

Reading Acquisition Order out of BabyLM Representations: Child-Scale Training Produces Developmentally Aligned Semantic Spaces

Chenyi Ding
University of California, Davis

Zachary N. Houghton
University of Oregon
Vail Systems, Inc

Masoud Jasbi
University of California, Davis

Abstract

The age at which children learn different words, also known as Age of Acquisition (AoA), can be predicted from several word-level properties such as word frequency and concreteness. We ask whether acquisition order information is also implicitly present in the internal representations of language models, and what linguistic factors drive this. We probe 18 public checkpoints that span three architecture families at matched size, data scales from 10 million to roughly 180 billion words, and the winning systems and official baselines of the three previous BabyLM challenges. A linear ridge read-out of frozen hidden states predicts Wordbank age of acquisition for 619 words in the communicative development inventory (CDI), with the probing layer chosen by nested cross-validation. Models trained on child-scale data match or exceed their full-scale counterparts in every family. Masked and hybrid training objectives yield representations from which acquisition order is easier to recover; challenge winners do not best capture children’s acquisition order. Geometric measures of semantic structure add predictive value beyond classic predictors only for child-scale models. AoA decodability thus provides a training-free measure of how closely a model’s representations track children’s real acquisition order.

1 Introduction

Current estimates suggest that by the time they are six year old, children hear somewhere between 7 to 80 million words, depending on what kind of environment they grow up in (Frank, 2023). Yet, they master their native language long before modern language models trained on thousands of times more text (Warstadt et al., 2023). The BabyLM Challenge treats this “data gap” as a research opportunity: models trained on developmentally plausible amounts of text can serve both

as more efficient learners and as candidate models of human language acquisition (Charpentier et al., 2025). One concrete way to test the second role is to ask how much of children’s actual learning is recoverable from such models. Parent-reported vocabularies collected in the Wordbank repository show that the variance in acquisition order is reliable (Braginsky et al., 2016), and that small set of word-level properties account for a third of it.

Language models have entered this literature from the behavioral side. Chang and Bergen (2022) track the surprisal of individual words across training checkpoints and derive an age of acquisition for the model itself, and Portelance et al. (2023) show that language model surprisal predicts children’s age of acquisition across five languages, with the strongest effects for predicates and function words. The third BabyLM Challenge adopted the checkpoint-based method as one of its hidden evaluation tasks (Charpentier et al., 2025). These approaches predict acquisition from model behavior, either over training time or at the output layer. What the trained representation itself encodes about acquisition order has, to our knowledge, been examined only for single models and never compared across training regimes.

In this paper, we treat the decodability of children’s age of acquisition (AoA) from frozen representations as a measurable property of a trained model: how much of children’s real acquisition order its representations recover. We call this property “developmental alignment”, and we measure it across 18 public checkpoints, spanning three architecture families with matched parameter counts (OPT-125M, GPT-2, RoBERTa), training data from 10 million to roughly 180 billion words, the winning systems of the 2023 and 2024 challenges (ELC-BERT and GPT-BERT), and the official baselines of all three challenge years, including the 2025 GPT-BERT baselines that differ only in the ratio of masked to causal training.

For each checkpoint we aggregate hidden states for 619 words in the Communicative Development Inventory (CDI) (Frank et al., 2017) over child-directed sentences from CHILDES, train a ridge read-out against parent-reported age of acquisition (AoA) from the Wordbank repository under cross-validation with nested layer selection, and compare how much acquisition order each checkpoint’s representations recover.

Three architecture families tell the same story: checkpoints trained on child-scale corpora match or exceed counterparts trained on three to four orders of magnitude more text in AoA decodability. The effect is largest in the GPT-2 family, where the BabyLM-trained baseline reaches $\rho = .62$ against .48 for the full-scale model. Training objective matters as much as data. At identical data and architecture, the masked-focused and mixed GPT-BERT variants reach the highest scores overall ($\rho = .637$ and $.639$), while the purely causal variant trails them. The winners of previous challenges, in contrast, sit in the lower half of the ranking, below several plain baselines. Finally, three geometric measures of semantic neighborhood structure add predictive value over the classic word-level predictors of Braginsky et al. (2016) for child-scale models, and add essentially nothing for the full-scale OPT model.

Our contributions are threefold. First, we introduce AoA decodability as a training-free way to measure how closely a model’s representations track children’s real acquisition order (requiring a single frozen checkpoint rather than a full trajectory of intermediate checkpoints). Second, we provide the first systematic comparison of this property across data scales, architectures, and training objectives, including the winners and baselines of all previous BabyLM challenges. Third, we show that leaderboard performance and developmental alignment dissociate, which has consequences for how the community evaluates progress toward human-like learning.

