chin Chien and Kenneth Wexler. 1990. Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. *Language acquisition*, 1(3):225–295.

Jinsun Choe and Kamil Deen. 2016. Children's difficulty with raising: A performance account. *Language Acquisition*, 23(2):112–141.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. 11. MIT press.

Noam Chomsky. 1993. *Lectures on government and binding: The Pisa lectures*. 9. Walter de Gruyter.

Noam Chomsky and 1 others. 1976. *Reflections on language*. Temple Smith London.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Stephen Crain. 1991. Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, 14(4):597–612.

Stephen Crain, Andrea Gualmini, and Luisa Meroni. 2000. The acquisition of logical words. *LOGOS and Language*, 1(1):49–59.

Stephen Crain and Cecile McKee. 1985. The acquisition of structural restrictions on anaphora. In *Proceedings of NELS*, volume 15, pages 94–110.

Stephen Crain and Mineharu Nakayama. 1987. Structure dependence in grammar formation. *Language*, pages 522–543.

Alejandrina Cristia, Emmanuel Dupoux, Michael Gurven, and Jonathan Stieglitz. 2019. Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child development*, 90(3):759–773.

Ewa Dąbrowska, Caroline Rowland, and Anna Theakston. 2009. The acquisition of questions with long-distance dependencies.

Jill De Villiers and Thomas Roeper. 1995. Relative clauses are barriers to wh-movement for young children. *Journal of child Language*, 22(2):389–404.

Jill De Villiers, Thomas Roeper, and Anne Vainikka. 1990. The acquisition of long-distance rules. *Language processing and language acquisition*, pages 257–297.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Emeryse Emond and Rushen Shi. 2025. The knowledge of binding principles in early child grammar: Experimental evidence from 30-month-old toddlers. *Language Acquisition*, 32(3):268–296.

Joffe Rachel Falmagne. 2013. Language and the acquisition of logical knowledge. In *Reasoning, necessity, and logic*, pages 111–131. Psychology Press.

Steven Y. Feng, Noah D. Goodman, and Michael C. Frank. 2024. Is child-directed speech effective training data for language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22055–22071, Miami, Florida, USA. Association for Computational Linguistics.

Michael Fetters and Jeffrey Lidz. 2017. Early knowledge of relative clause islands and island repair.

Cynthia Fisher, Yael Gertner, Rose M Scott, and Sylvia Yuan. 2010. Syntactic bootstrapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(2):143–149.

Penelope Floor and Nameera Akhtar. 2006. Can 18-month-old infants learn words by listening in on conversations? *Infancy*, 9(3):327–339.

Stephani Foraker, Terry Regier, Naveen Khetarpal, Amy Perfors, and Joshua Tenenbaum. 2009. Indirect evidence and the poverty of the stimulus: The case of anaphoric one. *Cognitive Science*, 33(2):287–300.

Michael C Frank. 2023. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 27(11):990–992.

Michael C Frank and Noah D Goodman. 2025. Cognitive modeling using artificial intelligence. *Annual Review of Psychology*, 77.

Heidi R Getz. 2019. Acquiring wanna: Beyond universal grammar. *Language Acquisition*, 26(2):119–143.

Jill Gilkerson, Jeffrey A Richards, Steven F Warren, Judith K Montgomery, Charles R Greenwood, D Kimbrough Oller, John HL Hansen, and Terrance D Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology*, 26(2):248–265.

Helen Goodluck, Michele Foley, and Julie Sedivy. 1992. Adjunct islands and acquisition. In *Island constraints: Theory, acquisition and processing*, pages 181–194. Springer.

Betty Hart and Todd R Risley. 1992. American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental psychology*, 28(6):1096.

Christopher Hirsch, Robin Orfitelli, and Kenneth Wexler. 2007. The acquisition of raising reconsidered. In *Language acquisition and development: Proceedings of GALA*, pages 253–262.

Mina Robinson Hirzel. 2022. *Island Constraints: What is there for children to learn?* Ph.D. thesis, University of Maryland, College Park.

Katherine Howitt, Sathvik Nair, Allison Dods, and Robert Melvin Hopkins. 2024. Generalizations across filler-gap dependencies in neural language models. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 269–279, Miami, FL, USA. Association for Computational Linguistics.

Michael Y. Hu, Jackson Petty, Chuan Shi, William Merrill, and Tal Linzen. 2025. Between circuits and Chomsky: Pre-pretraining on formal languages imparts linguistic biases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9691–9709, Vienna, Austria. Association for Computational Linguistics.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Haerim Hwang. 2024. Wanna contraction in first language acquisition, child second language acquisition, and adult second language acquisition. *Bilingualism: Language and Cognition*, 27(3):322–333.

