

TinyLibrary: Do Age-Graded Curricula Help Small Vision-Language Models?

Dheeraj Varghese
University of Amsterdam
d.varghese@uva.nl

Abstract

Developmentally plausible curricula are a recurring hypothesis in research on cognitively plausible language-model training, but it remains unclear which notions of difficulty produce useful training curricula. We ask whether annotations derived from children’s literature yield a usable ordering signal. We introduce TinyLibrary, a corpus of 46K illustrated children’s-book pages drawn from a collection for readers aged 3–13. We infer each book’s age band from its OCR text and use the resulting label to condition the generation of multimodal training data. This design lets us test whether the developmental structure of children’s literature can be used to generate synthetic multimodal training data as an effective curriculum signal. We train small vision–language models for up to 10 epochs, using a pretrained vision encoder and a language model trained from scratch, and compare five curriculum orderings at matched token budgets. We evaluate the resulting models on the zero-shot BabyLM 2026 suite. Our results show that the age progression only partly aligns with teacher-model predictability: readability increases monotonically as intended, whereas teacher perplexity is U-shaped, with text for the youngest readers among the least predictable per token. Consistent with this mismatch, age-based ordering shows no reliable advantage over random ordering. Teacher-perplexity ordering achieves the highest mean BLiMP score and exceeds its seed-matched random counterpart in 10 of 12 comparisons, but this advantage transfers unevenly across teacher families and is entangled with annotation format. No ordering yields a reliable grounding advantage: Winoground remains at chance, while VQA is above chance but variable across conditions and seeds. We release the annotations and code.

1 Introduction

Curriculum learning asks whether models learn more effectively when training examples are presented in a structured progression rather than in

random order (Elman, 1993; Bengio et al., 2009). This question is especially relevant to the BabyLM challenge (Warstadt et al., 2023; Hu et al., 2024; BabyLM Organizing Committee, 2026), where models must learn from developmentally plausible data budgets. A useful curriculum could improve learning without requiring more data or a larger model. Its success, however, depends on what counts as “easy”: input that is appropriate for a young reader is not necessarily predictable or useful to a language model. The central question is therefore not only whether curriculum learning helps, but whether a human developmental ordering provides the right difficulty signal for the learner.

Previous BabyLM studies have tested curricula based on properties such as sentence length, readability, or teacher-model perplexity, where a separate pretrained language model scores how predictable each sample is (Martinez et al., 2023; Saha et al., 2024). Reported gains have been weak or setting-dependent. These measures are typically applied *post hoc* to corpora that were not created around an intended progression. Children’s literature provides a complementary test case: books are written and illustrated for readers at different developmental levels, offering a source-derived, human-motivated ordering signal. We ask whether this age-targeted structure can be transferred into synthetic multimodal training data and used as an effective curriculum for small vision–language models. We address this question through three contributions:

1. **We introduce TinyLibrary, a multimodal corpus constructed to study age-based curriculum ordering.** Constructed from the English-language portion of the International Children’s Digital Library (ICDL; Druin, 2005), TinyLibrary contains 46,721 illustrated pages and 128,823 synthetic captioning, visual question-answering, and reasoning sam-

ples, totaling 21.7M words. Because reliable per-book age metadata was unavailable, we infer one of four coarse age bands from each book’s original OCR text. To create image-grounded VLM examples while preserving this age signal, we use the inferred band to condition the generation of captioning, visual question-answering, and reasoning samples.

2. **We find that inferred age order does not provide a stable curriculum advantage.** Using nanoVLM (Wiedmann et al., 2025), we compare developmental age order and its reverse with readability, teacher-perplexity, and random ordering under matched token budgets, data splits, and seeds. As intended, age conditioning produces progressively higher Flesch–Kincaid grades. However, mean per-token negative log-likelihood under a teacher language model is U-shaped across age bands: it is highest for the youngest and oldest bands and lowest for the middle bands. Neither developmental nor reverse-age ordering yields a stable BLiMP advantage across schedules and training durations; at 10 epochs, random ordering is higher in all 12 seed-matched comparisons.
3. **Teacher-perplexity ordering achieves the highest mean BLiMP, but does not establish a portable curriculum advantage.** Its advantage transfers unevenly across teacher-model families. Moreover, the resulting difficulty quartiles largely separate annotation types: the easiest samples are predominantly question answering and reasoning, while the hardest are almost entirely captions. The observed advantage therefore cannot be attributed confidently to a general notion of model-relative difficulty.

Together, these results distinguish human age appropriateness from model predictability: the former does not yield stable curriculum gains in our setting, while the apparent benefit of the latter remains entangled with teacher choice and task format.

2 Related Work

Learning under constrained data. The BabyLM challenge provides a standardized setting for studying how effectively language models learn from developmentally motivated data budgets (Warstadt et al., 2023; Hu et al., 2024; Charpentier

et al., 2025a; BabyLM Organizing Committee, 2026). Prior work shows that small models can acquire substantial grammatical knowledge from child-directed corpora (Huebner et al., 2021), while simplified synthetic data can also support strong learning under limited exposure (Eldan and Li, 2023). Since the introduction of the multimodal track, visual input has not consistently translated into stronger text-only grammatical performance (Hu et al., 2024). Findings from the third BabyLM challenge likewise emphasize architectural and objective choices among successful systems (Charpentier et al., 2025b). Our work instead holds the model and training budget fixed to study whether the order of multimodal training examples improves learning.

Curricula for language models. The “starting small” hypothesis proposes that learning improves when models encounter simpler or more restricted input before more complex data (Elman, 1993). Rohde and Plaut (1999) questioned the generality of this effect, while later work formalized curriculum learning as ordering examples by difficulty (Bengio et al., 2009; Platanios et al., 2019). In BabyLM, CLIMB (Martinez et al., 2023) found no reliable gains from vocabulary or data curricula, and Saha et al. (2024) reported only setting-dependent gains for heuristic multimodal curricula. Most closely related to our question, Feng et al. (2024) find that the local properties of child-directed speech affect learning, whereas its global developmental ordering does not. We extend this line of investigation to multimodal data derived from age-targeted children’s books. We compare inferred age order with readability and teacher perplexity, and by treating the ordering signal and its training schedule as separate design choices.

Developmentally inspired multimodal training. Developmental research links word learning to the structure of children’s visual experience: Yu and Smith (2012) find that toddlers’ learning of object names is associated with the named object dominating their view during naming. In computational work, Sheybani et al. (2023) find that presenting infant egocentric videos in developmental order improves self-supervised visual learning, with benefits linked to the slower-changing and spatially simpler views of younger infants. For multimodal learning, Vong et al. (2024) demonstrate word-referent learning from paired visual and linguistic input recorded from one child’s experience,

while BabyVLM (Wang et al., 2025) constructs synthetic child-directed training data for vision-language models. Our work examines a complementary source of developmental structure: illustrated books written for readers of different ages. We test whether inferred reader age provides an effective ordering signal for synthetic multimodal annotations derived from these books.

3 The TinyLibrary Corpus

Source and page filtering. TinyLibrary¹ is derived from the International Children’s Digital Library (ICDL; Druin, 2005), a curated collection of children’s books for readers aged 3–13. Of 2,801 OCR-processed books, 2,795 receive an age-band label using the procedure described below. We retain the 2,416 English-language books, comprising 251,410 scanned pages.

Because our goal is to construct image-grounded training examples, we select pages with substantial non-text visual content. We remove pages dominated by OCR-detected text (PaddleOCR; Du et al., 2020), near-blank pages, and pages whose visual complexity largely disappears after detected text is removed. This process yields 46,721 illustrated pages. Appendix A provides the full filtering criteria, and Appendix B shows example pages.