2 Background

Acquisition order and lexical structure. Which words children learn first is not arbitrary. Parent-report vocabularies collected in the Wordbank repository show that acquisition order is stable across children, (split-half $r^2 > 0.98$; Braginsky et al. 2016), and that a small set of word-level properties accounts for roughly half of

the variance. Beyond these word-level predictors, the structure of the developing vocabulary also predicts acquisition order. Hills et al. (2009) show that early semantic networks grow by preferential acquisition, with new words entering according to their semantic relatedness to words already in the child’s vocabulary, and Fournassi et al. (2020) find that semantic-network connectivity predicts acquisition across languages. Measures defined over a semantic space as a whole, rather than over isolated word properties, therefore have a natural place in accounts of vocabulary growth. This is why in this study, we define three measures (cohesion, distinctiveness, and diversity) over the representation space of language models to test their predictive accuracy against more traditional measures such as frequency.

From behavior to representations. Previous work connecting language models to acquisition order has so far been behavioral. Portelance et al. (2020) predict CDI-based AoA from the average surprisal of an LSTM trained on child-directed speech, with gains over frequency for predicates and function words. Chang and Bergen (2022) define an age of acquisition for a model as the training step at which the surprisal of a word crosses a learning threshold, and the third BabyLM Challenge turned this method into an evaluation task (Charpentier et al., 2025). Portelance et al. (2023) extend surprisal-based prediction to five languages. A parallel line elicits norm-like judgments from instruction-tuned LLMs, including adult AoA ratings (Trott, 2024) and batteries of lexical features tested against child AoA (Hagihara and Miyazawa, 2026). Behavioral measures have a clear advantage: they require nothing beyond output probabilities. They also have two costs. Trajectory-based measures need a full series of intermediate checkpoints, which public models rarely provide, and output probabilities compress the model’s complex representation of a given input into a single number per word; Bunzeck and Zarri   (2025) show that in-context surprisal can hide exactly the lexical deficits it is meant to measure. Probing the representation avoids both. Following Bommasani et al. (2020), we represent each word by its hidden state averaged over occurrences in context, and we ask how much of children’s acquisition order a linear read-out recovers from these frozen states. In preliminary experiments with a single child-scale model,

this read-out recovered acquisition order far better than fine-tuning the same model end to end on a corpus-based AoA signal derived from children’s first productions in CHILDES (Smolík and Filip, 2022), so we keep every model frozen throughout.

Within the workshop, Capone et al. (2024) argue against age of acquisition as a criterion for ordering training data, because linguistically similar words of comparable difficulty can be acquired at very different ages. Our use is diagnostic rather than curricular: AoA is the quantity we decode from a trained model, not a signal we train on.

3 Methods

We evaluate 18 public checkpoints across three families. For each checkpoint, we extract contextual word representations for 619 CDI words, construct a four-predictor AoA baseline model, and then train a nested cross-validated ridge read-out on the frozen hidden states. We then define three geometric measures of semantic neighborhood structure, namely cohesion, distinctiveness, and diversity, capturing how tightly a word clusters with semantically related words (cohesion), how distinct it is from unrelated ones (distinctiveness), and how densely populated its neighborhood is (diversity), and test whether they predict AoA above and beyond the baseline predictors.¹

Models. Table 1 lists the 18 public checkpoints. The core of the set is three architecture families with matched data-scale pairs: OPT-125M, GPT-2, and RoBERTa. Each family pairs child-scale checkpoints with the original full-scale release. For OPT-125M (Zhang et al., 2022), the child-scale checkpoints are trained on the BabyLM corpus at 10M and 100M words (Warstadt et al., 2023) and on the roughly 150M-token BabyLM v3 corpus (Houghton et al., 2026); the full-scale reference is the original release trained on roughly 180B tokens. For GPT-2 (Radford et al., 2019), the child-scale checkpoints are the 2025 baselines trained on 10M and 100M words (Charpentier et al., 2025); the full-scale counterpart is the original WebText checkpoint. For RoBERTa (Liu et al., 2019), the child-scale checkpoint is the 2023 BabyLM baseline (Warstadt et al., 2023), set against the original release. Beyond these pairs, the set includes the 2023 winning system ELC-BERT (Charpentier and Samuel, 2023), the 2024

winning architecture GPT-BERT (Charpentier and Samuel, 2024), the LTG-BERT baseline (Samuel et al., 2023; Hu et al., 2024), the BabyLlama distillation line (Timiryasov and Tastet, 2023; Hu et al., 2024), and the three 2025 GPT-BERT baselines, which share data and architecture and differ only in the ratio of masked to causal language modeling (Charpentier et al., 2025).

