

How Far Is the Training Corpus from Real Infant Input? Auditing BabyLM Against Day-Long Home Recordings, and What an Age-of-Acquisition Curriculum Does Not Buy

Imrul Huda

University of Chicago
imrulhuda@uchicago.edu

Abstract

The BabyLM Challenge asks models to learn from developmentally plausible data, but how closely the standard corpus resembles the language infants actually hear has not been measured. We compare the Strict-Small corpus against SEEDLingS, a corpus of day-long home recordings from 44 infants with 358,300 hand-annotated noun tokens, over the 299 words the two share. Corpus word frequencies match real infant input only moderately (Spearman’s $\rho = 0.42$; all correlations reported here are Spearman rank correlations), and real input frequency predicts human age of acquisition (AoA) far better ($\rho = -0.79$) than the training corpus’s own frequencies do ($\rho = -0.48$).

Breaking the corpus down by source locates the gap. CHILDES tracks real infant input at $\rho = 0.55$ while the five other sources fall between 0.14 and 0.31, and the pooled corpus falls below CHILDES alone. The sources are not ordered by how speech-like they are: written text ranks above adult conversation. What separates them is child-directedness.

Against this we set a curriculum intervention. We order the corpus by the AoA of the words each sentence contains, derived from the MacArthur–Bates Communicative Development Inventory (CDI), train GPT-2 Small on it and on a content-matched shuffled baseline, across three seeds and three training budgets up to 100M words. The curriculum reliably produces a model whose acquisition order matches the human one ($\rho = 0.63$ – 0.78 against 0.09 – 0.22), but the curriculum encodes human AoA at $\rho = 0.96$ by construction and the alignment falls to $\rho \approx 0.3$ once presentation order is partialled out. It confers no advantage on BLiMP at any budget. Improving what a corpus contains may matter more than the order in which it is presented.

1 Introduction

A central goal of the BabyLM Challenge is to close the data-efficiency gap between human and machine language learners by using developmentally plausible training corpora (Warstadt et al., 2023; Hu et al., 2024; Charpentier et al., 2025). Most gains so far have come from changes to model architectures or training objectives rather than to the data itself (Charpentier et al., 2025). One exception is the use of variation sets—consecutive rephrasings common in child-directed speech—which improved some evaluation scores under certain conditions, though the effect varied across benchmarks (Haga et al., 2024). This motivates investigating other structural properties of child input.

We focus on the order in which words are learned. Age of acquisition (AoA) is a well-documented regularity in early word learning: children tend to acquire concrete, frequent words such as *mama*, *ball*, and *dog* before abstract words, and early vocabulary is strikingly consistent across children (Frank et al., 2017). The BabyLM 2026 evaluation suite includes an AoA task (Choshen et al., 2026) based on the method of Chang and Bergen (2022). This method tracks word surprisal across training checkpoints to estimate a model’s implicit AoA and compares the resulting order with human norms from the MacArthur–Bates Communicative Development Inventory (CDI). The task therefore allows us to test directly whether a vocabulary curriculum shifts a model’s learning trajectory toward the human order.

This paper makes two contributions. First, we audit the developmental plausibility of the Strict-Small corpus itself, comparing its lexical statistics against SEEDLingS, a corpus of real infant input (Bergelson and Aslin, 2017; Bergelson et al., 2019; Kalenkovich et al., 2025). The comparison

requires no model and no training. We find that the corpus reproduces real input frequency only moderately, that real input frequency is a far better predictor of human AoA than any in-corpus measure, and, breaking the corpus down by source, that a single child-directed component carries most of what fidelity the corpus has while pooling the remaining sources reduces it.

Second, we test the intervention this audit bears on. We construct an AoA-staged curriculum, train it against a content-matched shuffled baseline across three seeds and three training budgets, and measure where each model’s word-learning order falls relative to children’s. The curriculum reliably reproduces human AoA order; we then show that most of this is the ordering reproducing itself, and that none of it reaches grammatical competence. Prior work using distributional measures has found that language model learning trajectories do not track children’s (Ficarra et al., 2025); our result is compatible with that once the mechanical component is removed.