Inferring age bands. We use Mistral-7B-Instruct (Jiang et al., 2023) to clean each book’s original OCR text and infer one of four target-age bands: 3–5, 6–8, 9–12, or 12+. Long books are divided into text chunks, classified separately, and assigned a band by majority vote.

Generating multimodal annotations. We generate annotations in two stages. First, CogVLM2 (Hong et al., 2024) produces a detailed description of each selected page image. Llama-3-8B-Instruct (Grattafiori et al., 2024) then converts the description into LLaVA-style (Liu et al., 2023) captioning, visual question-answering, and reasoning samples. The generation prompt is conditioned on the book’s inferred age band and specifies corresponding vocabulary, sentence structure, and reasoning complexity. The resulting corpus contains 128,823 samples and 21.7M words.

Corpus composition and ordering signals. Table 1 summarizes the corpus by inferred age band. The distribution is imbalanced: the 12+ band con-

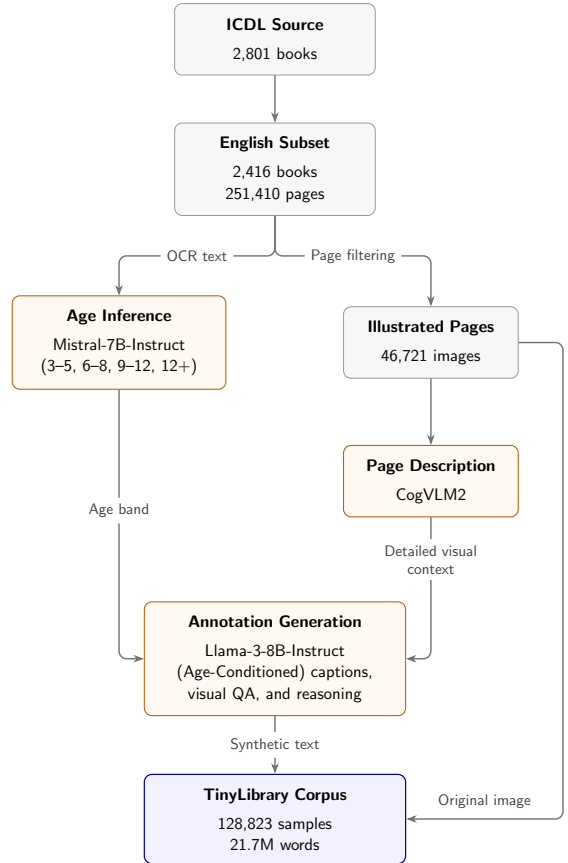


Figure 1: TinyLibrary construction. OCR text is used to infer a book-level age band, while illustrated pages are selected for annotation. Page descriptions are generated from images and converted into age-conditioned captioning, visual QA, and reasoning samples.

tains 60% of all words. We characterize the generated target text using Flesch–Kincaid grade (Kincaid et al., 1975) and mean per-token negative log-likelihood (NLL) under a SmolLM2-135M teacher language model (Allal et al., 2025), conditioned on the sample prompt.

As intended, the age conditioning produces progressively higher Flesch–Kincaid grades, from 3.51 for the 3–5 band to 10.76 for the 12+ band, indicating that the generated annotations follow the requested reading levels.

Teacher-model NLL follows a different pattern. It is highest for the 3–5 band, lowest for the two middle bands, and rises again for the 12+ band. The same band-level pattern is reproduced by Qwen3-0.6B and Gemma-3-270M, whose sample-level rankings correlate strongly with the SmolLM2 ranking (Spearman $\rho = 0.91$ and 0.96; Appendix F). Thus, inferred age order and teacher-NLL order encode different sample

¹Dataset on [Hugging Face](#); code on [GitHub](#).

Band	Samples	Words	Tok.	FK	NLL
3–5	34,623	3.36M	128	3.51	2.856
6–8	18,389	2.51M	170	5.34	2.542
9–12	15,488	2.70M	212	7.76	2.523
12+	60,323	13.1M	264	10.76	2.748
Total	128,823	21.7M	—	—	—

Table 1: TinyLibrary by inferred age band: sample and word counts, mean tokens per sample, mean Flesch–Kincaid grade, and mean per-token NLL under a SmolLM2-135M teacher. The 12+ band contains 60% of the corpus words.

progressions. This divergence motivates treating them as separate curriculum orderings.

4 Experimental Setup

Model and training. We use nanoVLM (Wiedmann et al., 2025), comprising a SigLIP2-base-512 vision encoder (Tschannen et al., 2025), a two-layer MLP projector, and a language model with the SmolLM2-360M architecture (Allal et al., 2025), for 460M parameters in total. The vision encoder is initialized from pretrained weights, while the projector and language model are initialized from scratch; all components are then trained jointly. Images are processed at 512×512 resolution, producing 64 visual tokens, and samples are packed into 4,096-token sequences. Appendix C provides the remaining training details.

Curriculum orderings. We distinguish the sample *ordering* from the schedule used to present that ordering during training. We compare five orderings:

1. **Developmental:** inferred age bands in ascending order, beginning with 3–5;
2. **Anti-developmental:** the reverse age-band order, beginning with 12+;
3. **Readability:** ascending Flesch–Kincaid grade;
4. **Teacher perplexity:** ascending per-token NLL under SmolLM2-135M; and
5. **Random:** a fresh random permutation of the corpus in each epoch.

The anti-developmental condition tests whether the direction of the inferred age progression matters. Readability and teacher perplexity provide continuous alternatives based on surface complexity and teacher-model predictability, respectively.

Schedule implementations. We implement each of the four non-random orderings using two schedules. The staged schedule partitions the corpus into four consecutive groups: the four age bands for the developmental conditions and score quartiles for readability and teacher perplexity. Samples are shuffled within each group, but group membership and group order remain fixed across epochs. The age-band groups retain their naturally unequal sizes, whereas the continuous-score quartiles contain equal numbers of samples. Each staged epoch presents every training sample once.

The competence-based schedule (Platanios et al., 2019) samples with replacement from an expanding prefix of the ranked corpus. At training step t out of T , the available fraction is

$$c(t) = \min\left(1, \sqrt{c_0^2 + \frac{t}{T}(1 - c_0^2)}\right), c_0 = 0.1.$$

Training therefore begins by sampling from the first 10% of the selected ordering, and the available pool expands gradually until the full corpus becomes available at the end. Because sampling is performed with replacement, early-ranked examples can be repeated and some later-ranked examples may never be observed.

The random condition does not use either schedule: it presents every sample once per epoch and reshuffles the full corpus between epochs. Crossing the four non-random orderings with the two schedules, together with the random baseline, produces nine core conditions.

Matched comparisons. All runs use the same model architecture, the same static 2% sample-level validation split, training hyperparameters, and consume the same total number of training tokens. We reuse the same three seed values across orderings, allowing each curriculum run to be paired with the corresponding random-ordering run.

In staged and random runs, each epoch is one complete corpus pass, so their comparison isolates presentation order. For competence-based runs, the 5- and 10-epoch labels denote equivalent token budgets rather than complete corpus passes. Because competence-based sampling changes both presentation order and unique-sample coverage, we base our primary ordering comparisons on the staged conditions and the random baseline and treat the competence-based results as secondary.