Word representations. For each checkpoint we extract hidden states at every layer. In the contextual condition, each CDI word is read inside up to 50 child-directed sentences drawn from CHILDES (MacWhinney, 2000); the hidden states of its subword tokens are averaged within each sentence and then across sentences (Bommasani et al., 2020). In the bare condition the word is presented alone. Because the 18 models use different tokenizers, we locate each target word by its character position in the original sentence rather than matching tokens directly. All representations are extracted once; no model is updated at any point.

Acquisition target and baseline predictors. The target is production-based age of acquisition for 619 items of the English CDI of the English Words and Sentences CDI, spanning sound effects and early nouns acquired by 16 months (*woof woof, dog, mommy*) through function words not reliably produced until 30 months (*if, but, would*), defined as the age in months at which half of children are reported to produce the word (Frank et al., 2017). The reference model is the four-predictor regression of Braginsky et al. (2016): log frequency in child-directed speech, mean length of utterance, concreteness (Brysbaert et al., 2014), and association with babies, which they call babiness (Perry et al., 2015). Under our cross-validation this baseline reaches $\rho = .707$.

Read-out. For every checkpoint, layer, and representation type, we fit a ridge regression ($\lambda = 10$) that uses the frozen hidden states to predict AoA, under five-fold cross-validation, features z-scored within each training fold. We operationalize score as the Spearman correlation between pooled held-out predictions and observed AoA. We use a linear probe because less expressive probes limit the risk that the probe itself, rather than the representation, carries the signal (Hewitt and Liang, 2019; Belinkov, 2022). We select a single layer rather than averaging across all layers because different depths encode different kinds of information, and

¹Data and analysis code are available at <https://github.com/znoughton/aoa-pipeline>.

Checkpoint	Family	Training data	Objective	Source
opt125-babylm150m	OPT-125M	BabyLM v3 150M	causal	Houghton et al. (2026)
opt125-babylm100m	OPT-125M	BabyLM 100M	causal	Warstadt et al. (2023)
opt125-babylm10m	OPT-125M	BabyLM 10M	causal	Warstadt et al. (2023)
opt125-full	OPT-125M	~180B tokens	causal	Zhang et al. (2022)
gpt2-babylm100m	GPT-2	BabyLM 100M	causal	Charpentier et al. (2025)
gpt2-babylm10m	GPT-2	BabyLM 10M	causal	Charpentier et al. (2025)
gpt2-full	GPT-2	WebText (40 GB)	causal	Radford et al. (2019)
babylama58-2023	Llama	BabyLM 10M	causal + distillation	Timiryasov and Tastet (2023)
babylama-100m	Llama	BabyLM 100M	causal + distillation	Hu et al. (2024)
babylama-10m	Llama	BabyLM 10M	causal + distillation	Hu et al. (2024)
roberta-babylm100m	RoBERTa	BabyLM 100M	masked	Warstadt et al. (2023)
roberta-full	RoBERTa	160 GB text	masked	Liu et al. (2019)
elcbert-100m-2023	LTG-BERT	BabyLM 100M	masked	Charpentier and Samuel (2023)
gptbert-100m-2024	GPT-BERT	BabyLM 100M	hybrid	Charpentier and Samuel (2024)
ltgbert-100m-2024	LTG-BERT	BabyLM 100M	masked	Samuel et al. (2023)
gptbert25-mixed	GPT-BERT	BabyLM 100M	hybrid, 50% causal	Charpentier et al. (2025)
gptbert25-causal	GPT-BERT	BabyLM 100M	hybrid, 93.75% causal	Charpentier et al. (2025)
gptbert25-masked	GPT-BERT	BabyLM 100M	hybrid, 6.25% causal	Charpentier et al. (2025)

Table 1: The 18 checkpoints, grouped by architecture family. HuggingFace identifiers are listed in Appendix A.

averaging would mix a strong signal at one depth with noise from the rest. Because the best layer varies across models, the reported numbers use nested cross-validation: within each outer training fold, an inner five-fold loop selects the layer, which is then evaluated on the held-out outer fold. This keeps the layer choice from ever seeing the words it is scored on, avoiding the optimistic bias of picking the best layer post hoc. We report 95% confidence intervals from 2,000 bootstrap resamples over words, and we compare each checkpoint to the full-data OPT reference on their common valid words.

