

Can Child Language Acquisition be Modeled by Linear Causal or Bidirectional Masked Models Trained on Child-Directed Input?: A Comparison of Baby Language Models and Children in Acquiring *Wanna* Contraction

Haerim Hwang

The Chinese University of Hong Kong

haerimhwang@cuhk.edu.hk

Abstract

Children’s remarkable ability to acquire language remains one of the most fascinating questions in cognitive science. How do they develop sophisticated syntactic knowledge? To contribute to this question, the present study investigates the role of input and learning mechanisms by modeling child language acquisition across increasing age. Specifically, this study trains 30 baby language models (LMs) on naturalistic input directed to children up to ages 3 through 7 and compares their syntactic knowledge directly with that of children aged 3 to 7, focusing on the constraint on *wanna* contraction, a phenomenon not previously tested in language modeling. Furthermore, the study examines two major LM architectures: an autoregressive causal LM, represented by GPT, and a bidirectional masked LM, represented by RoBERTa. The results showed that children were sensitive to the constraint and displayed age-related development, but neither GPT-based nor RoBERTa-based baby models showed the same pattern. In this respect, the baby LMs do not seem to fully capture human-like development, possibly suggesting that the learning mechanisms underlying child language acquisition do not entirely align with the architectures tested here and cannot be reduced to input-driven learning alone within the context of these models.

1 Introduction

The question of how children acquire grammatical knowledge has been central to research in linguistics and cognitive science (Hsu et al., 2011; Lidz and Gagliardi, 2015; Linzen and Baroni, 2021; Piantadosi, 2023; Yang, 2016; Yang et al., 2017). Is the linguistic input available to children sufficient to explain the grammatical knowledge they ultimately develop (see also Chang et al., 2025)? Some theoretical approaches hold that language acquisition is fundamentally input-driven, with domain-

general cognitive abilities supporting the abstraction of linguistic forms (e.g., Lewis and Elman, 2001; Tomasello, 2003). According to other approaches, the input children receive is too sparse and underdetermined to lead to the acquisition of abstract, hierarchical structures, making innate language-specific cognitive mechanisms necessary (e.g., Chomsky, 1968; Chomsky, 1973; Crain et al., 2006; Yang, 2004). The current study aims to contribute to this area of research by investigating how children and baby language models (LMs), trained on child-directed input, evaluate sentences that involve an abstract structural dependency. While a number of studies have moved in this direction (e.g., Padovani et al., 2025), few have included language performance data from human children, thereby limiting the extent to which additional insights can be gained.

The significance of the present study lies in two main respects. First, this study directly compares the performance of children with that of baby LMs to reveal whether the processes underlying child language acquisition are similar to, or different from, the mechanisms these models rely on for language learning. Specifically, the study examines two major architectures: an auto-regressive causal model, represented by GPT (Radford et al., 2019), and a bidirectional masked model, represented by RoBERTa (Zhuang et al., 2021). These two approaches learn in fundamentally different ways: the former learns by predicting each next token based on the preceding ones in a linear, left-to-right fashion, consistent with the error-based learning account that humans process language incrementally and learn when the incoming input diverges from their expectations (Chang et al., 2012); the latter learns by reconstructing masked tokens from their surrounding context, thereby building strong bidirectional text representations that may better capture structural dependencies existing in human languages across both visible and invisible

linguistic material (e.g., [Jawahar et al., 2019](#)).

Furthermore, the study systematically varies the amount of input and vocabulary size in LMs to simulate child language acquisition. As an initial step in examining the role of input in language acquisition, it trains five LMs on different datasets representing the input available up to ages 3 through 7, while comparing their performance against children's performance. Specifically, the study uses the CHILDES corpus ([MacWhinney and Erlbaum, 2000](#)), which contains naturalistic child-directed input, rather than relying on data that is commonly used in this area of research (e.g., Wikipedia).

The target linguistic phenomenon in this study is the contraction of *wanna*, which is a reduced form of *want to*. This construction yields a subtle contrast in grammaticality, as shown in (1) versus (2). Whereas *want to* cannot contract to *wanna* when the question is about the subject of the *to*-infinitival clause (e.g., (2)), such contraction is possible if the question is about any other constituents within the *to*-infinitival clause, such as a direct object of the verb (e.g., (1a)) or the preposition (e.g., (1b)). Some linguists assume that the contrast in *wanna* contraction results from the blocking effect created by *wh*-movement. In (2), the subject of the *to*-infinitival complement is extracted from the position between *want* and *to* and moved to the front. This movement leaves behind a gap (which is marked as *_*) in its original position, preventing *want* and *to* from contracting across it. In contrast, (1) involves the extraction of non-subject elements from positions elsewhere within the infinitival clause, not from between *want* and *to*, so no blocking effect arises. In this way, *wanna* contraction is subject to a structural constraint (cf., [Getz, 2019](#); [O'Grady et al., 2008](#)). The present study focuses on this phenomenon because it can be expressed using simple vocabulary, making it suitable for child participants and baby LMs, while still exhibiting structural dependencies.

- (1) a. Who do you [want to, wanna] kiss *_*?
b. Who do you [want to, wanna] dive with *_*?
([Zukowski and Larsen, 2011](#), p. 213, (1a)–(1b))
- (2) a. Who do you [want *_* to, *wanna] kiss you?
b. Who do you [want *_* to, *wanna] dive first?
([Zukowski and Larsen, 2011](#), p. 213, (2))

Child language acquisition research on *wanna* contraction showed somewhat inconsistent results, mostly using a production paradigm (cf., [Hwang, 2024](#)). [Thornton \(1990\)](#) reported that children as young as 3;6 demonstrated target-like knowledge of the constraint on *wanna* contraction. However, [Zukowski and Larsen \(2011\)](#), in a later study that addressed a methodological issue involving the use of different prompts across experimental conditions in [Thornton \(1990\)](#)'s study, suggested that young children (age: 4;00–7;03) do not possess grammatical knowledge of *wanna* contraction. A similar pattern was shown by [Getz \(2019\)](#), whose findings showed that children's use of *wanna* diverged from adults across age groups (younger group's age: 3;09–4;10; older group's age: 5;00–7;03). The study also found a frequency effect limited to impossible *wanna* contraction contexts, where children produced fewer impossible contractions when the embedded verb was more frequent. Although this finding appears to support the view that input frequency plays a decisive role in child language acquisition, it is important to note that the frequency effect emerged only in the case of impossible *wanna* contraction. If children learn through the storage of specific patterns and their associated frequencies, then a similar frequency effect should also be expected for possible *wanna* contraction, at least among younger children, who are still building abstract generalizations on the basis of item-based patterns ([Hwang, 2024](#)).

It is only in recent work that children's knowledge of *wanna* contraction has been tested in comprehension. Since production is more likely to be affected by performance factors, such as cognitive demands, and since an inability to produce a form does not necessarily reflect a lack of underlying knowledge, [Hwang \(2024\)](#) assessed children's knowledge using a child-friendly acceptability judgment task. The results indicated that children as young as 3;11 can distinguish between possible and impossible sentences involving *wanna* contraction, consistent with [Thornton \(1990\)](#).