Read together, the two contributions suggest that the composition of a developmentally plausible corpus is a more consequential design choice than the order in which its contents are presented.

2 Related Work

Curriculum-learning approaches have appeared repeatedly across BabyLM iterations but have not produced consistently strong aggregate results (Warstadt et al., 2023; Hu et al., 2024; Charpentier et al., 2025). CLIMB (Diehl Martinez et al., 2023) explored vocabulary, data and objective curricula together and reported no reliable gain over a non-curriculum baseline; Salhan et al. (2024) applied cognitively motivated curricula across languages with similarly mixed results. ConcreteGPT (Capone et al., 2024) ordered training data according to lexical concreteness ratings (Brysbaert et al., 2014), but it did not derive its ordering from child acquisition data or organize the data into developmental stages. Variation sets, by contrast, improved some benchmarks—including BLiMP and GLUE, but not EWoK—depending on their proportion in the training data (Haga et al., 2024). Our curriculum instead uses lexical acquisition order derived directly from CDI production norms, and our finding that it confers no benchmark advantage is consistent with this broader pattern.

For evaluation, Chang and Bergen (2022) intro-

duced a method for deriving model-internal AoA from surprisal trajectories across training checkpoints, and established that both model and human AoA are strongly conditioned by word frequency. The 2026 challenge adopts this method as an AoA task evaluated against CDI norms (Frank et al., 2017; Choshen et al., 2026). We apply the same core procedure using context sentences drawn from the training corpus itself rather than the evaluation pipeline’s released context set (§6). Ficarra et al. (2025) take a distributional approach to the same question and report that model learning trajectories do not correlate with children’s. Our contribution is not to revisit whether such alignment exists, but to ask what a corpus intervention can do about it, and to measure the corpus these interventions are applied to against a record of what infants actually hear.

We do not compare our models with the field-wide competition baseline (Charpentier and Samuel, 2024). Because our research question concerns the effect of word order within an otherwise fixed data-and-architecture setup, our control is a model with the same architecture and training content but a shuffled presentation order.

3 Method

3.1 AoA Curriculum and Baseline

We derive human AoA norms from the CDI production data distributed with the BabyLM 2026 evaluation package (Frank et al., 2017). The dataset provides, for each of 504 words, the proportion of children producing that word at each age from 16 to 30 months. We define a word’s human AoA as the earliest age at which production reaches 50%. Words that never reach 50% are assigned an AoA of 31 months as a proxy for late acquisition, mirroring the 50%-of-range criterion later used for the model.

We score every sentence in the six Strict-Small source files—CHILDES, BNC-Spoken, Switchboard, Gutenberg, OpenSubtitles, and Simple Wikipedia; 1,104,106 sentences in total—using the mean AoA of the words covered by the CDI. Sentences containing no CDI words are assigned an AoA of 31.

We group the scored sentences into five stages based on mean sentence AoA (Table 1) and concatenate them in stage order. We repeat the earlier stages to up-weight simpler vocabulary, reflecting its prominence in early child-directed speech.

Stage	AoA range	Sentences	Reps.
1	≤ 18 mo.	13,314	$\times 3$
2	19–21 mo.	33,922	$\times 2$
3	22–24 mo.	99,499	$\times 1$
4	25–27 mo.	192,727	$\times 1$
5	≥ 28 mo.	764,644	$\times 1$

Table 1: Curriculum stages based on mean sentence AoA. Stage 5 contains 69% of all sentences. Because the CDI covers only 504 words, many lexically simple sentences are assigned to this stage solely because they contain no CDI word (§6).

Within each stage, sentences are shuffled, so that stage membership is the only ordered variable and the curriculum does not also order the data by source corpus. The resulting curriculum contains 1,164,656 lines and 10,421,796 words, drawn from a unique corpus of exactly 10M words.

To isolate the effect of presentation order, the flat-shuffle baseline contains the identical multi-set of lines, including the same sentences and per-stage repetitions, but is shuffled randomly using seed 1337. The two conditions therefore have identical word frequencies, vocabularies, and corpus sizes.