Semantic structure. From each representation space we compute three measures per word. Let $\mathbf{w}_i$ be the representation of word i , let $s(i, j)$ be the cosine similarity between $\mathbf{w}_i$ and $\mathbf{w}_j$, and let $C(i)$ be the set of other words sharing the CDI category of word i . Cohesion is the mean similarity of a word to its own category,

$$\text{coh}(i) = \frac{1}{|C(i)|} \sum_{j \in C(i)} s(i, j), \quad (1)$$

distinctiveness is cohesion minus the mean similarity to all words outside the category,

$$\text{dist}(i) = \text{coh}(i) - \frac{1}{|\bar{C}(i)|} \sum_{j \in \bar{C}(i)} s(i, j), \quad (2)$$

where $\bar{C}(i)$ contains every word outside the category of i , and diversity is the entropy of the full similarity distribution of word i , computed with-

out category labels,

$$\text{div}(i) = - \sum_{b=1}^B p_b(i) \log p_b(i), \quad (3)$$

where $p_b(i)$ is the proportion of the similarities $\{s(i, j) : j \neq i\}$ that fall into bin b of a B -bin histogram over $[-1, 1]$, with $B = 20$. Each set of measures is added to the four baseline predictors in the same cross-validated regression.

Robustness. Hidden layer sizes range from 512 to 768, so we repeat the read-out after projecting every space to 256 principal components; the ranking of checkpoints is essentially unchanged (Spearman $\rho = .922$ with the main analysis). Contextual representations outperform bare ones throughout, consistent with Bommasani et al. (2020), and the main text reports contextual results.

4 Results

We organize the results around four central findings. First, child-scale checkpoints match or exceed full-scale checkpoints within every architecture family. Second, training objective matters as much as data scale for how much of children’s acquisition order a model’s representations capture, with masked and hybrid objectives ahead of purely causal training. Third, winning previous BabyLM challenges does not predict how well a model captures children’s acquisition order. Fourth, a word’s semantic neighborhood structure in a model’s representation space carries information about the children’s acquisition order above

Checkpoint	ρ [95% CI]	Δ vs. ref.
gptbert25-mixed	.639 [.584, .685]	+.079*
gptbert25-masked	.637 [.585, .685]	+.078*
gpt2-babylm100m	.626 [.572, .674]	+.068*
ltgbert-100m-2024	.623 [.568, .677]	+.064*
babylm100m	.618 [.563, .667]	+.058
gpt2-babylm10m	.604 [.545, .657]	+.043
roberta-babylm100m	.583 [.529, .638]	+.024
gptbert25-causal	.583 [.523, .640]	+.023
opt125-babylm100m	.577 [.518, .630]	+.017
opt125-babylm150m	.571 [.509, .627]	+.012
babylm10m	.571 [.510, .628]	+.010
opt125-full (ref.)	.560 [.500, .615]	—
roberta-full	.525 [.455, .592]	−.035
elcbert-100m-2023	.523 [.460, .581]	−.039
gptbert-100m-2024	.502 [.438, .563]	−.059
babylm58-2023	.489 [.423, .551]	−.070*
gpt2-full	.483 [.415, .545]	−.077*
opt125-babylm10m	.414 [.346, .481]	−.145*

Table 2: Nested read-out scores and pairwise differences against the full-data OPT reference on common valid words. Asterisks mark differences whose 95% bootstrap interval excludes zero.

and beyond what the classic word-level predictors capture, but only for child-scale checkpoints.

Figure 1 and Table 2 report the nested read-out score for every checkpoint. No read-out reaches the $\rho = .707$ of the Braginsky et al. (2016) baseline, which is expected: the baseline uses measured properties of exactly the kind corpora and adult raters provide, while the read-out sees nothing but the representation. The informative comparisons are between checkpoints.

Child-scale data matches or exceeds full-scale data. In every family the child-scale checkpoint at least matches its full-scale counterpart. The contrast is sharpest for GPT-2: the 2025 baseline trained on 100M words reaches $\rho = .626$ [.572, .674], while the WebText checkpoint reaches .483 [.415, .545]. Against the shared reference model, the former is reliably higher ($\Delta\rho = +.068$, 95% CI [.006, .124]) and the latter reliably lower (−.077 [−.141, −.016]). In the OPT family the 150M-token checkpoint reaches .571 against .560 for the 180B-token original, and in the RoBERTa family the BabyLM baseline reaches .583 against .525. Matched or better, in other words, despite three to four orders of magnitude less text. Data volume within the child-scale regime shows no consistent advantage either: the GPT-2 10M baseline (.604) is comparable to the 100M variant, while the OPT 10M baseline falls well below the reference (−.145 [−.219, −.073]), so the effect of

very small data interacts with the training recipe.

Objective mix matters as much as data scale.