Notably, while previous language modeling studies have examined linguistic phenomena involving structural dependencies, including parasitic gaps and across-the-board movement ([Lan et al., 2026](#)), as well as island effects ([Chang et al., 2025](#); [McInerney, 2025](#)), no study has specifically investigated *wanna* contraction, despite the considerable attention it has received in child language acquisition (see above). Moreover, this line of research

has yielded mixed findings. [McInnerney \(2025\)](#) reported limited evidence of island effects in causal and masked LMs, only when the models were trained on developmentally implausible quantities of data, including Wikipedia and the Pile ([Gao et al., 2020](#)). [Chang et al. \(2025\)](#) addressed this issue by using “predominantly” child-oriented and child-sized input to train their LMs and found that the LMs did not consistently achieve full structural generalization across the linguistic phenomena they tested. Building on the work of [Chang et al. \(2025\)](#) but using “only” child-oriented data, the present study extends the investigation to consider the following research questions (RQs).

- RQ 1: Can LMs trained on child-oriented, child-sized input acquire the constraint on *wanna* contraction?
- RQ 2: Do LMs exhibit child-like age-related development of sensitivity to the constraint on *wanna* contraction, as they are trained on datasets representing older children and a larger vocabulary size?
- RQ 3: Which LM type is better at learning *wanna* contraction: linear causal models or bidirectional masked models?

These RQs give rise to several plausible predictions. For RQ 2, it is conceivable that baby LMs trained on a dataset covering older children and employing a larger vocabulary size may align more closely with older children’s developmental patterns, thereby showing greater sensitivity than baby LMs trained on a dataset restricted to younger children and a smaller vocabulary size. As for RQ 3, GPT-based LMs’ performance may better resemble children’s because its auto-regressive architecture mirrors the incremental way in which children process language in real time and predict the upcoming information ([Chang et al., 2012](#)). Alternatively, RoBERTa-based LMs’ performance may align more closely with children’s because language itself relies on bidirectional dependencies. Even though language learners do not encounter sentences in both directions, findings from a preferential-looking experiment using eye-tracking indicate that they reanalyze and fix their initial parses and interpretations (e.g., [Abbot-Smith et al., 2017](#)). However, it is also possible that neither model architecture, nor any dataset designed to approximate human input,

will show human-like performance (RQ 1 and RQ 3). Taken together, the findings of this study may provide insight into whether language acquisition is solely based on autoregressive or bidirectional parsing mechanisms and/or whether it can be understood as purely input-driven within such learning architectures. Ultimately, they are expected to shed light on the mechanisms involved in child language acquisition.

2 Methodology

The methodology used in the present study draws on earlier work examining the grammatical knowledge of LMs (e.g., [Chang et al., 2025](#); [Howitt et al., 2024](#); [Ozaki et al., 2022](#); [Wilcox et al., 2018](#)). Prior to the beginning of the study, administrative and ethical approval had been obtained, along with informed consent from one parent of each child participant.

2.1 Child participants

Data from 46 child participants were drawn from [Hwang \(2024\)](#). All participants lived in the United States at the time of testing. The age distribution was as follows: 3-year-olds ($n = 2$), 4-year-olds ($n = 3$), 5-year-olds ($n = 13$), 6-year-olds ($n = 19$), and 7-year-olds ($n = 9$). These participants were recruited through flyers distributed to their parents at a local kindergarten. Upon completing the study, all participants were given stationery items equivalent to \$5 (e.g., stickers), following the convention in the area where they resided.

2.2 Baby language models

The current study sought to approximate the linguistic input available in child language acquisition by training a range of baby LMs that differed in three main respects: (1) model architecture, (2) vocabulary size, and (3) random seed. With respect to architecture, two types of LM were used: GPT-2-based models ([Radford et al., 2019](#)), which instantiate linear causal modeling, and RoBERTa-based models ([Zhuang et al., 2021](#)), which instantiate bidirectional masked modeling. This comparison enables to assess whether, and to what extent, the learning procedures associated with these model types can account for the acquisition of abstract syntactic knowledge. To ensure comparability, both models used the same architectural specifications, with 8 transformer layers, 8 attention heads, an embedding size of 512, and an intermediate feed-forward size of 2,048.

The specific model configurations in the present study are provided in Appendix A. For each LM type, 15 separate models were trained on authentic linguistic input directed to English-speaking children. Given that the child participants in this study resided in the United States at the time of testing, all data for training LMs were drawn from the North American English portion of CHILDES (MacWhinney and Erlbaum, 2000). On the basis of this corpus, five datasets were constructed, with each representing input available up to ages 3, 4, 5, 6, and 7 (see Appendix B).

Vocabulary size was also systematically varied so that it would approximate the vocabulary growth typically observed across these age groups. In the context of language modeling, vocabulary size, which is defined as the number of distinct token types that an LM can represent, determines how finely text is segmented into units, such as words or subwords. This factor is important for investigating LM performance because it directly affects sequence length, memory and computational cost, and the extent to which the model captures linguistic patterns rather than artifacts of tokenization. Following Shipley and McAfee (2015), vocabulary sizes were set to 900 for 3-year-olds, 2,000 for 4-year-olds, 10,000 for 5-year-olds, 13,000 for 6-year-olds, and 20,000 for 7-year-olds. For each of the two architectures, baby LMs were trained on each of the five datasets three times using different random seeds, namely 13, 30, and 42.

Importantly, the present modeling setup was designed to approximate age-conditioned learner states rather than to isolate the independent contribution of any single developmental factor. To approximate children’s linguistic experience as closely as possible, the age conditions were jointly varied in terms of exact input data, vocabulary size, and resulting parameter count, mirroring how these factors naturally change together during language development. In this respect, the models should be interpreted as cross-sectional snapshots of learners at different points in development, consistent with our cross-sectional design for child participants with different ages, rather than as sequential checkpoints from a single learner tracked longitudinally (Shah et al., 2024). Overall, the goal of this study is ecological approximation rather than causal decomposition. That is, the study focuses on whether models trained under age-indexed input conditions show behavior that aligns with children’s performance, rather than which individual component of

development is responsible for such behavior.

The codebase used in this study was adapted from Padovani et al. (2025)¹ solely for the purposes of this research. Following their setup, all models were trained from scratch for 100,000 steps using the AdamW optimizer (Loshchilov and Hutter, 2017), together with a linear scheduling scheme and a 40,000-step warm-up phase, which ensured consistent training conditions across models. A learning rate of 0.0001 was also applied uniformly across all experiments (Padovani et al., 2025). The data were split into training and validation sets in a 92/8 ratio. Each baby LM training run required roughly 1 hour and 50 minutes on an RTX 5090. Training and evaluation loss curves of the trained LMs are provided in Appendix C.