3.2 Model and Trajectory Evaluation

We train GPT-2 Small models with 124M parameters from scratch on each corpus with the standard GPT-2 tokenizer and AdamW. Training uses a cosine learning-rate schedule with 1% warmup, a learning rate of 5×10^{-4} , and an effective batch size of 32 sequences of 512 tokens. Each corpus is presented in its own fixed order, so that presentation order is the only respect in which the two conditions differ.¹ Every condition is trained with three random seeds (42, 43, 44), and we report the mean and standard deviation across seeds rather than a single run.

We report three training configurations. The first trains for one full pass over the Strict-Small curriculum (1,056 steps). The second repeats that curriculum for ten epochs (10,556 steps), and the third trains for one full pass over the Strict 100M curriculum (10,814 steps). The second and third hold total optimisation roughly constant while varying whether the additional compute buys repetition or new data. Word budgets are counted in

¹The version of this work submitted for review used a training script that shuffled the corpus at load time, which neutralised the ordering manipulation in both conditions. All results reported here use the corrected configuration.

words rather than tokens, using each corpus’s measured tokens-per-word ratio, and checkpoints are saved at the word-exposure intervals the challenge requires (Choshen et al., 2026).

For each CDI word, we sample up to 20 corpus sentences containing that word. Of the 504 CDI words, 502 occur in at least three corpus sentences; *vagina* and *teddybear* are excluded. At each checkpoint, we compute the summed teacher-forced surprisal, in bits, of the word’s subword tokens and average it across sentences to obtain a learning curve for each word.

Because a model’s surprisal on a word falls as it learns that word, the checkpoint at which the surprisal drops marks the model’s acquisition of it. We define a word’s model AoA as the first checkpoint at which its surprisal falls halfway between its value at initialisation and its minimum value. This mirrors the CDI’s 50%-of-range criterion, allowing the human and model measures to be defined analogously. We take the reference point at initialisation rather than at the first training checkpoint because a curriculum has, by then, already taught the words it front-loads: their surprisal has little room to fall further, so they never cross the threshold and are recorded as acquired last when they were in fact acquired first. An untrained reference makes the criterion well defined for every word regardless of when it is taught. Words that never cross the threshold are censored one checkpoint past the last. We report the Spearman correlation between model and human AoA, as well as a partial correlation controlling for log corpus frequency, a known confound in both measures. We treat $\alpha = 0.05$ as the significance threshold throughout.

3.3 SEEDLingS Ecological Validation

To assess how closely the training corpus reflects real infant experience, we compare it with SEEDLingS, a longitudinal corpus of day-long home recordings from 44 infants between 6 and 17 months of age (Bergelson and Aslin, 2017; Bergelson et al., 2019). Its publicly released noun-annotation dataset (Kalenkovich et al., 2025) contains 358,300 hand-annotated noun tokens with lemmatized `global_basic_level` labels. We use each lemma’s token count as a measure of its frequency in real infant input. This comparison requires no model and no training: it relates two frequency distributions over the same vocabulary.

Matching these lemmas to the CDI words in

our analysis yields 299 words with at least 10 SEEDLingS tokens. Over this set, we calculate three Spearman correlations: the correlation between training-corpus frequency and real input frequency; the correlation between human AoA and real input frequency; and the correlation between human AoA and training-corpus frequency. All analyses use log frequencies. Corpus frequency is the number of sentences containing the word. Lines in the CHILDES portion begin with a speaker code, which the tokenizer would otherwise read as an ordinary word, so these prefixes are stripped before counting.

We then repeat the same three correlations separately for each of the six source corpora, which asks not whether the corpus as a whole diverges from real infant input but which of its components do. Because the sources differ in size, we report the number of matched words behind each correlation. We run this breakdown on both the Strict-Small corpus and the larger Strict corpus, where the same words are estimated from roughly ten times as much text.

4 Results

4.1 How Far Is the Corpus from Real Infant Input?

Table 2 presents the SEEDLingS comparison across the 299 matched CDI nouns. Two findings stand out.

