

Learning from Correction: Data-Efficient Language Model Training Through Adaptive Teacher-Model Feedback

Sadri Islam¹, Tagore Rao Kosireddy², Kevin Li², Dylan Gaines¹,

¹Department of Computer Science, Kennesaw State University, Marietta, Georgia, USA

²Department of Computer Science, Michigan Technological University, Houghton, Michigan, USA

Correspondence: dgaine20@kennesaw.edu

Abstract

We present our submission to the STRICT-SMALL track of the 2026 BabyLM Challenge. We study whether a developmentally-inspired curriculum, particularly an adaptive teacher-feedback curriculum based on the student’s own productions, can improve language model pre-training under a 10M word budget. We trained a 58M-parameter student from scratch under five separate conditions that shared the same architecture and tokenizer, isolating different aspects of the training data and curriculum structure. We built an official baseline trained on a cleaned version of the official STRICT-SMALL corpus and also a synthetic baseline trained on teacher model generated data. Then, we developed a fixed curriculum to train a model on the same synthetic child-directed text in three developmental stages: infant, toddler, and child. Finally, we introduced an adaptive feedback curriculum in which the teachers generate additional training data by correcting the student’s productions during training, which adapts the training data according to the learner’s level of understanding. The fixed curriculum demonstrated superior performance on the BLiMP-Supplement, EWoK, and COMPS benchmarks, while the adaptive feedback curriculum substantially improved the Entity Tracking performance in one configuration. However, our official baseline model achieved the highest BLiMP, SuperGLUE, and overall macro-average scores, indicating that our curriculum methods did not produce the overall improvements that we had hypothesized. Our code is available at <https://github.com/dcgaines/babylm-2026>.

1 Introduction

Human children acquire language from an estimated two to seven million words per year, amounting to fewer than one hundred million words by early adolescence (Gilkerson et al., 2017). The learning regimes of modern language models are

disproportionate by any such measure: Chinchilla was trained on 1.4 trillion tokens (Hoffmann et al., 2022), and Llama 3.1 on 15.6 trillion (Grattafiori et al., 2024). Training at this scale is economically and ecologically costly and places state-of-the-art pretraining beyond the reach of most academic groups (Bender et al., 2021). The BabyLM Challenge confronts this efficiency gap directly, incentivizing researchers to optimize pretraining under data limitations inspired by human development (Warstadt et al., 2023; Hu et al., 2024; Charpentier et al., 2025). We participate in the STRICT-SMALL track, which caps pretraining at ten million words of unique data and evaluates submissions on a shared suite of metrics spanning grammatical knowledge, world knowledge, and downstream language understanding.

Conventional pretraining remains very different than the conditions of a human learning a language. The corpus is assembled, shuffled, and consumed on a fixed timetable, independent of the state of the learner. In real life, when a child produces an incorrect utterance, caregivers or teachers respond to and reformulate it. The caregiver frequently answers with a corrected rendering of the same intended message, so that a meaningful fraction of the child’s “training data” is adapted from their own productions (Chouinard and Clark, 2003; Clark, 2018).

In this work, we investigate whether a language model can benefit from a similar reformulation loop when limited training data is available. We also explore the efficacy of building a foundation for the student model through a curriculum of increasing grammatical complexity.

We conduct this comparison under a fixed 10M word budget, with the same architecture and tokenizer, so that curriculum structure is the only variable. Though the BASELINE retains the strongest grammatical knowledge, the ADAPTIVE CURRICULUM substantially improves the entity tracking

and inference-oriented fine-tuning. This suggests that correction-based feedback during pretraining makes referential reasoning and downstream adaptability stronger, though it might not help to improve zero-shot grammaticality.

2 Related Work

Curriculum learning. BabyLM participants have commonly explored developmental plausibility through curriculum learning (Bengio et al., 2009). Previous submissions have scheduled the vocabulary, data, and training objective over the course of training (Diehl Martinez et al., 2023), ordered corpora according to cross-linguistic theories of acquisition (Salhan et al., 2024), and masked tokens of different part-of-speech categories in a staged sequence that follows the typical word-acquisition order of children (Lucas et al., 2024). This accumulated evidence suggests that curricula deliver gains on particular task families but seldom lift aggregate performance above well-tuned non-curriculum models (Hu et al., 2024; Charpentier et al., 2025). These schedules share a structural property that we believe limits them: they are written before training begins. Such curricula advance as scheduled, whether or not the student has mastered the current stage, and have no means of compensating by directing additional effort toward the concepts the student struggles with.

Teacher-student modeling. A second line of work trains a small student model under the guidance of one or more larger teacher models. The idea originates in knowledge distillation (Hinton et al., 2015), where a compact student is trained to match the output distribution of a larger teacher. Here, the teacher supervises the student’s outputs without ever observing them. An interactive setup was also recently introduced in BabyLM (Mayer Martins et al., 2025), where a teacher model guided a small student through a storytelling task rather than transferring the teacher’s output distributions.

Sample-efficient language modeling. The BabyLM Challenge has encouraged language modeling research under training data limits, where the goal is not only to scale the models and corpora but also to improve how efficiently a model learns from limited input. In this setting, Timiryasov and Tastet (2023) introduced Baby Llama, a 58M-parameter student model trained with only 10M words. Their work showed that a compact

LLaMA-style student can benefit strongly from teacher guidance, especially when the training budget is too small for standard pretraining to be fully effective. Tastet and Timiryasov (2024) extended this direction with BabyLlama-2, using a larger student and ensemble-based teacher supervision to improve performance under the same motivation. In both systems, however, the feedback flows through logit distributions computed over the same static corpus where the teachers never observe output from the student.