2.3 Experimental design

The experiment employed an acceptability judgment task given that this paradigm is well established and widely validated in child language acquisition research (e.g., Ambridge et al., 2008). In the task, the two factors, Clause (Who vs. If) and Gap (Gap vs. No gap), were crossed in a 2×2 Latin square design. This yielded four conditions, as shown in (3). In total, 80 sentences were developed, with 20 sentences for each condition (for the full list, see Appendix D). The two If conditions served as the baseline for comparison with the two Who conditions in terms of grammaticality: the *If + Gap condition was contrasted with the Who + Gap condition, and the If + No gap condition was contrasted with the *Who + No gap condition.

- (3) a. Who + Gap: I wonder who you wanna work with.
- b. *Who + No gap: *I wonder who you wanna work.
- c. *If + Gap: *I wonder if you wanna work with.
- d. If + No gap: I wonder if you wanna work.

To reduce the cognitive demands placed on both child participants and baby LMs trained on child-directed input, all experimental sentences were kept short (6–7 words) and composed of simple, high-frequency vocabulary. The embedded verbs following *wanna* were either clearly intransitive ($k = 16$;

¹https://github.com/fpadovani/childes_vs_wiki

e.g., *work*) or verbs that can be used both transitively and intransitively but do not normally take a human object when transitive ($k = 4$; e.g., *read*, *walk*, *eat*, *drink*).

2.4 Experimental procedure

Child participants completed an acceptability judgment task where every sentence was presented both aurally and in writing (for the specific instructions that the child participants received, see Appendix E). For each trial, participants heard the sentence twice, separated by a one-second interval, while the corresponding sentence was shown on the screen. They then provided their judgments on a four-point smiley-face scale consisting of four response options, including "very bad," "bad," "good," and "very good," alongside an "I don't know" option, which was indicated by a question mark. To avoid practice effects, participants were not given different versions of the same item set (e.g., (3a), (3b), (3c), (3d))). Instead, they saw only one version from each item set, meaning that they were presented with five items per condition, for a total of 20 target items, in addition to 44 fillers.

The syntactic performance of each LM was assessed on the basis of sentence probabilities, following earlier studies (Linzen et al., 2016; Marvin and Linzen, 2018; Wilcox et al., 2018), using the "minicons" library (Misra, 2022). For linear causal models, the summed sequence log-probability with Beginning-of-Word (BOW) correction was used (Pimentel and Meister, 2024; Oh and Schuler, 2024). For bidirectional masked models, a likelihood score based on a within-word left-to-right masking strategy was used, as this reduces the tendency to overestimate token probabilities for words consisting of multiple tokens (Kauf and Ivanova, 2023). Each LM's probabilities were calculated for all 80 experimental sentences.

2.5 Analysis

Prior to conducting the statistical analyses, the children's "I don't know" responses were removed (5.22%). For analysis, "very bad," "bad," "good," and "very good," were respectively coded as 1, 2, 3, and 4. Next, z -transformations were applied to the children's acceptability scores and the LMs' sentence probabilities to place both measures on a common, standardized scale ($M = 0$, $SD = 1$) (e.g., Suijkerbuijk et al., 2025). This procedure minimizes differences in scale and variance between the raw ratings and model probabilities, enabling

direct comparisons across the two measures.

To evaluate LM's performance on *wanna* contraction (RQ 1), a linear mixed-effects regression model was constructed with Clause (Who vs. If), Gap (Gap vs. No gap), Learner type (Children vs. LMs) as categorical fixed effects, Age as a continuous fixed effect, and Item as a random effect. The interaction between Construction and Gap is crucial, as grammaticality aligns with the Gap condition in Who clauses but with the No Gap condition in If clauses; therefore, collapsing these into a single variable, such as grammaticality, could obscure important patterns in the data.

To examine age-related development in sensitivity to the constraint on *wanna* contraction (RQ 2), separate linear mixed-effects regression models were fitted to the child and LM data. These models included Clause (Who vs. If), Gap (Gap vs. No gap), and Age as fixed effects, and Item as a random effect. For a more detailed investigation of age-related effects, grammaticality sensitivity scores were computed for each Clause type by subtracting the acceptability scores of the ungrammatical condition from those of the grammatical condition: i.e., Who + Gap – *Who + No gap; If + No gap – *If + Gap. These grammaticality sensitivity scores were then analyzed using a simple linear regression model with Age as a predictor.

To explore any effect of model architecture (RQ 3), another linear mixed-effects regression model was constructed on the LM data. This model included Clause (Who vs. If), Gap (Gap vs. No gap), and Model architecture (linear causal model vs. bidirectional masked model) as fixed effects, and Item as a random effect.

3 Results

3.1 Knowledge of the constraint on *wanna* contraction

A visual inspection of Figure 1 indicates that the children aged 5 and older rated the Who + Gap and If + No gap conditions higher than the *Who + No gap and *If + Gap conditions, while the children aged 3 and 4 showed no clear pattern. In contrast, the baby LMs consistently rated the If + No gap condition highest, followed by the *If + Gap condition, then the *Who + No gap condition, which became rated more highly than the *If + Gap condition with increasing age—at age 4 for the linear causal model and age 5 for the bidirectional masked model—and finally the Who + Gap condition. This