First, word frequencies in the Strict-Small corpus correlate only moderately with frequencies in real infant input ($\rho = 0.42$). The corpus therefore approximates the overall scale of early linguistic experience but not its lexical distribution.

Second, real input frequency predicts human AoA much more strongly ($\rho = -0.79$, $p < 10^{-63}$) than the training corpus’s own frequencies do for the same words ($\rho = -0.48$). The frequency signal most strongly associated with children’s acquisition order, how often they hear a word, is therefore only weakly represented in the corpus from which a curriculum must be built.

4.2 Which Sources Diverge?

Table 3 repeats the comparison for each source corpus separately. Two further findings stand out.

First, CHILDES stands apart. Its word frequencies correlate with real infant input at $\rho = 0.55$ on the Strict corpus, against 0.27 ± 0.06 for the five remaining sources. The same ordering holds

Comparison ($n = 299$ nouns)	ρ	p
Corpus vs. real input freq.	0.42	5.8×10^{-14}
Human AoA vs. real input freq.	-0.79	4.6×10^{-64}
Human AoA vs. corpus freq.	-0.48	3.2×10^{-18}

Table 2: Ecological validation against real infant input. Values are Spearman correlations of log frequencies across CDI nouns matched to a SEEDLingS lemma with at least 10 tokens. The corpus only moderately reproduces real input frequency (row 1), while real input frequency is much more strongly associated with human AoA than corpus frequency is (rows 2 and 3).

Source	Strict-Small		Strict	
	n	ρ	n	ρ
CHILDES	298	0.48	298	0.55
Simple Wikipedia	261	0.29	298	0.29
OpenSubtitles	291	0.28	298	0.30
Gutenberg	264	0.25	293	0.31
BNC-Spoken	234	0.20	282	0.29
Switchboard	63	0.09	192	0.14
All pooled	299	0.42	299	0.45

Table 3: Correlation between each source corpus’s word frequencies and real infant input frequency, computed separately per source. n is the number of CDI nouns clearing the matching threshold in that source. CHILDES is markedly closer to real infant input than any other source, and the pooled corpus falls below CHILDES alone.

on Strict-Small, and CHILDES frequencies also predict human AoA far better ($\rho = -0.63$) than any other source. This is what one would expect of the only component that is itself child-directed speech.

Second, pooling lowers fidelity rather than raising it. The full corpus reproduces real input frequency at $\rho = 0.45$, below CHILDES alone at $\rho = 0.55$. Adding the non-child-directed sources therefore moves the corpus away from the distribution infants receive, even as it adds data.

The remaining five sources are not ordered by how speech-like they are. Written text (Gutenberg, $\rho = 0.31$) sits above adult conversation (BNC-Spoken, $\rho = 0.29$), and the lowest value belongs to a spoken corpus rather than a written one. What separates the sources is child-directedness, not spokenness. We note that Switchboard is the smallest source and clears the matching threshold for only 63 words on Strict-Small, so its value there is the least reliable in the table.

4.3 Does Ordering the Corpus Help?

Table 4 reports the curriculum and baseline models across the three training configurations. In every configuration the curriculum-trained model’s acquisition order aligns closely with human AoA order, and the flat-shuffle baseline’s does not. On one pass over Strict-Small the curriculum reaches $\rho = 0.77 \pm 0.02$ against 0.09 ± 0.03 for the baseline; the gap persists at ten epochs (0.63 ± 0.02 against 0.16 ± 0.02) and on the 100M corpus (0.78 ± 0.01 against 0.22 ± 0.02). Seed variability is small throughout, and the difference is significant at the level of training runs in all three configurations ($p < 10^{-5}$).

This alignment is, however, largely mechanical. The curriculum is constructed by ordering sentences by the AoA of the words they contain, so the position at which a word first appears in the training stream is itself almost perfectly correlated with its human AoA ($\rho = 0.96$). The model in turn acquires words roughly in the order it encounters them ($\rho = 0.78$ on one pass). Table 5 partials out corpus position: the residual alignment with human AoA falls to $\rho = 0.32$, and takes similar values at ten epochs (0.25) and on 100M words (0.29). Most of what the headline correlation measures is therefore the curriculum reproducing the ordering it encodes, though a smaller association survives that presentation order alone does not explain.