Prior BabyLM systems show that compact student models can benefit from teacher-guided training. However, most existing approaches rely on fixed teacher outputs or training data prepared before student learning begins. Our work addresses this limitation by comparing a curriculum-free baseline, a fixed developmental curriculum, and an adaptive curriculum where the student learns from a combination of pre-generated developmental data and teacher-generated corrective examples based on its own outputs.

3 Methodology

In this work, we compare three training strategies for a student model based on the Baby Llama architecture (Timiryasov and Tastet, 2023) in the limited data setting of BabyLM 2026:

- **BASELINE** - Standard training with no curriculum order or teacher feedback. We train two BASELINE models: one on a cleaned version of the official corpus, and one on the same synthetic data used in the other strategies, randomly shuffled.
- **FIXED CURRICULUM** - A static developmental curriculum trained on synthetic, teacher-generated data. The curriculum progresses through three phases of increasingly complex text to determine whether developmental ordering alone improves the learning.
- **ADAPTIVE CURRICULUM** - An adaptive feedback curriculum that divides each developmental phase into a fixed foundation part and an adaptive corrective part which is generated during the training. The teacher models iteratively observe the student’s productions and generate corrective examples.

The following subsections describe each condition in more detail.

3.1 Common Architecture and Training Setup

We use the official BabyLM 2026 Strict-Small dataset (Choshen et al., 2026), available on Hugging Face as BabyLM-2026-Strict-Small, to train our tokenizer. The raw corpus combines child-directed speech, Simple Wikipedia, OpenSubtitles, and associated sources for a total of 9,999,996 words. Since some of these sources are derived via transcription and web scraping, the corpus includes some markup and formatting artifacts that are not a part of natural language. We measured the primary artifacts in the raw corpus using a diagnostic scan; the artifact classes, their counts, and an example of each are shown in Table 1.

Our cleaning removed all of these artifact classes while preserving the utterances after them, collapsed repeated whitespace, and discarded lines that became empty (i.e. the line contained only the artifact). This reduced the corpus from 9,999,996 to 9,016,506 words, a 9.8% reduction. Using the cleaned corpus, we trained a byte-pair encoding (BPE) tokenizer (Sennrich et al., 2016) with a vocab size of 16K.

All conditions used the same tokenizer and student architecture so that the comparison could focus on the training strategy rather than model size or architectural differences. The student was a Llama-style decoder-only language model implemented with Hugging Face Transformers (Wolf et al., 2020) and based on the Baby Llama architecture (Timiryasov and Tastet, 2023). The model had 16 transformer layers, a hidden size of 512, an intermediate size of 1024, 8 attention heads, 8 key-value heads, and a context length of 512 tokens, with a total of 58.34M parameters. All student models were trained from random initialization using causal language modeling.

Across all conditions, we used the same general training pipeline. We tokenized the text with the shared 16K tokenizer and truncated each example to at most 128 tokens with no concatenation or packing across the sentences. We trained all models with the Hugging Face Trainer using AdamW optimization (Loshchilov and Hutter, 2019), cosine learning-rate scheduling, a weight decay of 0.1, bfloat16 precision, and an effective batch size of 256. We used an initial learning rate of 3×10^{-4} for the BASELINE, FIXED CURRICULUM, and the first phase of the ADAPTIVE CURRICULUM. All training was performed using NVIDIA RTX 5000 Ada Generation GPUs with 32GB of memory. When

Count	Artifact	Example
473K	Speaker labels	*MOT: more what?
74.3K	Bracketed directions	[giving sugar to Colin]
35.5K	Subtitle dashes	- And she did for a while.
22.9K	Unclear speech markers	*CHI: xxx Fraser napkin.
16.4K	Wiki section markers	= = = Method Man = = =
10.5K	Annotation lines	%add: COL

Table 1: Artifact classes found in the raw corpus by the diagnostic scan, with counts and one example of each. Speaker labels mark who is speaking in a transcript, unclear-speech markers are standard placeholders for speech that was inaudible on the recording or intentionally left untranscribed, and annotation lines carry transcription metadata such as the addressee.

possible, training was distributed across two GPUs, modifying the per-device batch size to maintain a consistent effective batch size.

3.2 Baseline

We trained one baseline model (BASELINE OFFICIAL) on the cleaned official corpus with the shared 16K tokenizer. With a corpus size of 9M words, we initially trained the model for 10 epochs. However, after observing signs of over-training, we reduced to 5 epochs for our final model, which improved performance and resulted in 45M total words of exposure. We use this baseline to measure the impact of our synthetic data. We note that the choice of 5 epochs and other hyperparameters was informed by evaluation on the final BabyLM pipeline rather than a held-out development set, which is a limitation discussed further in Section 6.

To assess the contribution of curriculum ordering, we additionally trained BASELINE SYNTHETIC using the same synthetic data described in Section 3.3, but randomly shuffled without any curriculum. Since it uses the same architecture, cleaning, tokenizer, training process, and hyperparameters as the other conditions, it allows us to measure whether our curriculum-based approaches provide benefits beyond standard shuffled training. We combined the synthetic corpus of 9,999,997 words and shuffled the sentences. We used the same BabyLlama architecture, shared 16K tokenizer, and optimization settings as in the BASELINE OFFICIAL model. We trained the model for 5 epochs, corresponding to approximately 50M words of exposure.

3.3 Fixed Curriculum

Instead of the official BabyLM corpus, the **FIXED CURRICULUM** condition used sentences created by a teacher model for three developmental stages: infant, toddler, and child. We defined each developmental stage by a separate instruction prompt that controlled the linguistic complexity of the teacher-generated sentences.