Xuân-Nga Cao Kam, Iglika Stoyneshka, Lidiya Tornyova, Janet D Fodor, and William G Sakas. 2008. Bigrams and the richness of the stimulus. *Cognitive science*, 32(4):771–787.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Elena Koulaguina and Rushen Shi. 2013. Abstract rule learning in 11-and 14-month-old infants. *Journal of psycholinguistic research*, 42:71–80.

Nur Lan, Emmanuel Chemla, and Roni Katzir. 2024. Large language models and the argument from the poverty of the stimulus. *Linguistic Inquiry*, pages 1–28.

Julie Anne Legate and Charles D Yang. 2002. Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review*, 19(1-2):151–162.

Bingzhi Li, Lucia Donatelli, Alexander Koller, Tal Linzen, Yuekun Yao, and Najoung Kim. 2023. SLOG: A structural generalization benchmark for semantic parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3213–3232, Singapore. Association for Computational Linguistics.

Jeffrey Lidz, Sandra Waxman, and Jennifer Freedman. 2003. What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition*, 89(3):295–303.

David Lightfoot. 1991. *How to set parameters: Arguments from language change*. Mit Press Cambridge, MA.

Deborah L Linebarger and Dale Walker. 2005. Infants' and toddlers' television viewing and language outcomes. *American behavioral scientist*, 48(5):624–645.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd edition. Lawrence Erlbaum Associates, Mahwah, NJ.

R. Thomas McCoy, 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.

R Thomas McCoy and Thomas L Griffiths. 2025. Modeling rapid language learning by distilling bayesian priors into artificial neural networks. *Nature communications*, 16(1):4676.

Cecile McKee. 1992. A comparison of pronouns and anaphors in italian and english acquisition. *Language acquisition*, 2(1):21–54.

Masato Mita, Ryo Yoshida, and Yohei Oseki. 2025. Developmentally-plausible working memory shapes a critical period for language acquisition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9386–9399, Vienna, Austria. Association for Computational Linguistics.

Jessica L Montag, Michael N Jones, and Linda B Smith. 2015. The words children hear: Picture books and the statistics for language learning. *Psychological science*, 26(9):1489–1496.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher Manning. 2023. Grokking of hierarchical structure in vanilla transformers. In *Proceedings*

of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 439–448, Toronto, Canada. Association for Computational Linguistics.

Mineharu Nakayama. 1987. Performance factors in subject-auxiliary inversion by children. *Journal of Child Language*, 14(1):113–125.

Elissa L Newport. 1990. Maturational constraints on language learning. *Cognitive science*, 14(1):11–28.

Miyu Oba, Akari Haga, Akiyo Fukatsu, and Yohei Oseki. 2023. BabyLM challenge: Curriculum learning based on sentence complexity approximating language acquisition. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 290–297, Singapore. Association for Computational Linguistics.

Miyu Oba, Yohei Oseki, Akiyo Fukatsu, Akari Haga, Hiroki Ouchi, Taro Watanabe, and Saku Sugawara. 2024. Can language models induce grammatical knowledge from indirect evidence? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20591–20603, Miami, Florida, USA. Association for Computational Linguistics.

Francesca Padovani, Jaap Jumelet, Yevgen Matusevych, and Arianna Bisazza. 2025. Child-directed language does not consistently boost syntax learning in language models. *arXiv preprint arXiv:2505.23689*.

Anna Papafragou and Julien Musolino. 2003. Scalar implicatures: experiments at the semantics–pragmatics interface. *Cognition*, 86(3):253–282.

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. 2024. Filtered corpus training (FiCT) shows that language models can generalize from indirect evidence. *Transactions of the Association for Computational Linguistics*, 12:1597–1615.

Lisa Pearl. 2022. Poverty of the stimulus without tears. *Language Learning and Development*, 18(4):415–454.

Lisa Pearl and Benjamin Mis. 2011. How far can indirect evidence take us? anaphoric one revisited. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Lisa S Pearl and Benjamin Mis. 2016. The role of indirect positive evidence in syntactic acquisition: A look at anaphoric one. *Language*, 92(1):1–30.

Amy Perfors, Joshua B Tenenbaum, and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338.

Massimo Piattelli-Palmarini, editor. 1980. *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky*. Harvard University Press.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*.

Geoffrey K Pullum and Barbara C Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The linguistic review*, 19(1-2):9–50.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Tian Qin, Naomi Saphra, and David Alvarez-Melis. 2024. Sometimes i am a tree: Data drives unstable hierarchical generalization. *arXiv preprint arXiv:2412.04619*.