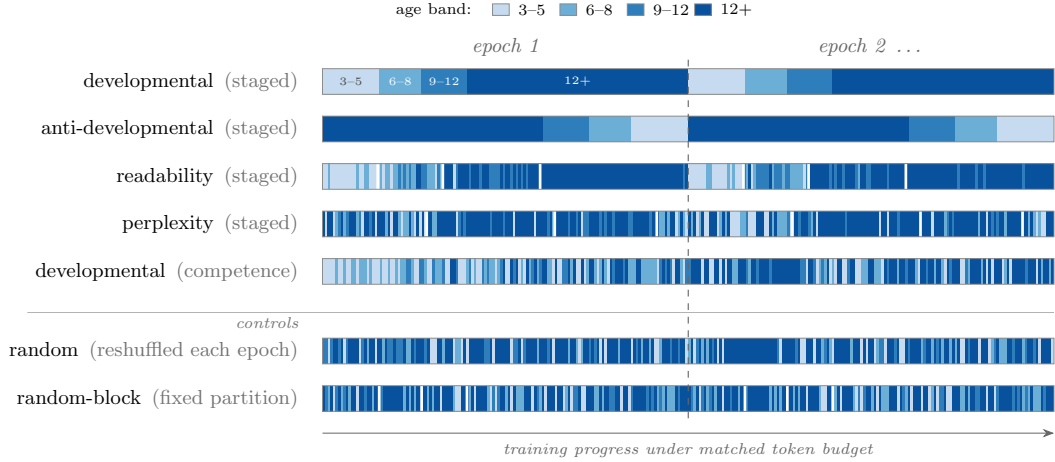


Figure 2: Age-band composition induced by the curriculum orderings and schedules. Color denotes inferred age band, and every row consumes the same training-token budget. Staged conditions repeat fixed groups across epochs, competence-based conditions sample from an expanding eligible pool, and random ordering reshuffles the complete corpus each epoch. The random-block control repeats fixed groups containing randomly assigned samples.

Runs and controls. We train the nine core conditions for 5 and 10 equivalent epochs using three seeds, yielding 54 runs. We treat the 10-epoch budget as primary and the 5-epoch budget as a second training duration at which to assess whether the direction of an effect persists.

We run three additional controls at the 10-epoch budget. First, the random-block control randomly assigns samples to four fixed groups matching the token masses of the age-band stages. Samples are shuffled within each group, but group membership and order are reused across epochs. This preserves the repeated staged structure without grouping samples by age. Second, we construct teacher-perplexity orderings using Qwen3-0.6B and Gemma-3-270M to test whether the SmoLLM2-derived ordering transfers across teacher families. Third, we train text-only models under developmental, teacher-perplexity, and random ordering, removing image tokens while retaining the corresponding text. These controls bring the total to 72 trained runs, in addition to three untrained model initializations.

Evaluation. For linguistic evaluation, we evaluate the trained language-model component without image inputs using the zero-shot subset of the official BabyLM 2026 pipeline (BabyLM Organizing Committee, 2026): BLiMP and BLiMP Supplement (Warstadt et al., 2020), EWoK (Ivanova et al., 2024), COMPS (Misra et al., 2023), and entity tracking (Kim and Schuster, 2023). For multimodal

evaluation, we use the complete vision-language model on the BabyLM VQA and Winoground tasks.

Final checkpoints are evaluated on the full task sets. To examine linguistic performance over training, we also evaluate checkpoints saved every 250 steps using the pipeline’s fast subsets.

5 Results

Table 2 reports final zero-shot performance at the primary 10-epoch budget. The first nine rows form the core experimental grid; the remaining rows report the random-block and alternative-teacher controls together with the untrained reference. Results for the 5-epoch budget and individual seeds are provided in Appendix D and Appendix E, respectively.

Metric sensitivity. With this 21.7M-word corpus, EWoK and COMPS remain near their 50% chance levels in every condition. BLiMP Supplement varies across conditions but does not produce a consistent curriculum ranking. Entity tracking is also difficult to interpret: the untrained model scores 41.8 ± 0.1 , while training generally lowers performance and produces substantial seed variance. We therefore use BLiMP as the primary outcome for comparing the curriculum orderings; the other linguistic metrics do not yield a consistent or interpretable ranking of those orderings.

Order	Schedule	BLiMP	Supplement	EWoK	COMPS	Entity tracking
Random	—	53.4 $\pm$ 0.4	44.5 $\pm$ 1.7	49.1 $\pm$ 0.6	49.8 $\pm$ 0.2	30.8 $\pm$ 5.5
Developmental	Staged	51.8 $\pm$ 0.6	46.5 $\pm$ 2.6	49.4 $\pm$ 0.4	50.2 $\pm$ 0.4	24.5 $\pm$ 15.2
Developmental	Competence	52.2 $\pm$ 1.0	46.0 $\pm$ 1.0	50.3 $\pm$ 0.5	50.0 $\pm$ 0.1	34.8 $\pm$ 5.7
Anti-developmental	Staged	51.5 $\pm$ 1.2	48.1 $\pm$ 0.7	50.5 $\pm$ 0.6	49.8 $\pm$ 0.1	24.7 $\pm$ 4.0
Anti-developmental	Competence	51.7 $\pm$ 0.6	47.1 $\pm$ 3.2	50.1 $\pm$ 0.3	49.9 $\pm$ 0.2	27.1 $\pm$ 11.1
Readability	Staged	52.8 $\pm$ 1.7	46.2 $\pm$ 1.3	50.1 $\pm$ 0.4	50.0 $\pm$ 0.4	27.2 $\pm$ 8.5
Readability	Competence	52.5 $\pm$ 0.6	44.6 $\pm$ 0.9	50.2 $\pm$ 0.3	50.3 $\pm$ 0.5	35.1 $\pm$ 3.4
Teacher perplexity	Staged	55.2 $\pm$2.1	48.0 $\pm$ 1.0	49.6 $\pm$ 0.3	50.1 $\pm$ 0.1	17.1 $\pm$ 2.9
Teacher perplexity	Competence	53.5 $\pm$ 1.0	46.1 $\pm$ 1.4	50.5 $\pm$ 0.4	50.1 $\pm$ 0.6	30.1 $\pm$ 2.8
Random-block (control)	Staged	52.8 $\pm$ 1.1	45.8 $\pm$ 1.6	50.1 $\pm$ 0.3	49.8 $\pm$ 0.0	36.7 $\pm$ 7.7
Teacher perplexity, Gemma (control)	Staged	54.0 $\pm$ 0.2	46.0 $\pm$ 3.0	50.4 $\pm$ 0.5	50.1 $\pm$ 0.5	17.3 $\pm$ 4.4
Teacher perplexity, Qwen3 (control)	Staged	52.6 $\pm$ 1.1	45.8 $\pm$ 2.5	50.1 $\pm$ 0.6	50.0 $\pm$ 0.1	25.8 $\pm$ 12.7
Untrained reference		50.0 $\pm$ 0.4	47.6 $\pm$ 1.0	50.0 $\pm$ 0.5	50.4 $\pm$ 0.3	41.8 $\pm$ 0.1

Table 2: **Teacher-perplexity/staged has the highest mean BLiMP score, while EWoK and COMPS remain near chance.** Zero-shot accuracy (%) after 10 epochs, reported as mean $\pm$ standard deviation over three seeds. The first nine rows form the core grid; random-block and the alternative-teacher orderings are 10-epoch controls. Bold indicates the highest mean BLiMP score.

5.1 Age-based ordering does not yield a stable advantage

In the primary staged comparison at 10 epochs, both age-based orderings finish below random ordering. Developmental/staged scores 51.8 $\pm$ 0.6 BLiMP, a mean seed-paired difference of -1.56 points from random, while anti-developmental/staged scores 51.5 $\pm$ 1.2, a paired difference of -1.84 . Every seed-level difference for both conditions is negative.

The competence-based age conditions also fall below random at 10 epochs: developmental/competence by -1.21 points and anti-developmental/competence by -1.69 points on average. Because these comparisons reflect differences in sample exposure as well as presentation order, we treat them as secondary. Across the two age directions, two schedules, and three seeds, all 12 paired differences favor random ordering at this budget.