The ordering also does not transfer to grammatical competence. On BLiMP the curriculum and baseline models are indistinguishable in every configuration, with differences of -1.00 , -0.10 and -0.19 points (all $p > 0.16$; Appendix A). The differences are uniformly small, uniformly negative, and shrink as training increases. We find no evidence that ordering the corpus by age of acquisition improves what the model learns about language, at any of the three budgets.

5 Discussion

Taken together, the two analyses point the same way. Ordering the corpus by age of acquisition reliably produces a model whose acquisition order matches the human one, but most of that match is the ordering reproducing itself, and none of it reaches grammatical competence. Meanwhile the corpus being ordered is only a moderate approximation of the language infants actually hear. The lever that was available to us turns out to matter

Corpus / schedule	Condition	ρ	partial
10M, 1 pass	Curriculum	0.77 ± 0.02	0.82
	Flat-shuffle	0.09 ± 0.03	-0.06
10M, 10 epochs	Curriculum	0.63 ± 0.02	0.58
	Flat-shuffle	0.16 ± 0.02	0.03
100M, 1 pass	Curriculum	0.78 ± 0.01	0.81
	Flat-shuffle	0.22 ± 0.02	0.09

Table 4: Spearman correlations between model AoA and human CDI AoA, as mean $\pm$ standard deviation over three random seeds. The partial column controls for log corpus frequency. In every configuration the curriculum-trained model aligns with human AoA order and the baseline does not.

	10M 1 pass	10M 10 ep.	100M 1 pass
Corpus position vs. human AoA	0.96	0.96	0.97
Model AoA vs. corpus position	0.78	0.53	0.79
Model AoA vs. human AoA	0.80	0.57	0.81
Residual, controlling position	0.32	0.25	0.29

Table 5: How much of the curriculum’s AoA alignment is attributable to presentation order. The curriculum encodes human AoA almost perfectly by construction (row 1) and the model largely follows the order it is given (row 2). Controlling for corpus position leaves a smaller residual association (row 4). All values are Spearman correlations over seed-averaged model AoA; $p < 10^{-8}$ for every residual.

less than the material it was applied to.

The sub-corpus breakdown locates the gap more precisely. It is not that the corpus is uniformly unlike infant input; it is that one component resembles it and the rest do not, and that pooling them together moves the whole further away. Nor is the difference explained by how speech-like a source is: written text ranks above adult conversation, and the lowest source is spoken. What distinguishes CHILDES is that it is addressed to children. This bears on how such corpora might be assembled. If the objective is a developmentally plausible corpus, the sampling weights across registers are a design parameter, and one that could be set against a measured distribution of real input rather than by availability.

We do not read this as an argument that child-directed speech should simply crowd out the other sources. Infants hear a great deal of speech not addressed to them, and day-long recordings capture that mixture as it actually occurs; this is precisely why SEEDLingS, rather than CHILDES alone, is the appropriate reference. The finding is not that

the corpus should contain more CHILDES, but that its mixture does not match the mixture infants receive, and that the mismatch is measurable.

Two caveats attach to the curriculum result. The first is that our curriculum is built from CDI norms and then evaluated against CDI norms, so a positive correlation is close to guaranteed; the position-controlled analysis is what separates the mechanical part from the remainder. The second is that the residual we do observe, at $\rho \approx 0.3$ across all three configurations, is not itself explained. It is stable under a tenfold change in either data or compute, which suggests it reflects something about the words rather than about the training budget, but we have not isolated what.

We therefore read the result as informative rather than encouraging. AoA curricula reproduce the order they encode and, on the evidence here, buy nothing beyond it. Improving the developmental fidelity of the training corpus itself may matter more than reordering its existing contents.

6 Limitations

The residual is unexplained. Partialling out corpus position leaves an association between model and human AoA of $\rho \approx 0.3$ in every configuration (Table 5). We do not know what produces it. It is stable under a tenfold change in either training data or compute, which argues against its being an artefact of training budget, but we have not established whether it reflects properties of the words themselves or a residual frequency effect that the position control does not absorb. The two controls we report, for frequency and for position, were applied separately rather than jointly.