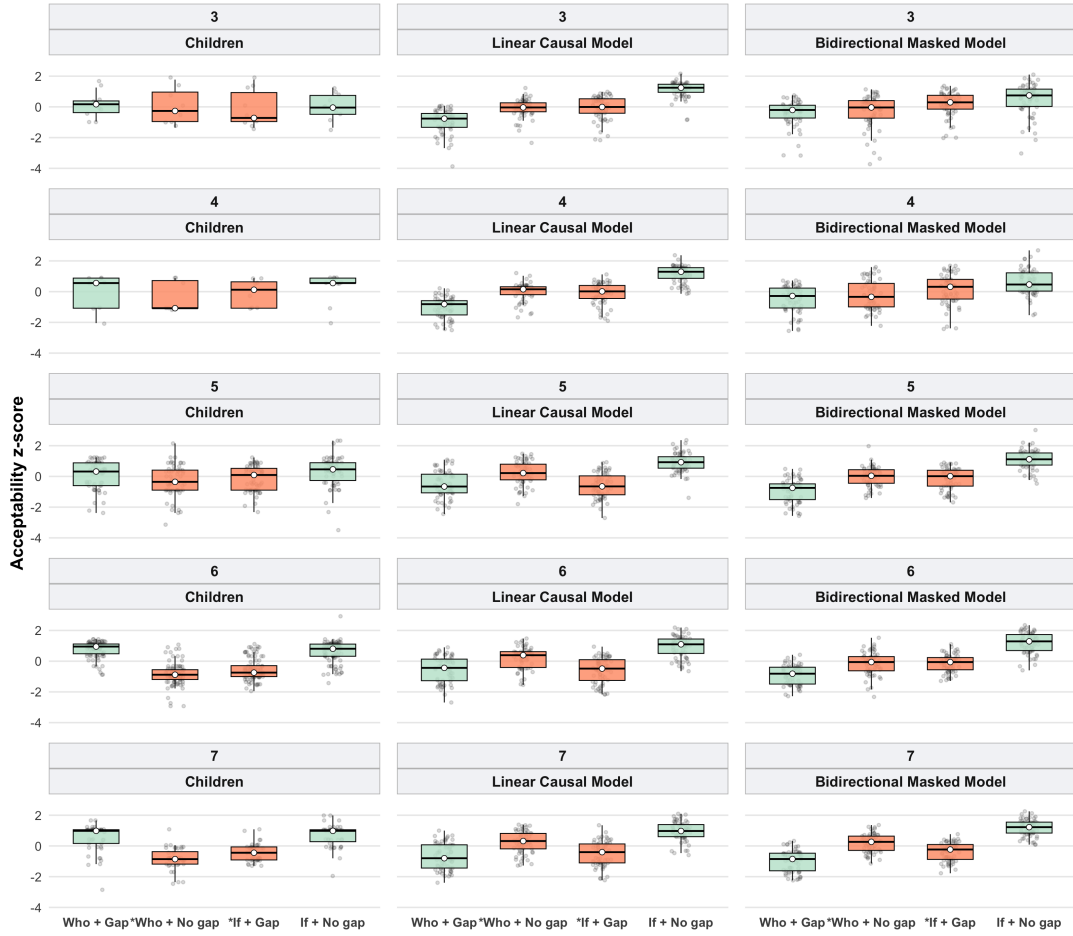


Figure 1: Mean acceptability z-scores by learner type and condition

Note. The results visualized for each language model with a distinct model architecture trained on input directed to children up to ages 3 through 7 are averaged over three random seeds.

pattern indicates that, although the children possess grammatical knowledge of the constraint on *wanna* contraction, the baby LMs developed for this study do not.

As shown in Appendix F, a linear mixed-effects regression model fitted to the acceptability z-scores revealed a significant effect of Clause ($b = .427$, $SE = .147$, $p = .004$), as well as significant interactions between Clause and Gap ($b = -1.06$, $SE = 0.29$, $p < .001$), between Clause and Learner type ($b = 1.29$, $SE = 0.29$, $p < .001$), between Clause, Gap, and Learner type ($b = 2.10$, $SE = 0.59$, $p < .001$), between Clause, Gap, and Age ($b = 0.40$, $SE = 0.05$, $p < .001$), between Clause, Learner type, and Age ($b = -0.13$, $SE = 0.05$, $p = .012$), and between Gap, Learner type, and Age ($b = 0.24$, $SE = 0.05$, $p < .001$). Crucially, the model also identified a four-way interaction among Clause, Gap, Learner type, and Age ($b = -0.62$, $SE = -6.05$, $p < .001$). These results suggest that the children and the LMs differed in their performance on ex-

perimental sentences involving *wanna* contraction, consistent with our visual inspection.

3.2 Age-related development of sensitivity to the constraint on *wanna* contraction

To further investigate the Learner type-related interaction observed in the previous section, separate follow-up linear mixed-effects regression models were constructed for each learner type. In the child model, a significant Clause $\times$ Gap interaction was found ($b = -2.52$, $SE = 0.65$, $p < .001$), as shown in Appendix G. Post hoc pairwise comparisons further indicated that children rated the Who + Gap condition higher than the *Who + No gap condition ($b = 1.09$, $SE = 0.08$, $p < .001$) and the If + No gap condition higher than the *If + Gap condition ($b = 0.89$, $SE = 0.08$, $p < .001$), suggesting robust knowledge of the constraint on *wanna* contraction. In addition, the model revealed a significant interaction among Clause, Gap, and Age ($b = 0.79$, $SE = 0.57$, $p < .001$). Subsequent simple regression

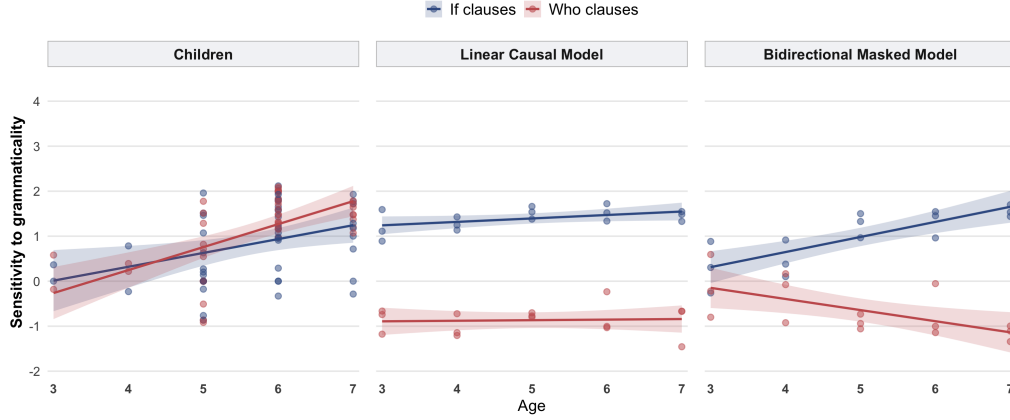


Figure 2: Relationship between age (whether measured in the case of children or approximated from the child-directed input in the case of language models) and grammaticality sensitivity scores by learner type

analyses showed that sensitivity to grammaticality increased with children’s age for both Who clauses ($b = 0.51$, $SE = 0.10$, $p < .001$) and If clauses ($b = 0.31$, $SE = 0.12$, $p = .012$), as shown in Figure 2.

A different pattern emerged from the linear mixed-effects regression model fitted to the baby LM data, as shown in Appendix H. The model revealed a significant effect of Clause ($b = 1.07$, $SE = 0.09$, $p < .001$), as well as significant interactions between Clause and Age ($b = -0.07$, $SE = 0.02$, $p < .001$) and between Gap and Age ($b = 0.16$, $SE = 0.02$, $p < .001$). There was also a significant interaction among Clause, Gap, and Age ($b = 0.09$, $SE = 0.03$, $p = .008$). However, subsequent simple regression analyses indicated that the pattern of age-related development in the baby LMs was different from the one found in children. These analyses showed that sensitivity to grammaticality tended to increase with the baby LMs’ age, as approximated by the input, for If clauses ($b = 0.21$, $SE = 0.05$, $p < .001$), but to decrease with the baby LMs’ age for Who clauses ($b = -0.12$, $SE = 0.06$, $p = .049$), although the pattern differed across the two model architectures (see Figure 2).