The three 2025 GPT-BERT baselines share data and architecture and differ only in the ratio of causal to masked training. The mixed ($\rho = .639$ [.584, .685]) and masked-focused (.637 [.585, .685]) variants top the entire ranking and are reliably above the reference (+.079 [.021, .137] and +.078 [.010, .143]); the causal-focused variant reaches .583, with a confidence interval spanning zero (+.023 [−.040, .087]). The purely masked LTG-BERT baseline is also reliably above the reference (+.064 [.001, .128]). Layer profiles separate the two model classes: encoder and hybrid checkpoints reach their peak in the upper third of the network, while decoder profiles are flatter across depth (Figure 2).

Benchmark winners do not best capture children’s acquisition order.

The winning systems of the first two challenges sit in the lower half of the ranking: ELC-BERT reaches $\rho = .523$ [.460, .581] and the 2024 GPT-BERT submission reaches .502 [.438, .563], below the naive GPT-2 baselines trained on the same corpus. Both systems were selected largely by aggregate accuracy on benchmark suites built around BLiMP (Warstadt et al., 2020). Whatever the challenge leaderboard selects for, it is not the ability to capture children’s real acquisition order. The nested protocol also acts as a check on optimistic layer selection here: the distilled babylm100m baseline leads the post hoc ranking ($\rho = .671$ at the first layer) but drops to .618 under nested selection, with unstable layer choices across folds, so we treat effects of distillation as unresolved rather than as a finding.

Semantic structure separates the two regimes most sharply.

The three neighborhood measures alone predict acquisition order at $\rho = .33$ to .52 from the representation spaces of child-scale checkpoints, but only .19 to .27 from the three full-scale spaces. Added to the four Braginsky et al. predictors, the measures raise the cross-validated correlation from .707 to as high as .756 (roberta-babylm100m) among child-scale checkpoints, while the full-scale OPT space adds nothing (.707 to .708). These analyses use the best layer of each space and are exploratory; the point is their consistency across fifteen child-scale checkpoints, not any single comparison.