Circular construction. The curriculum is built from CDI norms and evaluated against CDI norms, so some positive correlation is guaranteed by construction. This is why we report the position-controlled analysis alongside the raw correlation, and why we do not treat the headline value as evidence that the curriculum induces human-like learning. A design that ordered the corpus by a measure independent of the evaluation would test the underlying claim more directly.

Sparse and proxy norms. The CDI covers only 504 words. Consequently, any sentence containing none of these words is assigned to the latest AoA bin regardless of its actual complexity; such sentences constitute roughly 58% of the curricu-

lum, and the majority of stage 5 (Table 1). Extending the vocabulary coverage with the reading-AoA norms of Kuperman et al. (2012) could produce more precise later stages, but this extension fell outside the scope of the present study.

Operationalization of model AoA. Our definition—the first checkpoint at which surprisal crosses the 50%-of-range threshold from an untrained reference—is one of several possible choices. It may be sensitive to the number of context sentences used, up to 20, and to the particular contexts selected, which differ from the official pipeline’s fixed, released context set. We note that the choice of reference point is not innocuous: measuring the range from the first training checkpoint rather than from initialisation inverts the measure for precisely those words a curriculum front-loads (§3.2).

SEEDLingS matching. The ecological comparison is based on 299 CDI nouns matched by lemma to the SEEDLingS noun annotations (Kalenkovich et al., 2025). It therefore covers only nouns, the category annotated in the dataset, rather than the complete set of CDI words. Lemma-level token counts are also a coarse proxy for the frequency of a word in an individual child’s input. Nevertheless, the analysis draws on real longitudinal recordings from 44 infants rather than on corpus-derived proxies.

Uneven source sizes. The per-source breakdown estimates each correlation from the text available in that source, and the sources differ by two orders of magnitude in size. Switchboard clears the matching threshold for only 63 of the 299 words on Strict-Small, so its value there should be read as the least reliable in Table 3; the larger Strict corpus mitigates but does not remove this.

Scope. All experiments are in English. The curriculum results cover the Strict-Small and Strict corpora; we make no claims about the Multilingual track, and we do not compare against the official competition baseline, which we did not run.

Ethical Considerations

All training corpora come from the officially distributed BabyLM Strict-Small release. The CDI norms come from Wordbank’s public developmental data (Frank et al., 2017), and the SEEDLingS

noun annotations are available as a public, de-identified, CC-BY release (Kalenkovich et al., 2025). No new human-subject data were collected, and we did not access the raw recordings. We train and release only small, 124M-parameter models and make no claims regarding deployment.

Code and data availability. All corpus-construction, training and analysis code, together with the derived result files underlying every reported correlation, is available at <https://github.com/imrulhuda/babylm-aoa-audit>. The SEEDLingS noun annotations are obtained from the CC-BY release of Kalenkovich et al. (2025); the BabyLM corpora are used as distributed by the challenge organisers.

Acknowledgments

We thank the anonymous reviewers for comments that materially reshaped this paper, and the Bergelson Lab for releasing the SEEDLingS noun annotations that make the corpus comparison possible.