At 5 epochs, the age-based results are mixed: anti-developmental/staged is above random, anti-developmental/competence is below it, and the developmental conditions remain close to random (Appendix D). None of the positive five-epoch differences persists at 10 epochs. Thus, neither age direction provides an advantage that is consistent across schedules and training budgets.

Readability ordering likewise fails to produce a stable improvement. Although Flesch–Kincaid grade increases monotonically with inferred age band, neither readability condition exceeds the random mean at either training budget. Together, these

results provide no reliable evidence that either inferred reader age or its induced readability progression supplies a beneficial curriculum ordering in this setting.

5.2 Teacher-perplexity ordering has the highest mean but does not transfer reliably

Teacher-perplexity ordering produces the highest mean BLiMP score at both training budgets, although the strongest schedule differs. At 10 epochs, teacher-perplexity/staged reaches 55.2 $\pm$ 2.1, a mean seed-paired advantage of $+1.87$ points over random; the three paired differences range from -0.50 to $+4.15$. Teacher-perplexity/competence is effectively tied with random at this budget, with a mean paired difference of $+0.12$.

At 5 epochs, both teacher-perplexity conditions are above random: staged by $+1.62$ points and competence by $+2.42$. Across the two schedules, two budgets, and three seeds, teacher-perplexity ordering exceeds its seed-matched random counterpart in 10 of 12 comparisons. This repeated direction makes teacher perplexity the strongest ordering signal in the core grid, but the seed variation and schedule dependence prevent a stronger claim.

The alternative-teacher controls further qualify this result. A staged ordering derived from Gemma-3-270M reaches 54.0 $\pm$ 0.2, retaining a smaller mean paired advantage of $+0.66$ over random. The Qwen3-0.6B ordering reaches 52.6 $\pm$ 1.1 and retains no advantage, with a mean paired difference of -0.74 . The gain observed with the SmoILM2

teacher therefore does not transfer consistently across teacher-model families.

Teacher NLL is also strongly entangled with annotation format. The two easiest NLL quartiles contain almost no captions and consist almost entirely of visual question-answering and reasoning samples, whereas the hardest quartile is 99.6% captions (Appendix H). Consequently, under teacher-perplexity ordering, question-answering and reasoning samples are presented predominantly earlier in training, whereas captions are presented predominantly later. The observed BLiMP advantage could therefore reflect teacher-learner alignment, task-format order, or an interaction between them. These experiments do not establish teacher perplexity as a general, portable measure of curriculum difficulty.

5.3 Repeated fixed partitions are associated with late degradation

Figure 3 reports checkpoint performance on the fast BLiMP subset. These trajectories characterize changes during training; final performance in Table 2 is measured on the full task.

All conditions improve rapidly at the beginning of training. Mean Fast-BLiMP rises from 50.2 at initialization to approximately 53.6 within the first 250 steps, around half an epoch, and improves little thereafter. This early saturation leaves limited room for the curriculum orderings to separate.

Several staged conditions subsequently decline. Developmental and anti-developmental staging peak between steps 1,250 and 1,500 and lose 2.5–3.7 points by the final checkpoint. Readability/staged loses 2.2 points, while the fully reshuffled random baseline loses 1.3. In contrast, the SmolLM2 teacher-perplexity/staged condition loses only 0.4 points and retains the highest final mean.

The random-block control tests whether this decline requires age-homogeneous content. It reaches a mean Fast-BLiMP peak of 56.2 at step 500 and then declines by 3.4 points, despite its blocks containing randomly assigned samples. Its mean late-training slope is also more negative than random’s (−0.50 versus +0.03 points per 1,000 steps; Appendix G). The comparable decline shows that age-based grouping is not necessary for the pattern and is consistent with repeated partition membership contributing to it.

This control does not identify a mechanism or show that every fixed partition is harmful.

The SmolLM2 teacher-perplexity partition remains comparatively stable, whereas the Qwen3-teacher partition declines by 2.9 points after an early peak. Fixed partition reuse is therefore associated with the late degradation but is not sufficient to explain when it occurs.

5.4 Visual inputs do not consistently explain the BLiMP ranking, and grounding remains weak

Text-only control. We compare the full VLMs with text-only models trained on the same annotation targets under developmental, teacher-perplexity, and random ordering. The seed-paired VLM-minus-text-only BLiMP differences are $+0.5 \pm 1.1$ for random, $+2.1 \pm 2.7$ for teacher perplexity, and -1.9 ± 0.6 for developmental ordering. The direction of the difference therefore depends on the ordering, and the seed variation is substantial for the positive comparisons. Removing image tokens does not produce a consistent shift in BLiMP or explain the curriculum ranking. This control is not an equivalence test and does not rule out smaller modality effects.

Multimodal grounding. Across the 12 evaluated configurations, Winoground remains at its 50% chance level, with condition means between 49% and 53%. VQA requires selecting among eight candidate answers, giving a chance accuracy of 12.5%. Condition means range from 20% to 31%, but performance varies substantially across seeds, with standard deviations as large as 8 percentage points.

For the staged core orderings and the random-block control, the mean seed-paired VQA differences relative to random are positive. However, the pattern does not support a coherent curriculum effect: anti-developmental and teacher-perplexity ordering produce similarly high means, and the VQA ranking aligns with neither developmental order nor the BLiMP results.

We therefore find no reliable or interpretable evidence that a particular curriculum improves multimodal grounding. Winoground provides no evidence of compositional grounding above chance, while VQA performance is above chance but variable and does not support a coherent ordering effect. The linguistic effects, including the teacher-perplexity BLiMP advantage, do not translate into a consistent multimodal advantage. Appendix J reports the complete results.

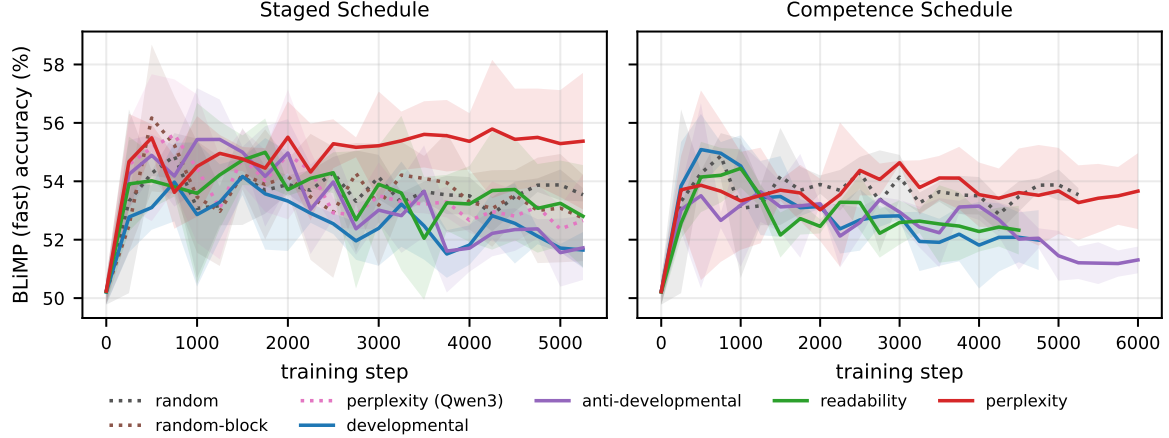


Figure 3: Fast-BLiMP accuracy over 10-epoch training, reported as the mean over three seeds with shading for ± 1 standard deviation. The random curve is repeated in both panels for comparison. The staged panel also includes the random-block and Qwen3-teacher controls. Several conditions with fixed partitions decline late in training, while the SmolLM2 teacher-perplexity/staged condition remains comparatively stable.