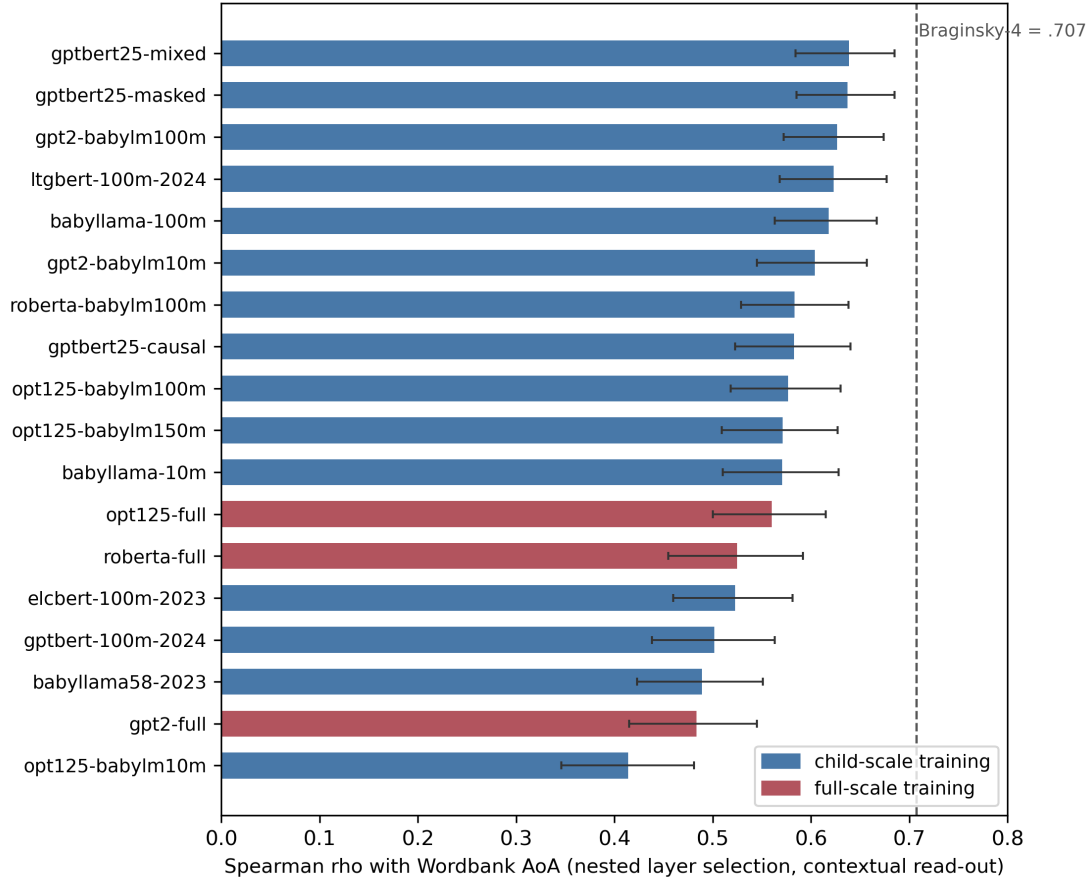


Figure 1: Nested read-out score per checkpoint (contextual representations), with 95% bootstrap confidence intervals. Blue bars are child-scale checkpoints; red bars are full-scale checkpoints. The dashed line marks the four-predictor baseline of Braginsky et al. (2016).

5 Discussion

We asked whether the order in which children learn words is recoverable from the internal representations of trained language models, and what kind of training makes it recoverable. Across 18 checkpoints, a linear read-out of frozen representations predicts children’s age of acquisition well above chance, and four results stand out. Child-scale training matched or exceeded full-scale training in all three architecture families, with the sharpest gap in GPT-2 ($\rho = .626$ against $.483$). Training objective mattered as much as data scale: masked and hybrid objectives produced representations from which acquisition order was easiest to recover, while purely causal training lagged behind. The winning systems of previous challenges, by contrast, did not best capture children’s acquisition order. And geometric measures of semantic structure carried acquisition-relevant information beyond the classic predictors only in child-scale representations. We take up each result in

turn below.

Why is more of children’s acquisition order recoverable from child-scale training? The most direct explanation is register. The BabyLM corpus consists mostly of transcribed or scripted speech, and about 40% of it comes from child-directed or child-appropriate sources (Warstadt et al., 2023), so the distributional statistics a child-scale model compresses are closer to the statistics children learn from. A more skeptical variant of this explanation attributes the effect to our extraction contexts: word representations are aggregated over CHILDES sentences, which are in-domain for child-scale models and out-of-domain for full-scale ones. Two observations speak against a purely skeptical reading. The full-scale models are far from floor (the weakest, GPT-2, still reaches $\rho = .483$), and the semantic-structure result concerns relative geometry within each space rather than fit to the extraction sentences. The decisive test is to repeat the extraction with register-neutral

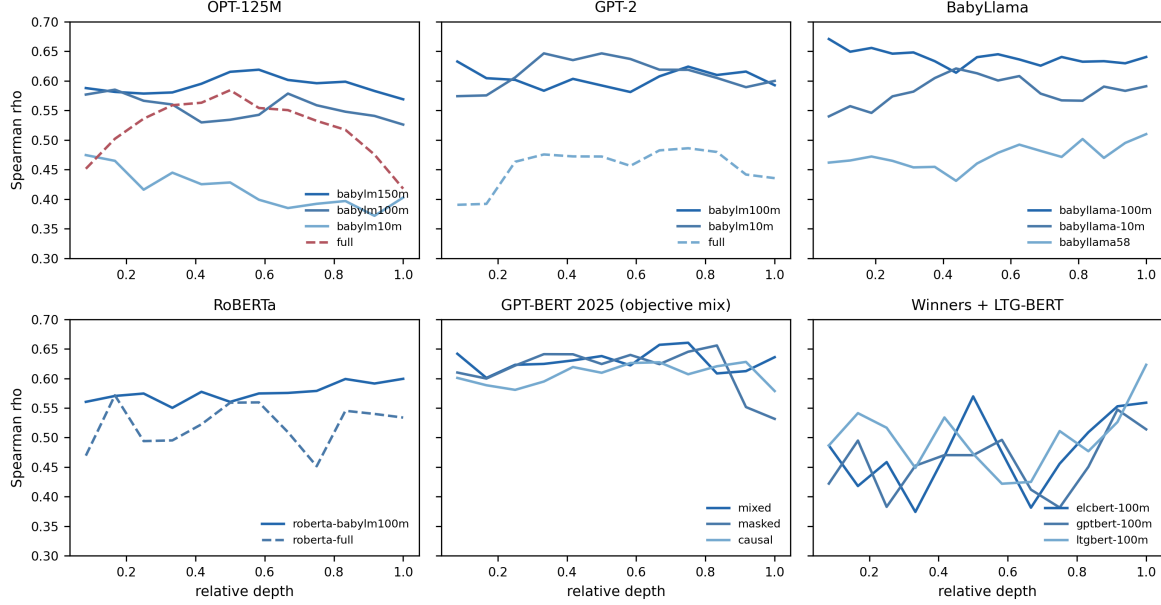


Figure 2: Read-out score by relative depth, per architecture family. Dashed lines are full-scale checkpoints; solid lines are child-scale checkpoints. The first three panels show purely causal models (OPT-125M, GPT-2, Llama); the remaining panels show masked (RoBERTa, LTG-BERT, ELC-BERT) and hybrid (GPT-BERT) objectives.

contexts, and we list it as the first item for future work. The register account also has a precedent with a twist: [Feng et al. \(2024\)](#) find that models trained on child-directed speech alone perform worst among size-matched datasets on syntactic and lexical-semantic benchmarks. Register match does not buy benchmark performance; our results identify what it does buy, namely developmental alignment of the semantic space.