References

- Elika Bergelson, Andrei Amatuni, Shannon Dailey, Sharath Koorathota, and Shaelise Tor. 2019. [Day by day, hour by hour: Naturalistic language input to infants](#). *Developmental Science*, 22(1):e12715.
- Elika Bergelson and Richard N. Aslin. 2017. [Nature and origins of the lexicon in 6-mo-olds](#). *Proceedings of the National Academy of Sciences*, 114(49):12916–12921.
- Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. [Concreteness ratings for 40 thousand generally known English word lemmas](#). *Behavior Research Methods*, 46(3):904–911.
- Luca Capone, Alessandro Bondielli, and Alessandro Lenci. 2024. [ConcreteGPT: A baby GPT-2 based on lexical concreteness and curriculum learning](#). In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 189–196. Association for Computational Linguistics.
- Tyler A. Chang and Benjamin K. Bergen. 2022. [Word acquisition in neural language models](#). *Transactions of the Association for Computational Linguistics*, 10:1–16.
- Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams. 2025. [Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data](#). In *Proceedings of the First BabyLM Workshop*, pages 399–420. Association for Computational Linguistics.
- Lucas Georges Gabriel Charpentier and David Samuel. 2024. [GPT or BERT: Why not both?](#) In *Proceedings of the 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*.
- Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Jaap Jumelet, Tal Linzen, Aaron Mueller, Suchir Salhan, Raj Sanjay Shah, Alex Warstadt, and Ethan Wilcox. 2026. [BabyLM turns 4 and goes multilingual: Call for papers for the 2026 BabyLM workshop](#). *arXiv preprint arXiv:2602.20092*.
- Richard Diehl Martinez, Zébulon Goriely, Hope McGovern, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2023. [CLIMB – curriculum learning for infant-inspired model building](#). In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 112–127. Association for Computational Linguistics.
- Filippo Ficarra, Ryan Cotterell, and Alex Warstadt. 2025. [A distributional perspective on word learning in neural language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11184–11207, Albuquerque, New Mexico. Association for Computational Linguistics.
- Michael C. Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A. Marchman. 2017. [Wordbank: An open repository for developmental vocabulary data](#). *Journal of Child Language*, 44(3):677–694.
- Akari Haga, Akiyo Fukatsu, Miyu Oba, Arianna Bisazza, and Yohei Oseki. 2024. [BabyLM challenge: Exploring the effect of variation sets on language model training efficiency](#). In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 252–261. Association for Computational Linguistics.
- Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. [Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora](#). In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21. Association for Computational Linguistics.
- Evgenii Kalenkovich, Sharath Koorathota, Shaelise Tor, Andrei Amatuni, Shannon Egan-Dailey, Charlotte Moore, Catherine Laing, Hallie Garrison, Gladys Baudet, Federica Bulgarelli, Sarp Uner, Lilianna Righter, and Elika Bergelson. 2025. [A year](#)

of nouns from English-learning infants’ daily lives: The SEEDLingS-Nouns dataset. *Behavior Research Methods*, 57(11):298.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44:978–990.

Suchir Salhan, Richard Diehl Martinez, Zébulon Goriely, and Paula Buttery. 2024. Less is more: Pre-training cross-lingual small-scale language models with cognitively-plausible curriculum learning strategies. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 174–188. Association for Computational Linguistics.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34. Association for Computational Linguistics.

A Zero-Shot Linguistic Evaluation

Table 6 reports filtered BLiMP accuracy for every model discussed in §4.3: two conditions, three random seeds, and three training configurations. We report BLiMP alone; the remaining zero-shot tasks in the challenge suite were not rerun for these models.

Two observations follow. First, accuracy rises with training as expected, from 60 on a single pass over Strict-Small to 69 at ten epochs and 75 on the 100M corpus, against a chance level of 50. The models underlying the curriculum comparison are therefore trained well past the point at which BLiMP is uninformative.² Second, the curriculum and baseline models are indistinguishable at every budget. The differences are -1.00 , -0.10 and -0.19 points, none significant at $n = 3$ per condition ($p = 0.169, 0.850, 0.447$), and they narrow as training increases.

Corpus / sched.	Condition	s42	s43	s44	mean
10M, 1 pass	Curriculum	59.29	60.15	60.84	60.09
	Flat-shuffle	61.42	61.54	60.31	61.09
10M, 10 ep.	Curriculum	68.86	69.04	67.90	68.60
	Flat-shuffle	69.34	68.43	68.32	68.70
100M, 1 pass	Curriculum	75.40	75.10	74.97	75.16
	Flat-shuffle	75.11	75.70	75.22	75.34

Table 6: Filtered BLiMP accuracy (%) at the final checkpoint, by condition, seed and training configuration. Chance is 50. Accuracy increases with training budget, and the curriculum confers no advantage over the shuffled baseline in any configuration.

²For reference, the version of this work submitted for review reported 55.99 on a single, shorter run; see §3.2.