The *infant* stage requested simple, child-directed speech aimed at children under four years of age. It emphasized single-clause sentences, usually following a subject–verb–object structure with present-tense, everyday verbs and concrete, familiar nouns. Abstract ideas, subordinate or relative clauses, and passive constructions were excluded. Although short sentences were encouraged, the prompt explicitly stated that this stage was defined by grammatical simplicity rather than sentence length, and therefore permitted simple coordination with *and*.

The *toddler* stage targeted an age level of approximately four to seven years and allowed one or two clauses joined by simple connectives such as *and*, *but*, *because*, *when*, and *so*. It also permitted descriptive adjectives, prepositional phrases, and a moderate everyday vocabulary, while restricting the content to concrete situations that it could understand.

The *child* stage targeted children aged approximately eight to twelve years old and introduced richer linguistic structures, including subordinate and relative clauses, varied verb tenses, and discourse connectives such as *however*, *therefore*, *although*, *despite*, *while*, and *whereas*. It also included abstract ideas and reasoning. At this stage, syntactic complexity rather than sentence length defined the target level.

The exact text for the prompts used can be found in Appendix A.

For every stage, we appended a manually-selected sentence from the official corpus as an example for the teacher model. The teacher was instructed to use it only as inspiration for topics and words, without copying, translating, or preserving its original meaning. We obtained outputs from two instruction-tuned teacher models, meta-llama/Llama-3.2-1B-Instruct (Grattafiori et al., 2024) and Qwen/Qwen2.5-1.5B-Instruct (Yang et al., 2024). We merged the output from the two teachers and removed duplicates to produce a

single dataset for each developmental stage.

The teachers first produced an initial corpus of approximately 12.3M words: 2.3M infant words, 5M toddler words, and 5M child words. We intentionally exceeded our 10M word budget with the initial teacher generation to create a buffer and allow for data cleaning. We cleaned this pool with the same process as the official corpus (as described in Section 3.2). In addition, we removed meta text from the instruction-tuned models, such as refusals, self-narration, and grammar commentary that did not follow the instructions for teacher-generated data. After cleaning, we subsequently downsampled the remaining corpus to the 10M word budget.

While downsampling, we used the complete cleaned generation pool for the infant phase. Then, we allocated the remaining budget equally between the toddler and child stages. The completed fixed-curriculum corpus comprised 2,301,826 infant words, 3,849,085 toddler words, and 3,849,086 child words, totaling 9,999,997 words which is a 23.0%, 38.5%, and 38.5% phase distribution for the infant, toddler and child stages. We allocated a lower proportion of our budget to the infant stage to avoid too much repetition because the simpler sentences tended to be shorter and addressed a smaller range of topics and ideas.

The student underwent sequential training across the developmental stages, completing 5 epochs in each stage before moving to the next. As with the **BASELINE OFFICIAL** model, we initially used 10 epochs, but found better performance with 5 epochs.

3.4 Adaptive Curriculum

We added an adaptive teacher feedback loop on top of the fixed curriculum for this approach. We divided each developmental stage into two phases, a fixed foundational phase and an adaptive feedback phase. We prepared the foundational phase before training to give the student an initial start for the current developmental stage. The adaptive feedback phase was generated throughout training by observing the student’s current outputs. In each developmental stage, the student first trained on the foundational phase for 5 epochs. Then, the student went through 10 iterations of the adaptive feedback phase, each training for 5 epochs. We used the same hyperparameters as previous conditions.

Each phase (both foundational and adaptive) was trained with a fresh optimizer initialization and cosine schedule, so that it began with a clean opti-

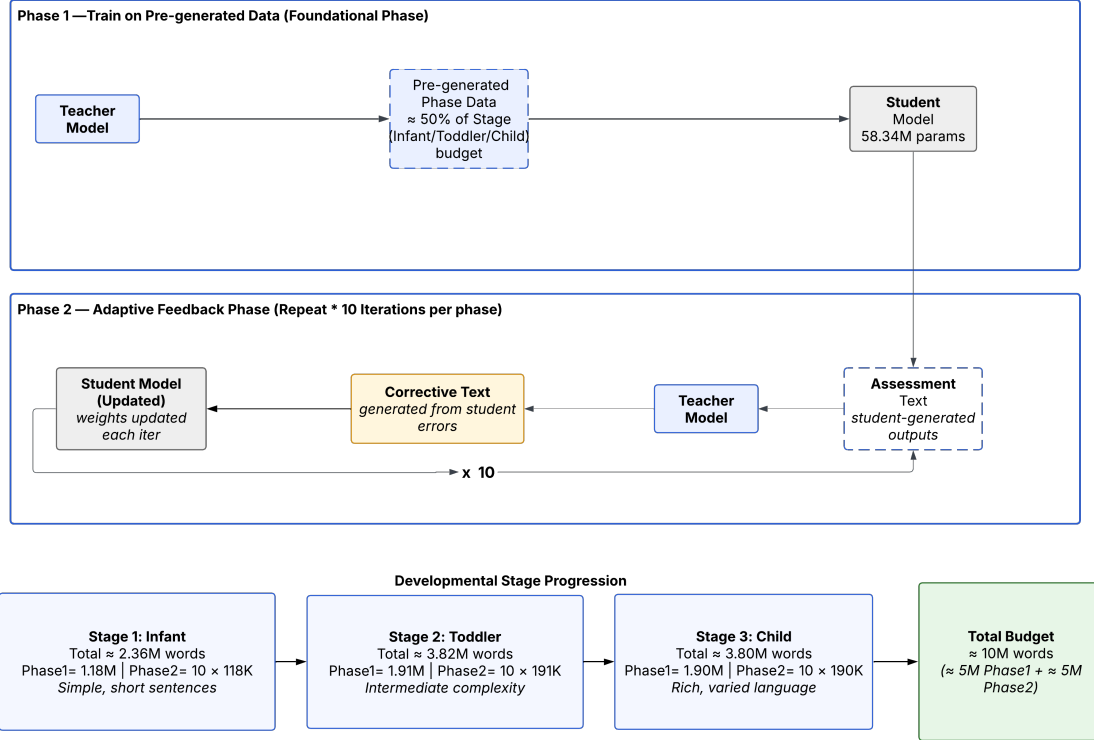


Figure 1: Overview of the adaptive feedback training process. The student is first trained on the foundational phase data. Then, it generates assessment text, which is provided to the teacher. The teacher produces corrective feedback examples, and the student is updated through repeated feedback iterations.