3.3 Effects of model architecture

Another linear mixed-effects model applied to the baby LM data showed significant main effects of Clause ($b = 0.73$, $SE = 0.02$, $p < .001$) and Gap ($b = 0.97$, $SE = 0.07$, $p < .001$) (see Appendix I). In addition, the model revealed significant interaction effects between Clause and Gap ($b = 0.43$, $SE = 0.05$, $p < .001$), between Clause and Model architecture ($b = 0.19$, $SE = 0.05$, $p < .001$), and between Gap and Model architecture ($b = -0.32$, $SE = 0.05$,

$p < .001$). Crucially, the analysis also identified a significant three-way interaction among Clause, Gap, and Model architecture ($b = -0.19$, $SE = 0.09$, $p = .049$). Further post hoc pairwise comparisons indicated that both the linear causal model and the bidirectional masked model gave higher ratings to the If + No gap condition than to the *If + Gap condition (linear causal model: $b = -1.29$, $SE = 0.08$, $p < .001$; bidirectional masked model: $b = -0.98$, $SE = 0.08$, $p < .001$), but to the *Who + No gap condition than to the Who + Gap condition (linear causal model: $b = -0.87$, $SE = 0.08$, $p < .001$; bidirectional masked model: $b = -0.64$, $SE = 0.08$, $p < .001$). The three-way interaction seems to be thus driven by the larger difference between the two If conditions in the linear causal model compared with the bidirectional masked model, as shown in Figure 2.

4 Discussion

Regarding RQ 1, the present study found that unlike children, baby LMs trained on child-oriented, child-sized input did not demonstrate knowledge of the constraint on *wanna* contraction. This finding is somewhat difficult to reconcile with the view that input is a key factor in the context of the model architectures tested in this study. In particular, the models struggled to accept the Who + Gap condition and to reject the *Who + No gap condition (see Figure 1). A possible reason is that probabilistic language learning may not be sufficient to capture the syntactic dependency between *who* and the gap. In sum, these results seem to be more compatible with the theoretical accounts positing that input-based, domain-general learning via linear causal

or bidirectional masked learning mechanisms may not alone be sufficient and that language-specific learning mechanisms are needed for the acquisition of complex syntactic knowledge (Chomsky, 1968; Chomsky, 1973; Crain et al., 2006; Yang, 2004; cf., Lewis and Elman, 2001; Piantadosi, 2023; Tomasello, 2003).

Furthermore, baby LMs did not exhibit child-like, age-related development in grammaticality sensitivity to the constraint on *wanna* contraction (RQ 2). The models were trained on a range of datasets designed to approximate the authentic input children receive, as well as the vocabulary sizes they develop at different ages. However, their performance did not reflect human-like developmental changes in grammatical sensitivity. More specifically, they exhibited largely similar patterns across age groups and vocabulary sizes, without showing the target-like grammatical contrast between the Who-Gap and *Who-No gap conditions; this contrasts with the child data, which showed clear developmental growth in the target grammatical sensitivity and suggested full attainment by age 5. Although children receive input that is sparse and often underspecified, they nonetheless come to acquire target-like knowledge of the *wanna* contraction constraint without explicit instruction, suggesting that the development of syntactic knowledge in LMs may differ fundamentally from its development in children.

Regarding RQ 3, neither linear causal models nor bidirectional masked models showed child-like performance on sentences involving *wanna* contraction. This result may be considered surprising, given psycholinguistic evidence that human language processing appears to share some features with the mechanisms used by these models. Humans, like linear causal models, use prior linguistic context to predict upcoming input, and mismatches between prediction and input may contribute to learning. In line with this, error-based learning has been argued to play a role in language development (Chang et al., 2006; Phillips and Ehrenhofer, 2015). On this account, learners continuously generate predictions, and when those predictions are disconfirmed, the gap between predicted and actual input provides implicit feedback that helps move the learner toward the target grammar. On the other hand, human sentence processing also involves re-analysis: parsers may go back to earlier parts of a sentence and revise initial interpretations to resolve syntactic dependencies, similar in some respects

to bidirectional masked models (Lopopolo et al., 2019; Pickering and Traxler, 1998).

Since both types of LMs failed to show robust knowledge of the *wanna* contraction constraint, unlike children, the present findings suggest that while these models may resemble humans in certain aspects of online processing, they still differ from them in the underlying mechanisms that support the acquisition of low-frequency or weakly evidenced grammatical knowledge. This interpretation also aligns with Murphy’s (2025) Representation, Operation, Structure, and Encoding (ROSE) model. In that framework, syntax is divided into two broad domains: a horizontal domain shaped by surface-based distributional regularities, and a vertical domain involving hierarchical and symbolic structure. One reasonable possibility is that LMs are effective at learning surface-level, horizontal patterns but struggle with deeper hierarchical structures, particularly when they are rarer (Hwang, 2026). *Wanna* contraction fits this description, since it depends on complex structural relations and does not provide strong distributional regularities (Hwang, 2024). If LMs are weak at acquiring this type of vertical structure, it is less surprising that they failed to show human-like syntactic knowledge involving structural dependency.

Focusing on vertical structure, Chang et al. (2025) reported a similar pattern of results. Their results showed that although some small GPT-2 models trained primarily on child-oriented input exhibited limited learning of filler-gap dependencies, they still did not achieve human-like island sensitivity. Chang et al. (2025) took this as potentially suggestive of a role for language-dedicated mechanisms in language development. The present findings may be broadly consistent with this interpretation, although it should be noted that alternative architectures and inductive biases remain to be further explored before any firm conclusions can be reached (see Section "Limitations").

5 Conclusion

In conclusion, baby LMs appeared to fall short of fully capturing human-like language development. This suggests some divergence between the learning mechanisms underlying child language acquisition and those instantiated in the tested architectures, as well as the possibility that input-driven learning alone may not be fully sufficient within these model architectures.

Limitations

The conclusions of the present study should be interpreted cautiously in light of several limitations outlined below, which future research should address. One limitation is that the inclusion of child participants and the systematic comparison across age groups made it difficult to examine a wider range of linguistic phenomena. Although the study investigated a novel phenomenon that has not previously been tested in LMs, future work should consider a broader set of phenomena in order to draw stronger generalizations about LMs' syntactic knowledge.

In addition, *wanna* contraction occurs more frequently in spoken than in written language, which may have given child participants an advantage, since the experimental sentences were presented to them in both spoken and written form, whereas the LMs were exposed only to written input. A useful next step would therefore be to investigate whether speech or multimodal LMs show different patterns of syntactic knowledge when auditory forms or both auditory and written forms of the experimental sentences are made available.

While CHILDES offers a large and authentic source of child-directed input, it cannot be assumed to fully reflect the complete linguistic environment to which children are exposed. A more ideal approach would involve collecting input data directly from the children participating in the study and training the language models on this data, so that both the LMs and the children could be tested on syntactic knowledge under more comparable conditions.

Furthermore, it is possible that the LMs developed for this study were too small in scale to adequately approximate the neural capacity that children bring to language learning. Also, the two types of learning mechanisms were represented only by GPT-2 and RoBERTa, corresponding to a linear causal model and a bidirectional masked model, respectively. Although the training data were varied in terms of input age and vocabulary size, the range of models considered remains narrow, which limits the generalizability of our conclusions. Other LMs with different inductive biases, such as LLaMA with decoder-only transformers or GPT-BERT with hybrid architectures, may show different patterns in how they learn syntactic constraints and structures (see also Chang et al., 2025).