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

Florencia Reali and Morten H Christiansen. 2004. Structure dependence in language acquisition: Uncovering the statistical richness of the stimulus. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26.

Terry Regier and Susanne Gahl. 2004. Learning the unlearnable: The role of missing evidence. *Cognition*, 93(2):147–155.

John Robert Ross. 1967. Constraints on variables in syntax.

Jerry Samet. 2008. The historical controversies surrounding innateness.

Laurent Sartran, Samuel Barrett, Adhiguna Kuncoro, Miloš Stanojević, Phil Blunsom, and Chris Dyer. 2022. Transformer grammars: Augmenting transformer language models with syntactic inductive biases at scale. *Transactions of the Association for Computational Linguistics*, 10:1423–1439.

Rose M Scott and Cynthia Fisher. 2009. Two-year-olds use distributional cues to interpret transitivity-alternating verbs. *Language and cognitive processes*, 24(6):777–803.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language*, pages 17–60. CRC Press.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2024. Using computational models to test syntactic learnability. *Linguistic Inquiry*, 55(4):805–848.

Colin Wilson. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive science*, 30(5):945–982.

Takateru Yamakoshi, Thomas L. Griffiths, R. Thomas McCoy, and Robert D. Hawkins. 2025. Evaluating distillation methods for data-efficient syntax learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14834–14847, Suzhou, China. Association for Computational Linguistics.

Aditya Yedetore, Tal Linzen, Robert Frank, and R. Thomas McCoy. 2023. How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9370–9393, Toronto, Canada. Association for Computational Linguistics.

Andrea Zukowski and Jaiva Larsen. 2011. Wanna contraction in children: Retesting and revising the developmental facts. *Language Acquisition*, 18(4):211–241.

A Statistics of the BABY and WIKI datasets

The detailed statistics for our WIKI BABY-F, and BABY datasets can be found in Table 9.

B Filtering the Corpus

This section details the procedure used to remove sentences containing *positive direct evidence* (PDE) for the target PoS phenomena. Among the five main constructions examined, three (Question Formation, Anaphoric *one*, and Binding) have identifiable PDE that could directly support the correct generalization. Our goal is to construct a conservative filtering procedure that minimizes the risk of leaving potential PDE in the training corpus, even at the cost of including some false positives.

Parsing and General Strategy All sentences in the BABY corpus were parsed using Stanza (Qi et al., 2020). For the filtered condition (BABY-F), we applied rule-based heuristics over dependency parses and POS tags. To ensure coverage, rules were deliberately designed to err on the side of over-filtering (i.e., allowing more false positives rather than false negatives).

Question Formation Sentences were removed if they (i) contained a relative clause modifying the subject, or (ii) were interrogative. For (i), we identified sentences containing both `acl:relcl` and `nsubj` dependencies where the token labeled `nsubj` precedes that labeled `acl:relcl`. For (ii), we detected questions based on the presence of a final question mark, a wh-word, or an auxiliary verb among the first two tokens.

Anaphoric *one* Direct evidence for anaphoric *one* requires discourse contexts with repeated complex noun phrases. Since such configurations are rare, we further disrupted potential co-reference cues by randomly shuffling word order within the relevant noun phrases, thereby eliminating patterns that could support the correct substitution rule. These examples were not restored in the unfiltered (BABY) version.

Binding To remove PDE for reflexive binding, we excluded sentences containing a reflexive ending in *self* that was preceded by at least two noun or pronoun tokens. POS tags for identifying reflexives and potential antecedents were extracted from Stanza parses.

C Corpus Token Count

The corpus token count with different vocabluary sizes can be found in Table 10.

D Hyperparameter setting

We experiment with models with 4 different architectures, which we call GPT2-small, GPT2-mini, and GPT2-xs and GPT2-xxs. We list their architecture information in Table 5.

Parameter	MINI	XS	XXS	SMALL
Hidden size	512	512	512	768
#Heads	8	8	4	12
#Layers	4	6	6	12
FFN dim	2048	2048	2048	3072

Table 5: Model configurations used in experiments.

In Experiment 3, the dynamic recency bias condition was trained and evaluated by *epochs* rather than by gradient steps, due to the implementation of the time-dependent bias update. We adopted our

Hyperparameter	Setting
learning rate	1e-4
batch size	32
context length	512
warmup steps	4,000
steps	100,000
patience	6,000
dropout	0.1
weight decay	0.1
learning rate scheduler	linear

Table 6: Hyperparameter settings for the experiments

standard GPT-2 mini architecture but followed the hyperparameter configuration of Mita et al. (2025) for comparability. Specifically, the maximum context length was set to 32 and the batch size to 512, as these settings yielded stable learning and positive performance in their reported experiments.