mization state and the model weights were carried forward. The combined budget for both phases was held to the same 10M-word budget as the FIXED CURRICULUM, so the two methods are directly comparable and differ only in how the data was constructed. We divided the budget equally between the two phases. The overall training process is shown in Figure 1.

As indicated in Section 4, this initial model did not perform as well as we had hoped. To try to allow for smaller changes during the feedback phase, we trained a variant of the ADAPTIVE CURRICULUM model that reduced the learning rate from 3×10^{-4} to 1×10^{-4} for the feedback phase (the foundation phase learning rate remained the same). We refer to this model as ADAPTIVE CURRICULUM (VARIABLE LR).

Foundational phase construction. We built the dataset for each foundational phase by randomly selecting half of the words from the same cleaned, teacher-generated corpus as the FIXED CURRICULUM model (Section 3.3). Here, the per-stage sizes were 1,180,005 infant words, 1,910,006 toddler words, and 1,900,006 child words, for a total of

4,990,017 foundational phase words.

Adaptive feedback phase construction. We performed each adaptive feedback iteration in the same manner. First, we sampled 3,000 seed lines from the pool of 452,928 lines in the cleaned official corpus between 5 and 30 words. We designed a series of prompts consisting of the token indicating the beginning of the sequence plus the first three words of each selected seed line. We used the student model to generate up to 20 new tokens from each prompt with sampling (temperature 0.9, top- p 0.95). These continuations were the student’s own, often poor productions, and they constituted the assessment text of the iteration.

Then, we used the two instruction-tuned teacher models, meta-llama/Llama-3.2-1B-Instruct (Grattafiori et al., 2024) and Qwen/Qwen2.5-1.5B-Instruct (Yang et al., 2024), to generate new training data with the goal of correcting mistakes made by the student. Each teacher produced half of the word budget of the iteration. Each student sentence was shown to a teacher once at a time. We asked the teacher to produce four or five grammatically

Stage	Student Generated Sentence	Teacher Suggested Sentences
Infant	<i>You want I to play with toy.</i>	<i>I want you to play with my toy. Please let me play with your toys. I need you to play with my teddy bear. Can you please play with my ball?</i>
Toddler	<i>they gonna want to ride the merry-go-round.</i>	<i>They want to ride the merry-go-round. The kids like to ride the merry-go-round. My friend wants to ride the merry-go-round. They want to ride the merry-go-round because they like it.</i>
Child	<i>Kleinsmith doing a great job in the field of secrecy.</i>	<i>Kleinsmith is doing a great job in the field of secrecy. Kleinsmith is very good at keeping secrets. In the field of secrecy, Kleinsmith is doing very well. Kleinsmith is a great secret keeper.</i>

Table 2: Examples of teacher feedback in the ADAPTIVE CURRICULUM. Each row shows one student-generated sentence and four selected teacher suggestions for the corresponding developmental stage.

accurate sentences expressing the same idea at the current developmental stage, with the same stage descriptions as the FIXED CURRICULUM condition. Example teacher corrections for each developmental stage are shown in Table 2, and the exact model prompts can be found in Appendix B.

The prompts also addressed two special cases. If the student-generated sentence was incoherent, the teacher was instructed to create well-formed sentences on a similar everyday topic. If the student sentence was already grammatically correct, the teacher was instructed to create new appropriate sentences on the same idea. Teachers decoded with temperature 0.75 and top- p 0.9, with a maximum of 128, 160, and 256 new tokens at the infant, toddler, and child levels. After training, we used our two training teachers, Qwen2.5-1.5B-Instruct and Llama-3.2-1B-Instruct, to independently classify all 90,000 student sentences recorded during the ADAPTIVE CURRICULUM training to assess how often these special cases occurred. We performed this assessment using the original 3,000 sentences from each of ten feedback iterations within each of three curriculum stages. We asked each teacher to assign one of three labels: *grammatically correct*, *coherent but with errors*, or *incoherent*. Our prompt included twelve examples presented in a mixed label order and used greedy decoding (full prompt in Appendix C). Qwen classified 40.10% of sentences as grammatically correct, 1.79% as coherent but with errors, and 58.11% as incoherent. Meanwhile, Llama assigned 3.11%, 1.27%, and 95.09% to these categories respectively. The two teachers agreed to the same label rating for 57.95% of the sentences. Most of these agreements were on sentences both labeled incoherent. Of the 36,088 sentences Qwen marked as correct, Llama agreed

on only 1,792 of them.

We split teacher output into one sentence per line, with common title abbreviations such as ‘Mr.’ and ‘Dr.’ protected from splitting. Each line was then run through a candidate filter. The filter stripped leading bullets, separators, and markdown emphases. Lines were removed if they matched the assistant-meta filter described in Section 3.3, had fewer than two words, or contained correction arrows, equals signs, parenthetical grammar comments, etc. A persistent key file recorded all sentences accepted in each iteration, and the teachers skipped any sentence that was already produced. This guaranteed that the adaptive feedback phase was filled with distinct sentences rather than near duplicates reproduced across the iterations.