6 Discussion

Developmental order is not necessarily learner-relative difficulty. Our results do not imply that children’s literature lacks meaningful developmental structure. They show that inferred reader age and the resulting readability progression do not necessarily correspond to what a small language model finds predictable or useful during training. Age grading reflects communicative, lexical, and conceptual demands on human readers, whereas teacher NLL measures conditional predictability. This mismatch offers one plausible interpretation of the age-ordering result, although our experiments do not establish that it causes the absence of an age-ordering advantage. More broadly, this finding complements prior work in BabyLM and related developmentally inspired settings showing weak or setting-dependent curriculum effects (Martinez et al., 2023; Saha et al., 2024; Feng et al., 2024): a human-motivated progression is not automatically an effective model-training curriculum.

Curriculum effects combine ranking, task composition, and schedule. Teacher-perplexity ordering produces the highest mean BLiMP scores, but its uneven transfer across teacher families does not support treating it as a portable difficulty measure. It also changes the order of annotation formats, placing question-answering and reasoning samples predominantly before captions. The random-block results introduce a second complication: they suggest that outcomes may also depend on how fixed sample groups are reused across

epochs, not only on the signal used to rank them. At the same time, the comparatively stable SmolLM2 teacher-perplexity staged schedule shows that fixed partitions alone do not explain the observed dynamics. Taken together, the tested curricula alter sample ranking, task composition, and repeated grouping across epochs. The present experiments cannot isolate which of these components accounts for the teacher-perplexity advantage or the late-training decline.

Scope and implications. The text-only controls provide no consistent evidence that visual inputs explain the BLiMP ranking, while the grounding evaluations remain weak and do not reveal a coherent curriculum effect. Our conclusions therefore concern linguistic curriculum behavior within a multimodal training setup, not successful grounded language acquisition from vision. Our results suggest that curriculum difficulty must be defined relative to the learner and objective while remaining separable from task format and schedule construction. In this setting, inferred age organizes the generated data along the intended readability progression but does not provide a reliable curriculum rule; teacher perplexity provides a stronger empirical signal, but the source of that advantage remains unresolved.

7 Conclusion

We introduced TinyLibrary and used it to test whether inferred reader age provides an effective curriculum signal for small vision–language mod-

els under matched token budgets. Across two training budgets, neither developmental nor reverse-age ordering yields a stable BLiMP advantage over random ordering, and no ordering produces a reliable grounding benefit. Teacher-perplexity ordering achieves the highest mean BLiMP scores, but its uneven transfer across teacher families and entanglement with annotation format do not support treating it as a portable measure of model-relative difficulty. Overall, developmentally meaningful structure for human readers does not automatically translate into an effective model-training curriculum. Curriculum effects can depend on the signal used to rank samples, the task formats that the ranking places earlier or later, and how the ranked samples are presented across epochs.

Limitations

Our conclusions concern synthetic age-conditioned annotations derived from model-inferred age bands, not the books’ original prose or publisher metadata. We test a single vision-language model architecture and scale, so the findings may not generalize to other model families or sizes.

Ethics Statement

The source books are works from the International Children’s Digital Library. We do not distribute the books or the page-image set; the three pages in Figure 4 are shown for illustration of what the corpus looks like. We release the processing pipeline, the synthetic annotations, difficulty scores, curriculum order files, and training code, which suffice to reproduce our comparisons on any similarly structured corpus. The synthetic annotations were generated by Llama-3 and may contain model biases; they are intended for research on training dynamics, not as children’s educational material. No personal or user data is involved.

Acknowledgments

The author thanks Yuki M. Asano for supervision and guidance. The author also thanks SURF (<https://www.surf.nl>) for the support in using the National Supercomputer Snellius.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček,

Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. 2025. [SmolLM2: When smol goes big – data-centric training of a small language model](#). *ArXiv preprint*, abs/2502.02737.

BabyLM Organizing Committee. 2026. [BabyLM turns 4: Call for papers for the 2026 BabyLM workshop](#). *ArXiv preprint*, abs/2602.20092.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. [Curriculum learning](#). In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, pages 41–48. ACM.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Wilcox, and Adina Williams. 2025a. [BabyLM turns 3: Call for papers for the 2025 BabyLM workshop](#). *ArXiv preprint*, abs/2502.10645.

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. 2025b. 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.

Allison Druin. 2005. What children can teach us: Developing digital libraries for children with children. *The Library Quarterly*, 75(1):20–41.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. 2020. [PP-OCR: A practical ultra lightweight OCR system](#). *ArXiv preprint*, abs/2009.09941.

Ronen Eldan and Yuanzhi Li. 2023. [TinyStories: How small can language models be and still speak coherent English?](#) *ArXiv preprint*, abs/2305.07759.

Jeffrey L. Elman. 1993. Learning and development in neural networks: the importance of starting small. *Cognition*, 48(1):71–99.

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*. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. 2024. [The Llama 3 herd of models](#). *ArXiv preprint*, abs/2407.21783.

- Wenyi Hong, Weihang Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. 2024. [CogVLM2: Visual language models for image and video understanding](#). *ArXiv preprint*, abs/2408.16500.
- 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](#). *ArXiv preprint*, abs/2412.05149.
- Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. [BabyBERTa: Learning more grammar with small-scale child-directed language](#). In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.
- Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R.T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyurek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. 2024. [Elements of world knowledge \(EWoK\): A cognition-inspired framework for evaluating basic world knowledge in language models](#). *ArXiv preprint*, abs/2405.09605.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. [Mistral 7B](#). *ArXiv preprint*, abs/2310.06825.
- Najoung Kim and Sebastian Schuster. 2023. [Entity tracking in language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.
- J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command, Millington TN Research Branch.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Richard Diehl Martinez, Hope McGovern, Zebulon Goriely, 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, Singapore. Association for Computational Linguistics.
- Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. [COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.
- Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. [Competence-based curriculum learning for neural machine translation](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1162–1172, Minneapolis, Minnesota. Association for Computational Linguistics.
- Douglas L.T. Rohde and David C. Plaut. 1999. Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition*, 72(1):67–109.
- Rohan Saha, Abrar Fahim, Alona Fyshe, and Alex Murphy. 2024. Exploring curriculum learning for vision-language tasks: A study on small-scale multimodal training. In *Proceedings of the BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*. *ArXiv*:2410.15509.
- Saber Sheybani, Himanshu Hansaria, Justin Wood, Linda Smith, and Zoran Tiganj. 2023. Curriculum learning with infant egocentric videos. *Advances in Neural Information Processing Systems*, 36:54199–54212.
- Michael Tschanen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, et al. 2025. [SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features](#). *ArXiv preprint*, abs/2502.14786.
- Wai Keen Vong, Wentao Wang, A. Emin Orhan, and Brenden M. Lake. 2024. Grounded language acquisition through the eyes and ears of a single child. *Science*, 383(6682):504–511.
- Shengao Wang et al. 2025. [BabyVLM: Data-efficient pretraining of VLMs inspired by infant learning](#). *ArXiv preprint*, abs/2504.09426.
- 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, Singapore. Association for Computational Linguistics.

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.

Luis Wiedmann, Andrés Marafioti, Lewis Tunstall, et al. 2025. nanoVLM: The simplest repository to train your VLM in pure PyTorch. <https://github.com/huggingface/nanoVLM>.

Chen Yu and Linda B Smith. 2012. Embodied attention and word learning by toddlers. *Cognition*, 125(2):244–262.

A Illustrated-page filtering.