Data quantity provides a second candidate mechanism. In the reading-time literature, surprisal from larger pre-trained transformers fits human reading times strictly worse within a model family ([Oh and Schuler, 2023](#)), and GPT-2 yields better psychometric fit than models a thousand times its size ([Shain et al., 2024](#)). [Oh and Schuler \(2023\)](#) link this trend to the capacity of larger models to memorize training sequences,² which pushes their predictions of open-class words beyond humanlike levels. Child-scale and full-scale checkpoints differ in register and in quantity at once, so our design does not separate the two accounts. Within the child-scale range the data lean toward register, since moving from 10M to 100M words does not consistently change alignment, but the memorization account concerns scales far be-

²This pattern of models relying on memorization where humans may not extends beyond the reading-time literature as well, such as in ordering preferences of binomials ([Houghton et al., 2025](#); [Houghton, 2025](#)).

yond this range. A scale ladder on a fixed corpus would separate the two accounts, and we list it alongside the register-neutral extraction as future work.

The objective effect has a candidate mechanism. A word’s masked-prediction loss depends on how well the model reconstructs it from surrounding words, so masked training organizes the space by substitutability in context. Substitutability is close to the neighborhood structure that predicts children’s acquisition ([Hills et al., 2009](#); [Fourtassi et al., 2020](#)), which may explain why masked and hybrid objectives yield spaces from which acquisition order is easier to read. Causal training optimizes next-word generation, and what it stores about a word is spread over the positions where that word helps prediction rather than concentrated in the word’s own representation.

For the BabyLM community the practical point is that AoA decodability is a cheap evaluation. It needs one frozen checkpoint, public child data, and a ridge regression: no intermediate checkpoints, no fine-tuning, no generation. It complements the official word-learning task ([Charpentier et al., 2025](#)), which reads acquisition from surprisal trajectories and therefore requires the full training run. The dissociation between leaderboard rank and how well a model captures children’s acquisition order gives the human-likeness

track a concrete target, since architectures that win on aggregate benchmarks do not automatically produce developmentally aligned representations. The method also transfers directly to the 2026 multilingual track: CDI instruments have been adapted well beyond English and Wordbank hosts norms for Dutch and Mandarin (Frank et al., 2017), the two new languages of that track (Choshen et al., 2026), and every step of the pipeline is language-agnostic.

6 Conclusion

This study measured how much of children’s word-learning order can be read out of the frozen representations of 18 public checkpoints. Child-scale training matched or exceeded full-scale training in all three architecture families, masked and hybrid objectives produced representations from which acquisition order was easiest to recover, and the winners of previous challenges did not best capture children’s acquisition order. Children’s acquisition order is already present in the representations of well-trained BabyLM systems; whether models can also be made to acquire words in a child-like order over training is the natural next question, and the intermediate checkpoints the community already releases make both questions answerable with the same data.

Limitations

First, all results concern English; the multilingual extension is planned but not yet run. Second, the acquisition target is parent-reported production AoA for 619 CDI items, which covers early vocabulary only and inherits the biases of parental report. Third, contextual representations are extracted from CHILDES sentences, which may favor models trained on similar registers; a replication with register-neutral extraction contexts is the clearest next check. Fourth, each checkpoint represents a single training run, so seed variance is unquantified, and the checkpoints differ in tokenizer and training details beyond the axes we analyze. Finally, decodability is correlational: a linear probe recovering acquisition order from a model’s representation does not show that children exploit the same structure.

References

Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Computational Lin-*

guistics, 48(1):207–219.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. [Interpreting pretrained contextualized representations via reductions to static embeddings](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781.

Mika Braginsky, Daniel Yurovsky, Virginia A. Marchman, and Michael C. Frank. 2016. From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

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.

Bastian Bunzeck and Sina Zarri . 2025. [Subword models struggle with word learning, but surprisal hides it](#). *arXiv preprint arXiv:2502.12835*.