Adaptive feedback phase budget. Each developmental stage’s (infant, toddler or child) adaptive feedback phase budget mirrored its foundational phase budget size. This adaptive feedback budget was split equally across ten iterations. This resulted in about 118K words per iteration in the infant stage, 191K in the toddler stage, and 190K in the child stage. After each iteration, we trained the student on the newly generated data for 5 epochs. Then, we used the updated student to generate the samples for the next iteration. This process trained the student model on 4,989,229 total adaptive feedback phase words across the 3 developmental stages, for a grand total of 9,979,246 training words for the model.

In this study, the teacher was used only to generate text examples. Unlike standard knowledge distillation approaches which directly transfer teacher distributions (Hinton et al., 2015), the student was not trained on teacher logits, hidden states, or probability distributions. The correction signal is signif-

icantly narrow, consisting of short, targeted corrections of the student’s own errors. The foundational phase provides broad lexical exposure, so the adaptive feedback phase can stay focused on correction.

Additional configurations. During our development process, we trained the ADAPTIVE CURRICULUM in a few different configurations, keeping the architecture, tokenizer, teachers, prompts, and filtering unchanged. This included a 10-epoch version — like in the other conditions, this did not work as well as the 5 epochs used in the final model. We also tried shifting the balance between the foundational phase and the adaptive feedback phase to 75/25 instead of 50/50, but this was not effective. In the final variant, we kept the 50/50 balance between the phases, but we mixed the teacher-generated foundational dataset with the cleaned official BabyLM corpus, taking 50% from the official corpus and 50% from the teacher models. This did not improve the model’s performance. Again, this did not improve the model’s performance.

4 Results

The BabyLM evaluation suite assesses models across multiple dimensions of linguistic competence, including grammatical knowledge (BLiMP), supplementary linguistic phenomena (BLiMP Supplement), world knowledge (EWoK), property knowledge and inheritance (COMPS), multi-step referential reasoning (Entity Tracking), and downstream language understanding (SuperGLUE finetuning).

Table 3 reports each model’s performance across the main evaluation metrics. We compare the BASELINE OFFICIAL, BASELINE SYNTHETIC, FIXED CURRICULUM, ADAPTIVE CURRICULUM, and ADAPTIVE CURRICULUM (VARIABLE LR) all using the 5-epoch configuration as our primary submission.

Zero-shot grammatical knowledge (BLiMP). BLiMP (Warstadt et al., 2020) evaluates zero-shot grammatical knowledge through minimal pairs, where a model must assign higher probability to the grammatical sentence over an ungrammatical one. The BASELINE OFFICIAL achieved the highest BLiMP score of 70.31, outperforming the BASELINE SYNTHETIC (66.71), the FIXED CURRICULUM (62.62), the ADAPTIVE CURRICULUM (VARIABLE LR) (59.28), and the ADAPTIVE CURRICULUM (57.77). For the BASELINE OFFICIAL,

FIXED CURRICULUM, and ADAPTIVE CURRICULUM (VARIABLE LR), this pattern holds for morphology (83.75 vs. 72.73 vs. 64.04) and syntax (65.29 vs. 57.01 vs. 53.03). However, the ADAPTIVE CURRICULUM (VARIABLE LR) achieved a higher semantics score (64.94) than the BASELINE OFFICIAL (62.60) and the FIXED CURRICULUM (59.00), despite its lower overall BLiMP score. The curriculum models therefore did not outperform either baseline on BLiMP under these conditions, and both adaptive variants scored below the FIXED CURRICULUM on BLiMP. We conjecture that training on teacher-generated synthetic text, which may differ in vocabulary and sentence structure from the natural BabyLM corpus, may have reduced the model’s exposure to the statistical patterns that BLiMP probes. BLiMP Supplement scores are similar across conditions, with the FIXED CURRICULUM narrowly leading (56.74), followed by the ADAPTIVE CURRICULUM (VARIABLE LR) (54.41), the BASELINE OFFICIAL (54.35), the BASELINE SYNTHETIC (53.28), and the ADAPTIVE CURRICULUM (52.93).

World knowledge (EWoK) and property knowledge (COMPS). EWoK (Ivanova et al., 2024) tests whether models have acquired real-world knowledge about physical and social domains through contextually appropriate minimal pairs. COMPS (Misra et al., 2023) evaluates property knowledge and whether models can infer that properties of superordinate concepts are inherited by subordinate concepts. Performance on EWoK and COMPS was similar across the five conditions, with scores differing by less than three points on each benchmark. The BASELINE SYNTHETIC achieved the highest scores on both EWoK (51.66) and COMPS (52.59), although its margins over the other conditions were small.

Entity tracking. Entity Tracking (Kim and Schuster, 2023) tests a model’s ability to follow referents across multiple operations and clause boundaries. The largest variation across the five conditions occurred on Entity Tracking, where the ADAPTIVE CURRICULUM (VARIABLE LR) achieved the highest score of 36.07, followed by the FIXED CURRICULUM (28.50), BASELINE OFFICIAL (26.70), ADAPTIVE CURRICULUM (17.66), and BASELINE SYNTHETIC (17.32). The best-performing variant also exceeded the challenge organizers’ GPT-2 Strict-Small reference score of 23.58. Although the fixed curriculum outperforms

Model	BLiMP	BLiMP Supp.	EWoK	COMPS	Entity Tracking	S.GLUE	Macroavg
BASLINE OFFICIAL	70.31	54.35	51.09	51.57	26.70	63.71	59.86
BASLINE SYNTHETIC	66.71	53.28	51.66	52.59	17.32	62.52	58.54
FIXED CURRICULUM	62.62	56.74	51.16	51.75	28.50	61.82	58.08
ADAPTIVE CURRICULUM	57.77	52.93	50.06	50.93	17.66	61.44	55.55
ADAPT. CURR. (VAR. LR)	59.28	54.41	49.53	49.83	36.07	62.54	56.44

Table 3: Evaluation results. Bold indicates best performance per column across our five models.