Finally, the hyperparameter configuration used

in this study was closely based on Padovani et al. (2025) to replicate and optimize model performance. However, computational limitations restricted the search space explored in this study, and thus a wider range of configurations should be explored in further research (see also Chang et al., 2025).

Supplementary material

The dataset and analysis codes used for the current study are available at <https://osf.io/p48ef> under a CC BY-NC-SA 4.0 license. The dataset shared should not be used outside of research contexts.

Acknowledgments

This study was funded by the Early Career Scheme supported to the author by Research Grants Council in Hong Kong (Reference: 24607925).

AI statement

For grammar checks and rephrasing of some sections, AI tools (e.g., Gemini-3, Perplexity) were used. All content remains the original work of the author.

References

- Kirsten Abbot-Smith, Franklin Chang, Caroline Rowland, Heather Ferguson, and Julian Pine. 2017. [Do two and three year old children use an incremental first-np-as-agent bias to process active transitive and passive sentences?: A permutation analysis](#). *PLOS ONE*, 12:Article e0186129.
- Ben Ambridge, Julian M Pine, Caroline F Rowland, and Chris R Young. 2008. [The effect of verb semantic class and verb frequency \(entrenchment\) on children's and adults' graded judgements of argument-structure overgeneralization errors](#). *Cognition*, 106:87–129.
- 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 15049–15065, Suzhou, China. Association for Computational Linguistics.
- Franklin Chang, Gary S. Dell, and Kathryn Bock. 2006. [Becoming syntactic](#). *psychological review*. *Psychological Review*, 113:234–272.
- Franklin Chang, Marius Janciauskas, and Hartmut Fitz. 2012. [Language adaptation and learning: Getting explicit about implicit learning](#). *Linguistics and Language Compass*, 6:259–278.

- Noam Chomsky. 1968. *Language and Mind*. Cambridge University Press, Cambridge, UK.
- Noam Chomsky. 1973. Problems of knowledge and freedom. *Philosophy*, 48:194–195.
- Stephen Crain, Takuya Goro, and Rosalind Thornton. 2006. [Language acquisition is language change](#). *Journal of Psycholinguistic Research*, 35:31–49.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. [The pile: An 800gb dataset of diverse text for language modeling](#). *arXiv*, 2101.00027:1–39.
- Heidi R. Getz. 2019. [Acquiring wanna: Beyond universal grammar](#). *Language Acquisition*, 26:119–143.
- Kevin Howitt, Sunil Nair, Amelia Dods, and Robert M. 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. Association for Computational Linguistics.
- Anne S. Hsu, Nick Chater, and Paul M.B. Vitányi. 2011. [The probabilistic analysis of language acquisition: Theoretical, computational, and experimental analysis](#). *Cognition*, 120:380–390.
- Haerim Hwang. 2024. [Wanna contraction in first language acquisition, child second language acquisition, and adult second language acquisition](#). *Bilingualism: Language and Cognition*, 27:322–333.
- Haerim Hwang. 2026. Can large language models be viewed as a cognitive model of human language? not yet, regardless of reasoning capability and size. *Acta Psychologica*, page Article 107025.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.
- Carina Kauf and Anna Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 925–935, Florence, Italy. Association for Computational Linguistics.
- Nur Lan, Emmanuel Chemla, and Roni Katzir. 2026. [Large language models and the argument from the poverty of the stimulus](#). *Linguistic Inquiry*, 57:315–342.
- John D. Lewis and Jeffrey L. Elman. 2001. Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proceedings of the 26th annual Boston University conference on language development*, pages 359–370, Boston, MA. Cascadilla Press.
- Jeffrey Lidz and Annie Gagliardi. 2015. [How nature meets nurture: Universal grammar and statistical learning](#). *Annual Review of Linguistics*, 11:333–353.
- Tal Linzen and Marco Baroni. 2021. [Syntactic structure from deep learning](#). *Annual Review of Linguistics*, 7:195–212.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. [Assessing the ability of lstms to learn syntax-sensitive dependencies](#). *Transactions of the Association for Computational Linguistics*, 4:521–535.
- Alessandro Lopopolo, Stefan L. Frank, Antal van den Bosch, and Roel Willems. 2019. [Dependency parsing with your eyes: Dependency structure predicts eye regressions during reading](#). In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 77–85.
- Ilya Loshchilov and Frank Hutter. 2017. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*, pages 195–212.
- Brian MacWhinney and Lawrence Erlbaum. 2000. *The CHILDES project . Volume I: Tools for analyzing talk: Transcription format and programs*. Lawrence Erlbaum, Mahwah, NJ.
- Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.
- Andrew McInerney. 2025. Transformer modeling of syntactic island conditions across training. *University of Pennsylvania Working Papers in Linguistics*, 31:119–128.
- Kanishka Misra. 2022. [minicons: Enabling flexible behavioral and representational analyses of transformer language models](#). *arXiv*:2203.13112.
- Elliot Murphy. 2025. [Rose: A universal neural grammar](#). *Cognitive Neuroscience*, 16:49–80.
- Byung-Doh Oh and William Schuler. 2024. [Leading whitespaces of language models’ subword vocabulary pose a confound for calculating word probabilities](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472, Miami, Florida. Association for Computational Linguistics.
- Satoru Ozaki, Daniel Yurovsky, and Lauren Levin. 2022. How well do lstm language models learn filler-gap dependencies? In *Proceedings of the Society for Computation in Linguistics 2022*, pages 76–88. Association for Computational Linguistics.
- William O’Grady, Michiko Nakamura, and Yasuko Ito. 2008. [Want-to contraction in second language acquisition: An emergentist approach](#). *Lingua*, 118:478–498.

- Francesca Padovani, Jaap Jumelet, Yevgen Matuskevych, and Arianna Bisazza. 2025. [Child-directed language does not consistently boost syntax learning in language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19735–19756, Suzhou, China. Association for Computational Linguistics.
- Colin Phillips and Lara Ehrenhofer. 2015. [The role of language processing in language acquisition](#). *Linguistic Approaches to Bilingualism*, 5:409–453.
- Steven Piantadosi. 2023. [Modern language models refute chomsky’s approach to language](#). In *From fieldwork to linguistic theory: A tribute to Dan Everett*, pages 353–414. Language Science Press.
- Martin J. Pickering and Matthew J. Traxler. 1998. [Plausibility and recovery from garden paths: An eye-tracking study](#). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24:940–961.
- Tiago Pimentel and Clara Meister. 2024. [How to compute the probability of a word](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). *openai*. pages 1–24.
- Raj Sanjay Shah, Khushi Bhardwaj, and Sashank Varma. 2024. [Development of cognitive intelligence in pre-trained language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9632–9657, Miami, FL. Association for Computational Linguistics.
- Kenneth G. Shipley and Julie G. McAfee. 2015. *Assessment in speech-language pathology: A resource manual*. Plural Publishing, San Diego, CA.
- Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. [Blimp-nl: A corpus of dutch minimal pairs and acceptability judgments for language model evaluation](#). *Computational Linguistics*, 51:1267–1301.
- Rosalind J. Thornton. 1990. *Adventures in long-distance moving: The acquisition of complex wh-questions (Unpublished PhD dissertation)*. University of Connecticut, Storrs, CT.
- Michael Tomasello. 2003. *Constructing a language: a usage-based theory of language acquisition*. Harvard University Press, Cambridge, MA.
- 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.
- Charles D. Yang. 2004. [Universal grammar, statistics or both?](#) *Trends in Cognitive Sciences*, 8:451–456.
- Charles D. Yang. 2016. *The price of linguistic productivity: How children learn to break the rules of language*. MIT Press, Cambridge, MA.
- Charles D. Yang, Stephen Crain, Robert C. Berwick, Noam Chomsky, and Johan J. Bolhuis. 2017. [The growth of language: Universal grammar, experience, and principles of computation](#). *Neuroscience Biobehavioral Reviews*, 81:103–119.
- Liu Zhuang, Lin Wayne, Shi Ya, , and Zhao Jun. 2021. [A robustly optimized bert pre-training approach with post-training](#). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.
- Andrea Zukowski and Jaiva Larsen. 2011. [Wanna contraction in children: Retesting and revising the developmental facts](#). *Language Acquisition*, 18:211–241.