E Bayesian mixed model results

The Baysian mixed model results are reported in Table 7.

Predictor	Estimate	95% CrI
Intercept	0.61	[0.56, 0.66]
Size (z)	0.04	[0.02, 0.06]
Category: Anaphoric one	-0.11	[-0.17, -0.04]
Category: Binding	0.12	[0.09, 0.16]
Category: Island	0.00	[-0.03, 0.04]
Category: Question formation	-0.04	[-0.08, -0.00]
Filtered (no vs. yes)	-0.02	[-0.04, 0.01]
Source (baby vs. wiki)	0.02	[0.01, 0.04]
Interaction Effects		
Size (z) × Filtered	-0.01	[-0.04, 0.01]
Anaphoric one × Filtered	-0.07	[-0.13, -0.01]
Binding × Filtered	0.01	[-0.01, 0.04]
Island × Filtered	0.02	[-0.01, 0.05]
Question formation × Filtered	0.03	[-0.01, 0.07]
Group-level standard deviations		
Dataset intercept SD	0.04	[0.01, 0.12]
Residual SD (σ)	0.19	[0.18, 0.20]

Table 7: Bayesian linear mixed-effects model with random intercepts for dataset. Entries are posterior estimates with 95% credible intervals. For two-level factors, positive estimates indicate higher values for the first level (no vs. yes; baby vs. wiki).

F Fine-grained Results

The subcateogry-wise results can be found in Table 8.

		10M			30M			50M			100M	
Category	**Phenomenon**	BABY-F	BABY	WIKI	BABY-F	BABY	WIKI	BABY-F	BABY	WIKI	BABY	WIKI
Anaphoric One	anaphoric-one-syntax	52.7	47.1	28.3	65.7	64.5	29.5	64.7	57.3	35.5	49.2	28.8
Island	island-adjunct	71.9	68.2	69.7	83.1	82.9	70.7	89.7	88.2	78.7	90.6	89.4
	island-complex-np	51.1	53.2	53.5	45.9	62.3	71.8	46.1	53.1	62.5	71.8	55.4
	island-wh	92.9	89.4	86.1	83.2	87.7	86.6	82.0	86.6	84.1	83.0	83.2
Question Formation	question-formation_or	61.3	61.4	65.4	54.7	55.7	61.4	51.8	49.9	58.0	51.2	59.6
	question-formation_rr	54.9	56.4	63.2	56.3	53.8	64.9	55.5	56.0	65.5	61.6	70.8
	question-formation_sr	62.4	59.8	60.2	61.1	58.9	60.2	61.7	61.5	65.4	63.8	63.6
Binding	principle_a_command	42.5	46.7	30.5	45.6	39.5	45.6	50.8	41.2	41.5	55.6	33.0
	principle_a_locality	65.0	64.5	68.6	79.3	79.7	81.7	83.3	83.5	84.5	92.4	83.2
Wanna	wanna	59.3	64.0	56.1	63.3	43.8	56.6	76.3	62.7	55.2	89.0	96.8
Average		61.4	61.1	58.2	63.8	62.9	62.9	66.2	64.0	63.1	70.8	66.4

Table 8: Category-wise results grouped by training size (columns) and category (rows).

Dataset	**10M**		**30M**		**50M**		**100M**	
	#Sents	#Words/Sent	#Sents	#Words/Sent	#Sents	#Words/Sent	#Sents	#Words/Sent
WIKI	430k	23.2	1.3M	23.52	2.1M	23.41	4.3M	23.51
BABY	1.4M	7.13	4.0M	7.53	6.6M	7.46	11.6M	8.55
BABY-F	1.4M	7.11	4.0M	7.50	6.6M	7.43	NA	NA

Table 9: Statistics of training corpora by data size. Each size group reports the number of sentences and average sentence length.

Vocab size	Dataset	10M	30M	50M	100M
8192	WIKI	17,341,898	54,230,974	90,741,059	181,490,178
	baby	16,216,427	50,045,572	83,924,585	159,613,479
32768	WIKI	14,442,880	45,544,347	76,384,301	153,006,220
	baby	153,18,669	46,823,177	78,091,016	146,768,608
49152	WIKI	13,989,838	44,114,584	74,020,389	148,359,082
	baby	15,181,258	46,370,173	77,257,389	144,939,275
65536	WIKI	13,741,202	43,308,251	72,687,684	145,743,672
	baby	15,114,124	46,128,756	76,809,625	143,962,866

Table 10: The corpus token count (CTC) of BPE tokenizers with different vocab size trained with different datasets