Model	BoolQ	MultiRC	RTE	WSC	MRPC	QQP	MNLI
BASLINE OFFICIAL	69.48	62.79	57.55	65.38	71.57	72.55	46.62
BASLINE SYNTHETIC	69.05	62.91	53.24	65.38	71.57	69.13	46.37
FIXED CURRICULUM	68.56	63.78	51.80	61.54	70.59	71.40	45.05
ADAPTIVE CURRICULUM	69.11	61.10	53.24	61.54	70.10	71.12	43.91
ADAPTIVE CURRICULUM (VARIABLE LR)	69.54	62.71	54.68	63.46	70.59	72.06	44.72

Table 4: Language understanding (SuperGLUE) sub-task results (accuracy). Bold indicates best performance per column across our five models.

both baselines, the two adaptive variants showed markedly different results. This contrast suggests that the effectiveness of adaptive training on this task may depend on the learning-rate schedule, though an isolated experiment with repeated measures would be required to confirm. The observed gains in the ADAPTIVE CURRICULUM (VARIABLE LR) model’s Entity Tracking cannot be attributed to the corrective feedback loop alone, and further controlled comparisons would be needed to establish the direct impact of developmental ordering and corrective feedback.

Language Understanding (SuperGLUE). The SuperGLUE benchmark represents a composite of difficult language understanding tasks that each model is finetuned and then evaluated on (Wang et al., 2020). The breakdown of each model’s performance on these tasks is shown in Table 4. The ADAPTIVE CURRICULUM (VARIABLE LR) achieved the highest BoolQ accuracy (69.54) and scored higher than the ADAPTIVE CURRICULUM on all seven tasks. It also performed competitively on QQP (72.06), although the BASELINE OFFICIAL led with 72.55. The BASELINE OFFICIAL also achieved the highest scores on RTE (57.55) and MNLI (46.62), and shared the highest WSC (65.38) and MRPC (71.57) scores with the BASELINE SYNTHETIC. The FIXED CURRICULUM achieved the highest MultiRC score (63.78). Overall, we found that no single condition led across all tasks, but the BASELINE OFFICIAL achieved the highest aggregate SuperGLUE score (63.71).

5 Discussion

The results reveal different strengths across the five conditions. The BASELINE OFFICIAL achieves the highest BLiMP and aggregate SuperGLUE scores, while the ADAPTIVE CURRICULUM (VARIABLE LR) achieves the best entity tracking despite lower BLiMP accuracy. The BASELINE SYNTHETIC performs best on EWoK and COMPS, whereas the FIXED CURRICULUM leads on BLiMP Supplement and MultiRC. A possible contributor to the lower BLiMP scores is the shift from natural to synthetic training text. Teacher-generated text may reflect the vocabulary and syntactic preferences of meta-llama/Llama-3.2-1B-Instruct (Grattafiori et al., 2024) and Qwen/Qwen2.5-1.5B-Instruct (Yang et al., 2024). However, the synthetic baseline also outperforms all three curriculum conditions on BLiMP, suggesting that the use of synthetic text alone does not explain the full performance gap.

The ADAPTIVE CURRICULUM (VARIABLE LR) achieves an entity tracking score of 36.07, exceeding the challenge organizers’ GPT-2 Strict-Small reference score (23.58) and approaching the corresponding base teacher models evaluated on the same benchmark (Llama-3.2-1B: 38.55, Qwen2.5-1.5B: 39.87), despite training on much less data. However, the standard ADAPTIVE CURRICULUM scores much lower on this task, suggesting that the gain may also depend on the learning rate schedule. The SuperGLUE results show a more limited advantage that the variable learning rate variant leads on BoolQ (69.54), narrowly above the BASELINE OFFICIAL (69.48), while the

latter outperforms both adaptive variants on MNLI and MRPC and retains the highest aggregate score. Together, these results suggest that the benefits of adaptive training may depend on the tasks and learning rate scheduling, with an advantage appearing on entity tracking in one variant.

On the overall macroaverage, the curriculum learning models failed to surpass either baseline model. As we discuss in Section 6, the curriculum used in this work was not strongly grounded in linguistic or child development literature, which may have led to this shortcoming. Additionally, the ADAPTIVE CURRICULUM models performed worse than the FIXED CURRICULUM overall, suggesting the feedback phase wasn’t as beneficial as we had hoped. The post-hoc ratings of the student-generated sentences reported in Section 3.4 showed that the teacher models frequently rated the sentences as “incoherent” (over 95% for Llama). A few cherry-picked examples include “I can’t believe how much I’m looking forward to going on an exciting adventure!” and “Well, you talked about it yesterday and now he’s excited.”, which are grammatically correct except for a missing comma in the second sentence. This suggests that the teacher models may not have been able to provide quality corrections during the feedback phase, limiting its effectiveness.

6 Conclusions

We investigated five pretraining strategies for a 58M-parameter student under a 10M word training budget: an official baseline, a synthetic baseline, a fixed developmental curriculum, an adaptive feedback curriculum, and the adaptive curriculum with a variable learning rate. In the adaptive curriculum, teacher models generate, corrective training data in response to the student’s own productions. The baseline retained the strongest grammatical knowledge as measured by BLiMP, while the adaptive curriculum with the variable LR achieved substantially stronger entity tracking, nearly matching the teacher models on that benchmark despite being trained on far less data. These findings suggest that with proper scheduling correction-based feedback during pretraining may strengthen reasoning and referential tracking, even when it does not improve overall grammatical accuracy, and may be most valuable in settings where downstream adaptability matters more than zero-shot grammatical judgment.