Luca Capone, Alessandro Bondielli, and Alessandro Lenci. 2024. [ConcreteGPT: A baby GPT-2 based on lexical concreteness and curriculum learning](#). In *Proceedings of the BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 189–196.

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.

Lucas Georges Gabriel Charpentier and David Samuel. 2023. [Not all layers are equally as important: Every Layer Counts BERT](#). In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.

Lucas Georges Gabriel Charpentier and David Samuel. 2024. [GPT or BERT: why not both?](#) In *Proceedings of the 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 Gotlieb Wilcox. 2026. BabyLM turns 4 and goes multilingual: Call for papers for the 2026 BabyLM workshop. *arXiv preprint arXiv:2602.20092*.

- 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*.
- Abdellah Fourtassi, Yuan Bian, and Michael C. Frank. 2020. [The growth of children’s semantic and phonological networks: Insight from 10 languages](#). *Cognitive Science*, 44(7):e12847.
- 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, 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.
- Hiromichi Hagihara and Kazuki Miyazawa. 2026. [How well do large language models mirror human cognition of word concepts?: A comparison of psychological ratings for early-acquired English words](#). *Behavior Research Methods*, 58:58.
- John Hewitt and Percy Liang. 2019. [Designing and interpreting probes with control tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2733–2743.
- Thomas T. Hills, Mounir Maouene, Josita Maouene, Adam Sheya, and Linda B. Smith. 2009. [Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition?](#) *Psychological Science*, 20(6):729–739.
- Zachary N. Houghton, Yuhuan Zhou, Devin Pluth, and Vijay K. Gurbani. 2026. [The holistic storage of verb+up phrases in text-based and audio-based language models](#). *arXiv preprint arXiv:2606.13993*.
- Zachary Nicholas Houghton. 2025. *Multi-Word Representations in Minds and Models: Investigating the Storage of Multi-Word Phrases in Humans and Large Language Models*. University of California, Davis.
- Zachary Nicholas Houghton, Kenji Sagae, and Emily Morgan. 2025. [The role of abstract representations and observed preferences in the ordering of binomials in large language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 695–702.
- 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 *Proceedings of the BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [RoBERTa: A robustly optimized BERT pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd edition. Lawrence Erlbaum Associates.
- Byung-Doh Oh and William Schuler. 2023. [Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times?](#) *Transactions of the Association for Computational Linguistics*, 11:336–350.
- Lynn K. Perry, Marcus Perlman, and Gary Lupyan. 2015. [Iconicity in English and Spanish and its relation to lexical category and age of acquisition](#). *PLOS ONE*, 10(9):e0137147.
- Eva Portelance, Judith Degen, and Michael C. Frank. 2020. [Predicting age of acquisition in early word learning using recurrent neural networks](#). In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, pages 1057–1063.
- Eva Portelance, Yuguang Duan, Michael C. Frank, and Gary Lupyan. 2023. [Predicting age of acquisition for children’s early vocabulary in five languages using language model surprisal](#). *Cognitive Science*, 47(9):e13334.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#). Technical report, OpenAI.
- David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. [Trained on 100 million words and still in shape: BERT meets British National Corpus](#). In *Findings of the Association for Computational Linguistics: EACL 2023*.
- Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. [Large-scale evidence for logarithmic effects of word predictability on reading time](#). *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.
- Filip Smolíková and Maroš Filip. 2022. [Corpus-based age of word acquisition: Does it support the validity of adult age-of-acquisition ratings?](#) *PLOS ONE*, 17(5):e0268504.
- Inar Timiryasov and Jean-Loup Tastet. 2023. [Baby llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty](#). In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.
- Sean Trott. 2024. [Can large language models help augment English psycholinguistic datasets?](#) *Behavior Research Methods*, 56:6082–6100.

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*.

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.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A Checkpoint identifiers

HuggingFace identifiers, in the order of Table 1: znhoughton/opt-babylm-125m-20eps-seed964; babylm/opt-125m-strict-2023; babylm/opt-125m-strict-small-2023; facebook/opt-125m; BabyLM-community/babylm-baseline-100m-gpt2; BabyLM-community/babylm-baseline-10m-gpt2; openai-community/gpt2; timinar/baby-llama-58m; babylm/babyllama-100m-2024; babylm/babyllama-10m-2024; babylm/roberta-base-strict-2023; FacebookAI/roberta-base; lgcharpe/ELC_BERT_baby_100M; Itg/gpt-bert-babylm-base; babylm/Itgbert-100m-2024; BabyLM-community/babylm-baseline-100m-gpt-bert-mixed; BabyLM-community/babylm-baseline-100m-gpt-bert-causal-focus; BabyLM-community/babylm-baseline-100m-gpt-bert-masked-focus.