Starting from 251,410 scanned pages, we apply three filters to identify pages containing substantial visual content beyond printed text. First, we sum the areas of the text boxes detected by PaddleOCR (Du et al., 2020) and divide by the total page area. We remove pages for which this estimated text coverage is at least 30%.

Second, we convert each page to grayscale and compute a normalized Canny edge map using thresholds of 100 and 200. We define the page’s edge-content score as the standard deviation of this map and remove pages with scores below 0.2 as near-blank.

Finally, we create a temporary copy of each remaining page in which the detected text regions are filled with a mean color sampled from the surrounding image. We recompute the edge-content score and retain pages with scores of at least 0.3. This final step removes pages whose visual complexity is primarily attributable to printed text. The modified copies are used only for filtering; the original page images are retained for training. The complete procedure yields 46,721 illustrated pages.

B Corpus Examples

Figure 4 shows three source pages; Figure 5 shows the age-conditioned register realized for one page

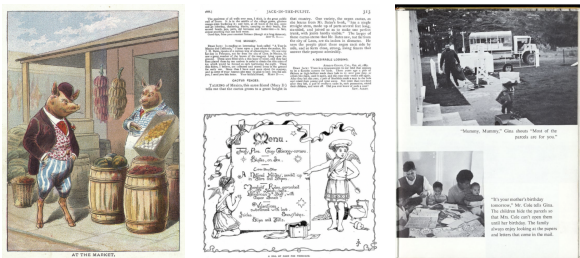


Figure 4: Illustrative sample pages from the corpus. Pages are shown for illustration only.

Captions generated for the *same* page, by band:

3–5: Pigs buy fruits at the market.

6–8: Pigs at the market: one pig sells fruits, the other talks.

9–12: Two well-dressed pigs in fancy clothes at a market.

12+: Two pigs, one stylishly dressed, the other selling fruits.

Reasoning, 3–5: “Why so many parcels?” — “It’s almost her mom’s birthday. . . Gifts come on special days.”

VQA, 12+: “Why is the cactus important?” — “The letter mentions a cactus that grows in Mexico. . . notable for its unusual height. . .”

Figure 5: Age-conditioned synthetic annotations.

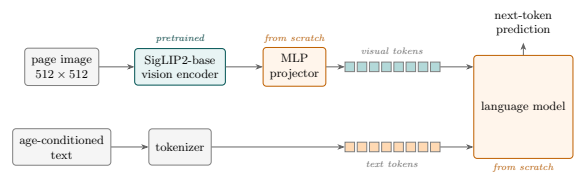


Figure 6: Model architecture. Only the vision encoder is pretrained; the projector and language model are trained from scratch on TinyLibrary. Components are subsequently trained jointly.

across all four bands.

C Training Details

Training uses bfloat16 precision on a single NVIDIA A100 40GB GPU, with a micro-batch size of 2 and eight gradient-accumulation steps. We use AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay 0.01) and clip gradients to a maximum norm of 1.0. Peak learning rates are 5.12×10^{-3} for the projector, 5×10^{-5} for the vision encoder, and 3×10^{-4} for the language model. Each learning rate uses a 3% linear warmup followed by cosine decay to 10% of its peak value. The step limit is set to 97% of the estimated capacity of each generated training sequence to prevent the data iterator from restarting. Final evaluations use the last checkpoint saved at a 250-step interval. Training takes approximately three hours for the 5-epoch budget and six hours for the 10-epoch budget.

D Five-Epoch Grid

Table 3 reports the complete results at the 5-epoch budget. Figure 7 shows the corresponding checkpoint trajectories.

E Seed-Matched Differences from Random Ordering

As the same seeds are used across conditions, each run can be compared with the random-ordering run

Order	Schedule	BLiMP	Supplement	EWoK	COMPS	Entity tracking
Random	—	53.2 $\pm$ 0.9	46.0 $\pm$ 1.6	50.1 $\pm$ 0.7	50.4 $\pm$ 0.3	25.0 $\pm$ 6.2
Developmental	Staged	53.9 $\pm$ 0.5	45.8 $\pm$ 0.8	50.2 $\pm$ 0.2	50.0 $\pm$ 0.4	27.7 $\pm$ 14.1
Developmental	Competence	53.3 $\pm$ 1.7	46.6 $\pm$ 0.7	50.2 $\pm$ 0.8	50.1 $\pm$ 0.4	32.6 $\pm$ 13.7
Anti-developmental	Staged	54.4 $\pm$ 1.9	45.6 $\pm$ 0.8	50.0 $\pm$ 0.9	49.9 $\pm$ 0.1	26.2 $\pm$ 14.8
Anti-developmental	Competence	52.2 $\pm$ 0.9	45.8 $\pm$ 2.3	50.0 $\pm$ 0.6	50.2 $\pm$ 0.5	30.8 $\pm$ 5.6
Readability	Staged	53.1 $\pm$ 0.7	47.6 $\pm$ 3.3	50.3 $\pm$ 0.7	50.1 $\pm$ 0.6	38.5 $\pm$ 3.6
Readability	Competence	52.8 $\pm$ 1.5	45.6 $\pm$ 3.5	49.3 $\pm$ 0.3	50.0 $\pm$ 0.3	28.9 $\pm$ 11.7
Teacher perplexity	Staged	54.8 $\pm$ 2.0	47.4 $\pm$ 1.3	50.5 $\pm$ 0.1	50.0 $\pm$ 0.1	27.5 $\pm$ 13.8
Teacher perplexity	Competence	55.6 $\pm$ 1.0	47.9 $\pm$ 1.8	50.8 $\pm$ 0.2	50.0 $\pm$ 0.7	20.6 $\pm$ 5.6

Table 3: Zero-shot accuracy (%) at the 5-epoch budget, reported as mean $\pm$ standard deviation across three seeds. Bold indicates the highest mean BLiMP score.

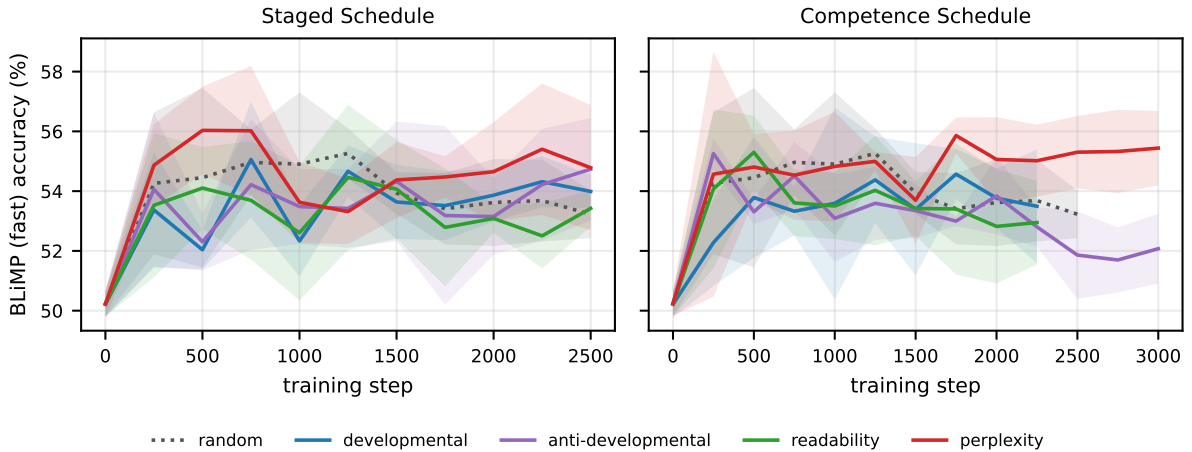


Figure 7: Fast-BLiMP accuracy over the 5-epoch budget, reported as the mean across three seeds with shading for ± 1 standard deviation. The random curve is repeated in both panels. The late declines visible in several 10-epoch staged runs are less developed at this shorter budget.

using the same seed and training budget. Table 4 reports these paired BLiMP differences; positive values favor the listed condition. With only three seeds, we treat the differences as descriptive rather than as formal statistical tests.