While our curriculum models’ results showed promise in some specific task areas, the overall evaluations failed to match the curriculum-free baseline models. This shows that our approach was not as effective as we hypothesized. To improve our method in future work, we could isolate the effect of the curriculum from the overall dataset by using large instruction models to order the curated official dataset by increasing complexity. We could also leverage the use of two teacher models by having them take distinct, specialized roles; student productions could be routed to the teacher best suited to correcting a specific error type. Finally, we could explore a wider range of adaptive phase hyperparameters to better balance grammaticality and reasoning performance, or alternate between the foundational and feedback phases within each developmental stage.

Our five models and their intermediate checkpoints are publicly available on the Hugging Face Hub: the OFFICIAL BASELINE,¹ the FIXED CURRICULUM,² the ADAPTIVE CURRICULUM VARIABLE LR,³ the ADAPTIVE CURRICULUM with a learning rate of 3×10^{-4} ,⁴ and the SYNTHETIC BASELINE.⁵

Limitations

This work had a few limitations that should be noted. The hyperparameters used in training were chosen based on prior experience and observations made during the experiments, but they were not rigorously tuned. This includes the learning rate, number of epochs, and batch size, as well as the number of feedback iterations and the split between the foundational and adaptive phases in FEEDBACK CURRICULUM. It is possible that a more thorough grid search over hyperparameters may have resulted in improved model performance.

Additionally, the ADAPTIVE CURRICULUM model used the same teacher models to generate the adaptive feedback as well as to generate the foundational data. This may have resulted in the synthetic data from both phases reflecting teacher-

¹<https://huggingface.co/ksu-help/babylm2026-strict-small-baseline-ep5>

²<https://huggingface.co/ksu-help/babylm2026-strict-small-fixed-ep5>

³<https://huggingface.co/ksu-help/babylm2026-strict-small-adaptive-synth-ep5>

⁴<https://huggingface.co/ksu-help/babylm2026-strict-small-adaptive-lr3e-4-ep5>

⁵<https://huggingface.co/ksu-help/babylm2026-strict-small-synthetic-baseline-ep5>

specific shortcomings or stylistic biases. It may have been beneficial to use a separate teacher model for each purpose, to allow one to correct any mistakes learned from the other.

Finally, we defined our three developmental stages through roughly equal age divisions and progressions of grammar rules, without strong grounding in prior linguistic or child development literature. The choice of three discrete stages was practical rather than theoretically motivated; a continuous or finer-grained complexity ordering might better reflect how children actually acquire language. Additionally, even young children are exposed to adult-to-adult speech that is considerably more complex than our infant-stage prompts, so removing complex constructions entirely may not accurately reflect naturalistic input. Future work could ground stage boundaries in established language acquisition milestones (Gilkerson et al., 2017), gradually increase the proportion of complex utterances rather than using hard stage transitions, or compare discrete versus continuous complexity schedules directly.

Acknowledgements

The authors generated the ideas, workflow, experimental design, and claims of this work. AI tools were used as follows: OpenAI and Anthropic assisted with code debugging and grammar correction in the manuscript. Grammarly was also used to increase the readability of the paper. All suggestions were reviewed, verified, and edited if necessary by the authors before inclusion and no confidential or personally identifiable data were used.

References

- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48.
- 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*, Suzhou, China. Association for Computational Linguistics.
- 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: Call for papers for the 2026 babyLM workshop](#). *Preprint*, arXiv:2602.20092.
- Michelle M. Chouinard and Eve V. Clark. 2003. Adult reformulations of child errors as negative evidence. *Journal of Child Language*, 30(3):637–669.
- Eve V. Clark. 2018. Conversation and language acquisition: A pragmatic approach. *Language Learning and Development*, 14(3):170–185.
- Richard Diehl Martinez, Zébulon Goriely, Hope McGovern, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2023. CLIMB: Curriculum learning for infant-inspired model building. *arXiv preprint arXiv:2311.08886*.
- Jill Gilkerson, Jeffrey A. Richards, Steven F. Warren, Judith K. Montgomery, Charles R. Greenwood, D. Kimbrough Oller, John H. L. Hansen, and Terrance D. Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, 26(2):248–265.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. [Distilling the knowledge in a neural network](#). *arXiv preprint arXiv:1503.02531*.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. [Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora](#). In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.
- Anna Ivanova, Carina Kauf, Greta Tuckute, Nancy Kanwisher, Joshua B Tenenbaum, Evelina Fedorenko,

- and Jacob Andreas. 2024. EWoK: Elements of world knowledge for evaluation of language model world modeling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.
- Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 3835–3855.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Evan Lucas, Dylan Gaines, Tagore Rao Kosireddy, Kevin Li, and Timothy C. Havens. 2024. [Using curriculum masking based on child language development to train a large language model with limited training data](#). In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 221–228, Miami, FL, USA. Association for Computational Linguistics.
- Jonas Mayer Martins, Ali Hamza Bashir, Muhammad Rehan Khalid, and Lisa Beinborn. 2025. [Once upon a time: Interactive learning for storytelling with small language models](#). In *Proceedings of the First BabyLM Workshop*, Suzhou, China. Association for Computational Linguistics.
- Kanishka Misra, Julia Taylor 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 2920–2938.
- Suchir Salhan, Richard Diehl Martinez, Zébulon Goriely, and Paula Buttery. 2024. [Less is more: Pre-training cross-lingual small-scale language models with cognitively-plausible curriculum learning strategies](#). In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, Miami, FL, USA. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Jean-Loup Tastet and Inar Timiryasov. 2024. [BabyLlama-2: Ensemble-distilled models consistently outperform teachers with limited data](#). In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, Miami, FL, USA. Association for Computational Linguistics.
- 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*, pages 279–289, Singapore. Association for Computational Linguistics.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. [Superglue: A stickier benchmark for general-purpose language understanding systems](#). *Preprint*, arXiv:1905.00537.
- 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 Mo-hanbey, 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.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pi-er-ic Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Trans-formers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. [Qwen2.5 technical report](#). *arXiv preprint arXiv:2412.15115*.