Appendices

A Model configurations in the study

The model configurations in the present study are summarized in Table 1. Note that the model setup aims to approximate the bundle of changes that co-occur across development in children’s input environment. In naturalistic contexts, older children are exposed to more language input, possess larger vocabularies, and likely have access to richer representational resources for language acquisition. Consequently, the models here therefore represent cross-sectional simulations of development, rather than checkpoint-based alignments of a single learner’s development.

B Dataset construction

The datasets for this study were created through two main steps: data extraction followed by pre-processing. First, child-directed input for baby LM training was extracted from CHILDES (MacWhinney and Erlbaum, 2000) in R using the "childesr" package. To ensure that only data from children living in the United States were included, we used the following command: `collection = "Eng-NA"`.

The extracted files, which were in .cha format, were then preprocessed in Python to obtain the child’s age from the @ID: lines. Each utterance was cleaned by removing CHAT annotations, such as those indicating body language or other annotations, and unwanted symbols. Only utterances produced by non-child speakers were retained, so

	Linear causal model	Bidirectional masked model
Architecture	GPT-2	RoBERTa
Layers		8
Attention heads		8
Intermediate size		2048
Max sequence length		512
Total parameters	10.88M for input data to 3-year-olds; 11.16M for input data to 4-year-olds; 13.21M for input data to 5-year-olds; 13.98M for input data to 6-year-olds; 15.77M for input data to 7-year-olds	11.18M for input data to 3-year-olds; 11.75M for input data to 4-year-olds; 15.85M for input data to 5-year-olds; 17.39M for input data to 6-year-olds; 20.98M for input data to 7-year-olds
Learning rate		0.0001
Scheduler type		linear
Training batch size		16
Evaluation batch size		16
Gradient accumulation step		2
Vocabulary size and input data	900 for input data to 3-year-olds; 2,000 for input data to 4-year-olds; 10,000 for input data to 5-year-olds; 13,000 for input data to 6-year-olds; 20,000 for input data to 7-year-olds	
Random seeds		13, 30, 42

Table 1: Model configurations in the study

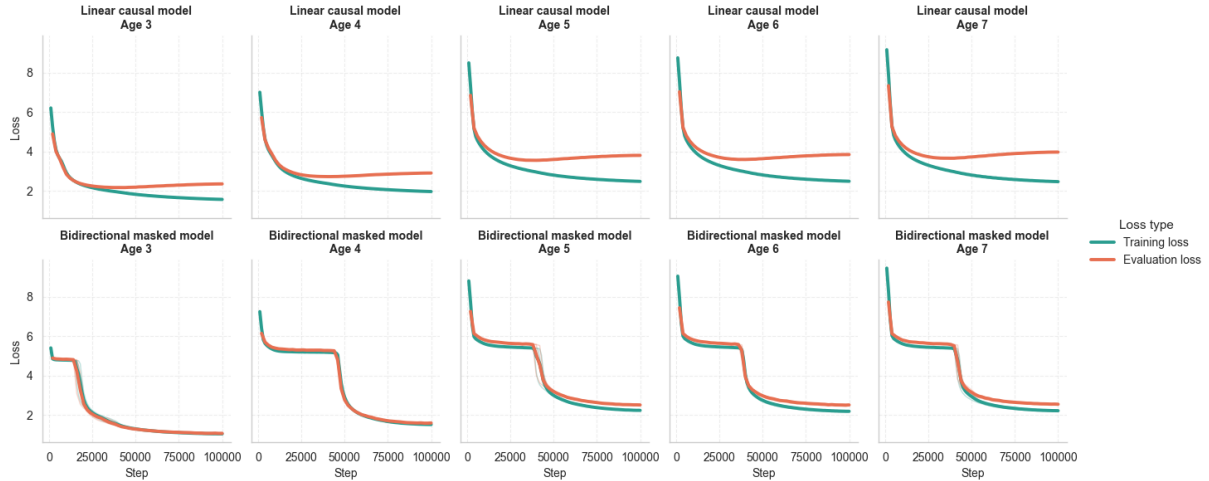


Figure 3: Training and evaluation loss curves

that the resulting dataset reflected child-directed input. For each retained utterance, metadata were recorded, including age, speaker code, corpus, subfolder, transcript filename, file path, and the cleaned utterance text. The complete dataset was saved as "CHILDES_sorted.csv", and cumulative age-based datasets in CSV files were then generated for all utterances available up to ages 3, 4, 5, 6, 7, and 8. The sizes of these datasets, in terms of word tokens and utterances, are summarized in Table 2.

Dataset	Word tokens	Utterances
Input up to age 3	790,259	4,189,835
Input up to age 4	1,007,504	5,432,165
Input up to age 5	1,334,574	7,379,696
Input up to age 6	1,423,082	7,919,487
Input up to age 7	1,468,227	8,188,443

Table 2: Size of five datasets extracted from CHILDES and used for training baby language models

C Training and evaluation loss curves of the baby language models

Training and validation losses were logged at 100-step intervals. The resulting loss curves for the models trained in this study are shown in Figure 3. Training a single model took about 1 hour and 50 minutes on an RTX 5090.

D Experimental stimuli

Below is the full set of experimental sentences employed in this study.