F Robustness across Teacher Models

Figure 9 visualizes the mismatch between inferred age order and teacher NLL under the primary SmolLM2-135M teacher. To test whether this pattern depends on the teacher model, we rescore every sample using Qwen3-0.6B and Gemma-3-270M. We follow the same procedure for all three teachers, computing the mean per-token NLL of the target text conditioned on the prompt. As the models use different tokenizers, we compare their band-level patterns and sample rankings rather than their absolute NLL values.

All three teachers produce the same band-level pattern, with the highest NLL for the 3–5 band, the lowest NLL for the middle bands, and an increase

for the 12+ band. Their sample-level rankings are also strongly correlated with the SmolLM2 ranking, with Spearman correlations of 0.91 for Qwen3 and 0.96 for Gemma-3. Although the samples assigned to the easiest and hardest quartiles are not identical across teachers. Agreement is lower for the easiest quartile, particularly between SmolLM2 and Qwen3.

Despite this agreement in estimated difficulty, the corresponding curriculum benefit does not transfer consistently. As reported in Table 2, the Gemma-derived ordering retains a smaller BLiMP advantage over random, whereas the Qwen3-derived ordering retains no advantage. Cross-teacher agreement in sample difficulty therefore does not guarantee equivalent training outcomes.

G Fixed-Partition Analysis

The random-block control tests whether the late decline observed in several staged conditions requires samples to be grouped by difficulty. Both ran-

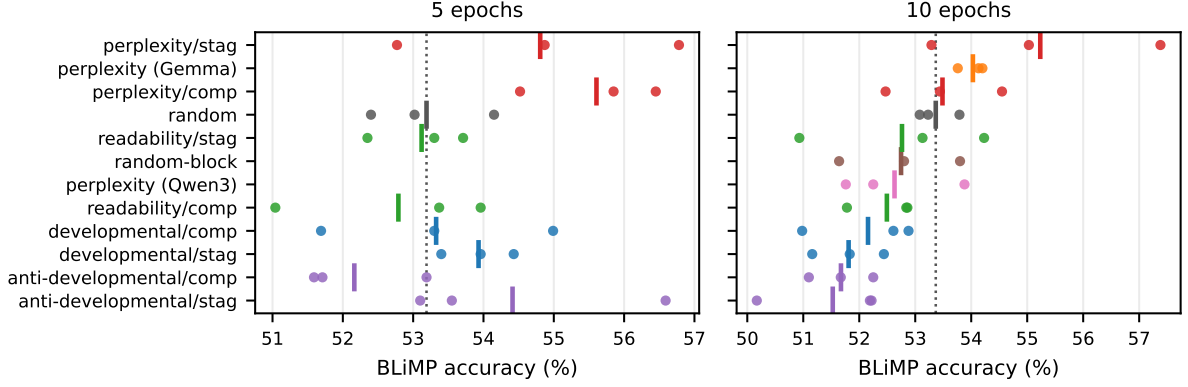


Figure 8: Final BLiMP scores by seed at the 5-epoch and 10-epoch budgets. Dots represent individual seeds, horizontal marks represent condition means, and dotted lines represent the random-ordering mean. The controls were trained only at the 10-epoch budget.

Condition	5-epoch budget		10-epoch budget	
	Seed differences		Seed differences	Mean
Developmental / Staged	+2.0, +0.4, -0.2	+0.74	-2.6, -0.6, -1.4	-1.56
Developmental / Competence	-0.7, +0.3, +0.8	+0.14	-0.9, -2.1, -0.6	-1.21
Anti-developmental / Staged	+0.7, +0.5, +2.4	+1.22	-1.6, -2.9, -1.0	-1.84
Anti-developmental / Competence	-0.7, -1.4, -1.0	-1.03	-2.7, -0.8, -1.6	-1.69
Readability / Staged	+0.9, +0.7, -1.8	-0.07	-2.9, +1.2, -0.1	-0.60
Readability / Competence	+1.6, -2.0, -0.8	-0.40	-2.0, -0.2, -0.4	-0.87
Teacher perplexity / Staged	+0.4, +1.9, +2.6	+1.62	-0.5, +2.0, +4.2	+1.87
Teacher perplexity / Competence	+3.5, +1.5, +2.3	+2.42	+0.8, +0.4, -0.8	+0.12
Random-block / Staged	—	—	-1.0, -1.4, +0.6	-0.62
Teacher perplexity (Gemma) / Staged	—	—	+0.4, +1.1, +0.5	+0.66
Teacher perplexity (Qwen3) / Staged	—	—	-2.0, +0.8, -1.0	-0.74

Table 4: Seed-matched BLiMP differences from random ordering, in percentage points. The mean is computed across the three paired differences.

	SmolLM2 135M	Qwen3 0.6B	Gemma-3 270M
NLL, band 3–5	2.86	3.14	2.96
NLL, band 6–8	2.54	2.67	2.62
NLL, band 9–12	2.52	2.57	2.57
NLL, band 12+	2.75	2.71	2.76
Spearman ρ vs. SmolLM2	—	0.91	0.96
Easiest-quartile overlap	—	0.70	0.82
Hardest-quartile overlap	—	0.91	0.92

Table 5: Mean per-token NLL by inferred age band and sample-level agreement with the SmolLM2-derived ordering. All three teachers reproduce the same band-level pattern, but their sample rankings are not identical.

dom and random-block ordering present every sample once per epoch without using a content-based ranking. Random ordering reshuffles the complete corpus each epoch, whereas random-block assigns samples to four fixed groups matching the token masses of the age-band stages and reuses those assignments across epochs. The comparison

therefore isolates the association between repeated group membership and training dynamics.

Because peak-to-final differences can exaggerate decline by selecting a noisy maximum, we also calculate two measures that do not depend on identifying a peak. The first is the linear Fast-BLiMP slope over the second half of training, expressed in points per 1,000 steps. The second is the difference between mean performance during the final fifth of training and mean performance from steps 500–1,250.

Random-block has a more negative mean slope and late–early difference than the fully reshuffled random baseline. This pattern is consistent with repeated group membership contributing to late degradation even when the groups contain randomly selected samples.

The effect is not uniform across fixed-partition conditions. Anti-developmental ordering also declines, whereas teacher-perplexity ordering re-

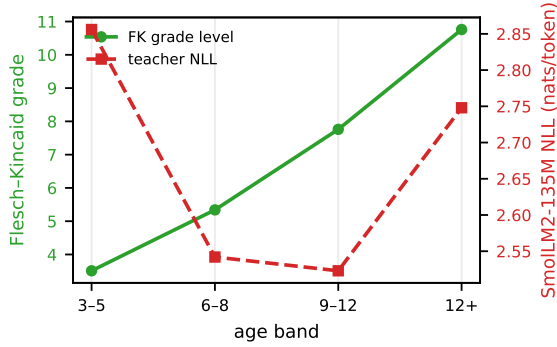


Figure 9: Readability and teacher-model perplexity encode different age-band progressions. Flesch–Kincaid grade increases monotonically with inferred age, whereas mean per-token NLL under the primary SmolLM2-135M teacher is highest for the youngest and oldest bands and lowest for the middle bands.