A Fixed Curriculum Prompts

Infant.

Generate language for very young children (under about 4 years old), in the style of simple child-directed speech.

Generate several grammatically simple sentences a young child would naturally say or hear.

Style of this level:

- Single-clause sentences, usually subject-verb-object.
- Present tense and basic everyday verbs (eat, run, see, play, sleep, want, have, look, go, like).
- Concrete, familiar nouns. No abstract ideas, no subordinate or relative

clauses, no passive voice.

- Sentences may be short or slightly longer (simple "and" coordination is fine). What defines this level is grammatical simplicity, NOT the number of words.

Examples of the target style (note the varied lengths):

- I see you.
- The cat sees the bird.
- Mama gives me milk.
- The dog runs fast.
- I want my ball and my cup.

Use the source passage below only as inspiration for topic and words. Do not copy it, translate it, or try to preserve its meaning.

Output only the new sentences, one per line, no numbering and no bullets.

Source passage:

Toddler.

Generate language for children aged about 4 to 7 years old.

Generate several sentences at this developmental level.

Style of this level:

- Mostly one or two clauses, joined with simple connectives (and, but, because, when, so).
- Descriptive adjectives, prepositional phrases, and a moderate everyday vocabulary.
- Concrete situations a young child understands. Avoid technical or abstract vocabulary.
- Sentences may be short or longer. A short sentence can still show this level (e.g. a simple "because" clause), and a longer one can stay simple. Grammatical style, not length, defines the level.

Examples of the target style (note the varied lengths):

- She cried because she fell.
- The big yellow bus stopped at the corner.
- My little brother is sad because his balloon flew into the tree.
- When it rained, the children ran inside to stay dry.
- He wanted to play outside, but he had to finish his lunch first.

Use the source passage below only as inspiration for topic and words. Do not copy it, translate it, or try to preserve its meaning.

Output only the new sentences, one per line, no numbering and no bullets.

Source passage:

Child.

Generate richer language for children aged about 8 to 12 years old.

Generate several sentences at this developmental level.

Style of this level:

- Subordinate clauses, relative clauses, and varied tenses.
- Discourse connectives (however, therefore, although, despite, while, whereas).
- Abstract ideas and reasoning are allowed.
- Sentences may be short or long. A short sentence can still be syntactically complex (e.g. "He left before she could answer"), and complexity, not length, is what defines this level.

Examples of the target style (note the varied lengths):

- He left before she could answer.
- Despite the long flight, the team arrived in good spirits.
- The library, which had been closed for renovations, reopened with new study spaces.
- Although the experiment had failed twice, the students believed small changes would finally produce clearer results.
- She wondered whether the rumor about the old house was actually true.

Use the source passage below only as inspiration for topic and words. Do not copy it, translate it, or try to preserve its meaning.

Output only the new sentences, one per line, no numbering and no bullets.

Source passage:

B Adaptive Curriculum Prompts

Infant.

A small language model in training produced the sentence below while learning to talk. It may contain grammar mistakes.

Generate 4 to 5 correct sentences that express the same idea properly, in the style of speech to a child under about 4 years old: one clause, present tense, concrete everyday words, no subordinate or relative clauses.

If the sentence is already correct, generate 4 to 5 simple sentences on the same idea at this level.

If it cannot be understood, generate 4 to 5 simple well-formed sentences at this level on a similar everyday topic.

Write each sentence on its own line.

Toddler.

A small language model in training produced the sentence below while learning to talk. It may contain grammar mistakes.

Generate 4 to 5 correct sentences that express the same idea properly, at the level of a child aged about 4 to 7: one or two clauses joined by simple connectives (and, but, because, when, so), everyday vocabulary, concrete situations.

If the sentence is already correct, generate 4 to 5 sentences on the same idea at this level.

If it cannot be understood, generate 4 to 5 well-formed sentences at this level on a similar everyday topic.

Write each sentence on its own line.

Child.

A small language model in training produced the sentence below while learning to write. It may contain grammar mistakes.

Generate 4 to 5 correct sentences that express the same idea properly, at the level of a child aged about 8 to 12: they may use subordinate or relative clauses, varied tenses, and discourse connectives (however, although, because, while), and may express abstract ideas.

If the sentence is already correct, generate 4 to 5 sentences on the same idea at this level.

If it cannot be understood, generate 4 to 5 well-formed sentences at this level on a similar topic.

Write each sentence on its own line.

C Student Sentence Rating Prompt

Decide which of the three categories fits the sentence best: Grammatically correct, Incoherent, or Coherent but with errors.

Examples: "the the baby was very happy." -> Coherent but with errors "They got us a new game." -> Grammatically correct "the when park going that is" -> Incoherent "what are you waiting for?" -> Grammatically correct "it was a on the day." -> Incoherent "He go to school." -> Coherent but with errors "The Great War was parked outside." -> Incoherent "And I'd have trouble remembering what I had planned." -> Grammatically correct "where are the book." -> Coherent but with errors "she gave you the name." -> Grammatically correct "was I alive is my favorite food." -> Incoherent "Well,how fast were with you?" -> Coherent but with errors

Answer with only one of these three: Grammatically correct, Incoherent, or Coherent but with errors. No explanation needed

Sentence: sentence