- (1) a. I wonder who you wanna read with.
b. *I wonder who you wanna read.
c. *I wonder if you wanna read with.
d. I wonder if you wanna read.
- (2) a. I wonder who you wanna dance with.
b. *I wonder who you wanna dance.
c. *I wonder if you wanna dance with.
d. I wonder if you wanna dance.
- (3) a. I wonder who you wanna sing with.
b. *I wonder who you wanna sing.
c. *I wonder if you wanna sing with.
d. I wonder if you wanna sing.
- (4) a. I wonder who you wanna walk with.
b. *I wonder who you wanna walk.
c. *I wonder if you wanna walk with.
d. I wonder if you wanna walk.
- (5) a. I wonder who you wanna stay with.
b. *I wonder who you wanna stay.
c. *I wonder if you wanna stay with.
d. I wonder if you wanna stay.
- (6) a. I wonder who you wanna swim with.
b. *I wonder who you wanna swim.
c. *I wonder if you wanna swim with.
d. I wonder if you wanna swim.
- (7) a. I wonder who you wanna run with.
b. *I wonder who you wanna run.
c. *I wonder if you wanna run with.
d. I wonder if you wanna run.
- (8) a. I wonder who you wanna smile with.
b. *I wonder who you wanna smile.
c. *I wonder if you wanna smile with.
d. I wonder if you wanna smile.
- (9) a. I wonder who you wanna start with.
b. *I wonder who you wanna start.
c. *I wonder if you wanna start with.
d. I wonder if you wanna start.
- (10) a. I wonder who you wanna talk with.
b. *I wonder who you wanna talk.
c. *I wonder if you wanna talk with.
d. I wonder if you wanna talk.
- (11) a. I wonder who you wanna eat with.
b. *I wonder who you wanna eat.
c. *I wonder if you wanna eat with.
d. I wonder if you wanna eat.
- (12) a. I wonder who you wanna travel with.
b. *I wonder who you wanna travel.
c. *I wonder if you wanna travel with.
d. I wonder if you wanna travel.
- (13) a. I wonder who you wanna drink with.
b. *I wonder who you wanna drink.
c. *I wonder if you wanna drink with.
d. I wonder if you wanna drink.
- (14) a. I wonder who you wanna live with.
b. *I wonder who you wanna live.
c. *I wonder if you wanna live with.
d. I wonder if you wanna live.
- (15) a. I wonder who you wanna cry with.
b. *I wonder who you wanna cry.
c. *I wonder if you wanna cry with.
d. I wonder if you wanna cry.
- (16) a. I wonder who you wanna laugh with.

- b. *I wonder who you wanna laugh.
 c. *I wonder if you wanna laugh with.
 d. I wonder if you wanna laugh.
- (17) a. I wonder who you wanna wait with.
 b. *I wonder who you wanna wait.
 c. *I wonder if you wanna wait with.
 d. I wonder if you wanna wait.
- (18) a. I wonder who you wanna shout with.
 b. *I wonder who you wanna shout.
 c. *I wonder if you wanna shout with.
 d. I wonder if you wanna shout.
- (19) a. I wonder who you wanna pray with.
 b. *I wonder who you wanna pray.
 c. *I wonder if you wanna pray with.
 d. I wonder if you wanna pray.
- (20) a. I wonder who you wanna work with.
 b. *I wonder who you wanna work.
 c. *I wonder if you wanna work with.
 d. I wonder if you wanna work.

E Instructions for child participants

The specific instructions that the child participants were presented with were as follows: "This is a 'smiley face' choice task. In this task, you will decide whether sentences are good/possible or bad/impossible in English. First, you will see a sentence on the screen while hearing that same sentence twice on the headphones. After you hear the sentence for the second time, a smiley face scale will pop up on the screen. It is at this point that you can press a number on the keyboard to rate the sentence on a 4-point smiley face scale with an additional "I don't know" option."

F Results of the linear mixed-effects regression for the entire data

Table 3 summarizes the results of the linear mixed-effects regression constructed for the entire data, including Clause, Gap, Learner type, and Age as fixed effects, as well as Item as a random effect.

Table 3: Linear mixed-effects regression results for the entire data

Predictor	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
(Intercept)	0.02	0.11	0.18	.857
Clause	0.43	0.15	2.90	.004
Gap	0.24	0.15	1.57	.117
Learner type	−0.04	0.15	−0.28	.781
Age	0.00	0.01	−0.32	.753
Clause × Gap	−1.06	0.29	−3.59	< .001
Clause × Learner type	1.29	0.29	4.37	< .001
Gap × Learner type	−0.16	0.30	−0.53	.599
Clause × Age	0.00	0.03	−0.14	.893
Gap × Age	0.04	0.03	1.68	.092
Learner type × Age	0.01	0.03	0.32	.753
Clause × Gap × Learner type	2.10	0.59	3.56	< .001
Clause × Gap × Age	0.40	0.05	7.77	< .001
Clause × Gap × Learner type	−0.13	0.05	−2.51	.012
Gap × Age × Learner type	0.24	0.05	4.57	< .001
Clause × Gap × Age × Learner type	−0.62	0.10	−6.05	< .001

G Results of the linear mixed-effects regression for the child data with the Age factor added

Table 4 summarizes the results of the linear mixed-effects regression constructed for the child data, including Clause, Gap, and Age as fixed effects, as well as Item as a random effect.

Table 4: Linear mixed-effects regression results for the child data with the Age factor added

Predictor	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
(Intercept)	0.05	0.17	0.29	.769
Clause	-0.25	0.32	-0.78	.438
Gap	0.38	0.33	1.17	.244
Age	-0.01	0.03	-0.35	.730
Clause × Gap	-2.52	0.65	-3.90	< .001
Clause × Age	0.07	0.06	1.21	.227
Gap × Age	-0.08	0.06	-1.50	.134
Clause × Gap × Age	0.79	0.11	7.05	< .001

H Results of the linear mixed-effects regression for the baby language model data with the Age factor added

Table 5 summarizes the results of the linear mixed-effects regression constructed for the baby LM data, including Clause, Gap, and Age as fixed effects, as well as Item as a random effect.

Table 5: Linear mixed-effects regression results for the baby language model data with the Age factor added

Predictor	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
(Intercept)	0.00	0.12	0.00	1.000
Clause	1.07	0.09	12.41	< .001
Gap	0.16	0.11	1.52	.132
Age	0.00	0.01	0.00	1.000
Clause × Gap	-0.01	0.17	-0.06	.952
Clause × Age	-0.07	0.02	-4.11	< .001
Gap × Age	0.16	0.02	9.73	< .001
Clause × Gap × Age	0.09	0.03	2.67	.008

I Results of the linear mixed-effects regression for the baby language model data with the Model architecture factor added

Table 6 summarizes the results of the linear mixed-effects regression constructed for the baby LM data, including Clause, Gap, and Model architecture as fixed effects, as well as Item as a random effect.

Table 6: Linear mixed-effects regression results for the baby language model data with the Model architecture factor added

Predictor	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
(Intercept)	0.00	0.11	0.00	1.000
Clause	0.73	0.02	30.75	< .001
Gap	0.97	0.07	14.08	< .001
Model architecture	0.00	0.02	0.00	1.000
Clause × Gap	0.43	0.05	9.13	< .001
Clause × Model architecture	0.19	0.05	4.08	< .001
Gap × Model architecture	-0.32	0.05	-6.69	< .001
Clause × Gap × Model architecture	-0.19	0.09	-1.97	.049