Condition	Late-half slope/1k	Late–early
Random	+0.03 $\pm$ 0.33	-0.32 $\pm$ 0.55
Random-block	-0.50 $\pm$ 0.42	-1.38 $\pm$ 0.97
Teacher perplexity	+0.06 $\pm$ 0.65	+0.83 $\pm$ 2.08
Developmental	-0.22 $\pm$ 0.39	-1.13 $\pm$ 1.11
Anti-developmental	-0.45 $\pm$ 0.14	-2.94 $\pm$ 0.86
Readability	+0.01 $\pm$ 0.81	-0.60 $\pm$ 2.62

Table 6: Late-training Fast-BLiMP dynamics, reported as mean $\pm$ standard deviation over three seeds. Late-half slope is calculated over the second half of training; late–early compares the final fifth of training with steps 500–1,250.

mains approximately flat; readability has a near-zero mean slope but substantial variation across seeds. Fixed-partition reuse is therefore associated with late degradation in some conditions, but it is not sufficient to explain when degradation occurs.

H Teacher-NLL Quartiles Largely Track Annotation Format

The annotation formats differ substantially in predictability under the SmolLM2-135M teacher. Captions have a mean per-token NLL of 3.64, compared with 2.25 for visual question answering and 2.29 for reasoning. Consequently, the equal-sized NLL quartiles largely separate samples by annotation format. Under ascending teacher-NLL ordering, VQA and reasoning samples are therefore presented predominantly earlier in training, while captions are presented predominantly later.

I Entity-Tracking Results

Entity tracking does not provide a stable basis for comparing the curriculum orderings. On the full

NLL quartile	Caption	VQA	Reason.	Med. tok.
Q1 (easiest)	0.3%	52.1%	47.6%	211
Q2	0.7%	51.9%	47.5%	289
Q3	31.6%	30.0%	38.4%	215
Q4 (hardest)	99.6%	0.1%	0.3%	73

Table 7: Annotation-format composition of the equal-sized teacher-NLL quartiles. The two lowest-NLL quartiles consist almost entirely of VQA and reasoning samples, whereas the highest-NLL quartile consists almost entirely of captions.

evaluation, the untrained reference scores 41.8 ± 0.1 , whereas the trained models score between 17.1 and 36.7 at the 10-epoch budget. Performance also varies substantially across seeds, with standard deviations as large as 15.2 points.

The checkpoint trajectories in Figure 10 show that scores generally decrease during training, but our experiments do not establish why. The pattern could reflect properties of the evaluation at this model scale, the architecture, or the training data. Without a known strong reference model under the same evaluation protocol, the untrained score cannot be interpreted as evidence of entity-tracking ability. We therefore do not use this metric to draw conclusions about the relative effectiveness of the curriculum orderings.

J Multimodal and Text-Only Controls

Multimodal evaluation. Following the official BabyLM evaluation pipeline, we rank candidate answers by their conditional log-likelihood given the image. This evaluation uses the complete VLM. VQA contains 25,230 questions with approximately eight candidate answers each, while Winoground contains 746 two-way comparisons.

Winoground remains near chance across all orderings. VQA performance exceeds chance but varies substantially across seeds. Relative to seed-matched random runs, the staged differences are $+5.7 \pm 4.9$ for developmental, $+8.6 \pm 5.5$ for anti-developmental, $+2.8 \pm 2.0$ for readability, $+8.0 \pm 5.3$ for teacher perplexity, and $+4.9 \pm 3.4$ for random-block.

Although these mean differences are positive, their ranking does not follow developmental order, teacher perplexity, or the BLiMP results. Together with the chance-level Winoground scores and substantial seed variation, they do not provide coherent evidence that a particular ordering improves multimodal grounding.

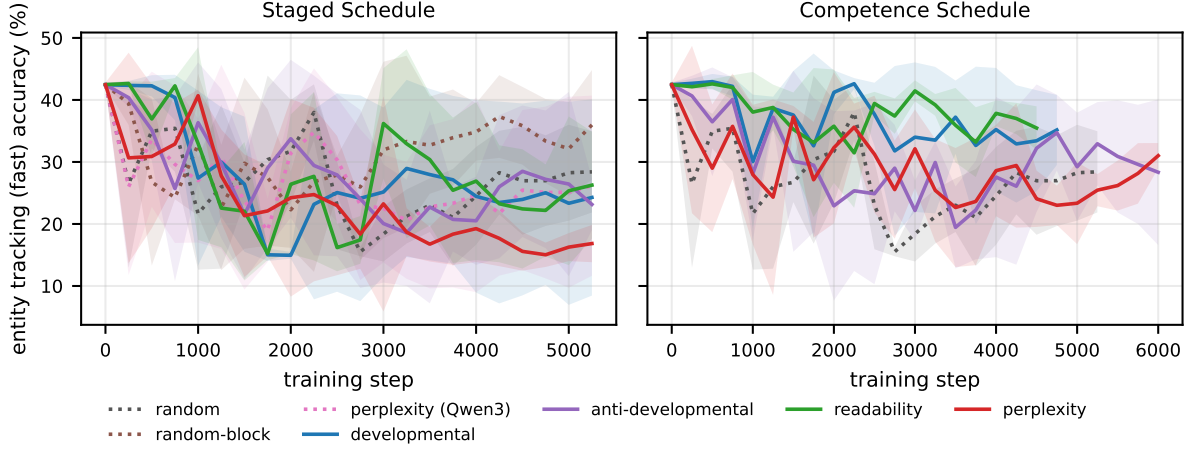


Figure 10: Fast entity-tracking accuracy over 10 epochs, reported as the mean across three seeds with shading for ± 1 standard deviation. Performance generally decreases from its initial level and varies substantially across seeds.

Ordering and schedule	VQA	Winoground
Random	20.3 $\pm$ 1.2	52.2 $\pm$ 1.7
Random-block / staged	25.2 $\pm$ 4.6	52.5 $\pm$ 1.1
Developmental / staged	26.0 $\pm$ 5.6	51.5 $\pm$ 0.5
Anti-developmental / staged	28.9 $\pm$ 4.8	50.8 $\pm$ 0.9
Readability / staged	23.1 $\pm$ 2.3	52.0 $\pm$ 1.9
Teacher PPL / staged	28.3 $\pm$ 4.8	52.6 $\pm$ 0.4
Qwen3 PPL / staged	25.5 $\pm$ 2.5	51.7 $\pm$ 2.6
Gemma PPL / staged	23.3 $\pm$ 0.3	51.6 $\pm$ 1.2
Developmental / competence	21.8 $\pm$ 8.7	52.4 $\pm$ 1.2
Anti-developmental / competence	23.8 $\pm$ 6.9	51.5 $\pm$ 1.9
Readability / competence	31.3 $\pm$ 4.7	49.2 $\pm$ 2.0
Teacher PPL / competence	30.6 $\pm$ 7.5	50.8 $\pm$ 0.5
Chance	12.5	50.0

Table 8: Multimodal zero-shot accuracy (%) after 10 epochs, reported as mean $\pm$ standard deviation over three seeds.

Text-only control. We train text-only models under developmental, teacher-perplexity, and random ordering using the same decoder architecture and annotation targets but without image tokens. The training budget is adjusted to preserve the same exposure to annotation text. After 10 epochs, the text-only models score 53.7 BLiMP under developmental ordering, 53.2 under teacher-perplexity ordering, and 52.9 under random ordering. The corresponding seed-paired VLM-minus-text-only differences are -1.9 ± 0.6 , $+2.1 \pm 2.7$, and $+0.5 \pm 1.1$. Because the direction varies across orderings, the control provides no evidence of a uniform benefit or penalty from visual inputs. It does not establish modality independence: the teacher-perplexity advantage over random decreases from 1.87 points in the VLM runs to 0.27 points in the text-only runs.