

BabySteps at BabyLM 2026: Stronger Optimization and Causal Baselines for Data-Constrained Pretraining

Siddhi Satheesh* Jorge Murillo* Yubei Chen

University of California, Davis

{svsatheesh, joemurillo, ybchen}@ucdavis.edu

Abstract

We present three BabySteps systems that strengthen baselines for data-constrained pretraining in the strict-small (10M words) and strict (100M words) NLP tracks of the 2026 BabyLM Challenge. Across our GPT-BERT experiments, optimizer choice and learning-rate tuning produced our largest gains, highlighting the importance of well-tuned baselines for evaluating progress in sample-efficient pretraining. The curriculum, synthetic-data, teacher-supervision, and corpus-mixing interventions that we tested did not improve the final aggregate score.

We also test whether the official GPT-2 checkpoint is a sufficiently strong reference for causal language modeling. A 127.6M-parameter Llama-style decoder trained with Muon and Liger-optimized kernels obtains a seven-suite NLP average of 56.18, compared with 52.79 for our completed reevaluation of the official BabyLM GPT-2 strict checkpoint, and scores higher on all seven categories. This is not a controlled architectural ablation: architecture, optimizer, implementation, and hyperparameters change together. Instead, it provides a stronger practical causal baseline. In addition, we report failed data interventions, a numerical failure caused by batches with no supervised target tokens, and substantial variance between small evaluation suites.

1 Introduction

The BabyLM Challenge (Warstadt et al., 2023; BabyLM Organizing Committee, 2026) asks whether language models can be made dramatically more sample-efficient by restricting pretraining to developmentally plausible budgets of 10M (strict-small) or 100M (strict) words. Plausible approaches to sample-efficient pretraining include: curate a better corpus, order it like a curriculum, generate synthetic text targeted at the evaluations,

or distill knowledge from a teacher. We pursued those intuitions, and this paper is in part a report on how they failed for us.

Our results point to a broader lesson about baseline quality. Within GPT-BERT (Charpentier and Samuel, 2024), we obtained larger and more reliable gains from Muon-family optimization and learning-rate tuning (Jordan et al., 2024; Si et al., 2025) than from any of the data-centric interventions we tested. Separately, we found that the official GPT-2 baseline was not a strong representation of what a modern causal decoder could achieve. We therefore trained a Llama-style causal model (Touvron et al., 2023) using Muon and a Liger-optimized implementation (Hsu et al., 2024), primarily to provide a stronger practical control rather than to propose an architectural innovation.

We make three contributions:

- Across our experiments, Muon-family optimization and learning-rate tuning produced our strongest results; none of the tested data interventions improved the final scores.
- We introduce a modern Llama-style causal baseline that improves the completed official GPT-2 strict result from 52.79 to 56.18 NLP average and outperforms it on every individual NLP benchmark.
- We document evaluation and reproducibility pitfalls that can materially distort conclusions at BabyLM scale.

We submit **BabySteps MurphysLaw-10M-mixed**, a strict-small GPT-BERT trained with AdaMuon; **BabySteps SidBert-100M-mixed**, a strict GPT-BERT trained with Muon; and **BabySteps Chatty BERT**, a strict Llama-style causal decoder trained with Muon using memory-efficient fused kernels.

* Equal contribution.

2 Method

2.1 Model families

2.1.1 GPT-BERT reconstruction

Two of our three submissions use GPT-BERT (Charpentier and Samuel, 2024), which trains a single transformer under a mixture of masked-language-modeling and causal objectives and can subsequently be evaluated in either causal or masked-next-token-prediction (MNTP) modes. We reimplemented the training loop—on both single and multi-GPU runs—using the published baseline configurations (the “reconstruction” in our internal naming), matched the BabyLM baseline model sizes for each track, and verified the implementation by reproducing baseline-level zero-shot scores before changing the optimization recipe.

Both GPT-BERT submissions use the *mixed* objective family and are evaluated in MNTP mode. In our sweeps, this combination provided the strongest overall balance across the 2026 evaluation suites, although causal-focus configurations remained competitive on several individual benchmarks.

2.1.2 Modern causal baseline: Llama with Liger

The third system, *BabySteps Chatty BERT*, is a decoder-only Llama-style model (Touvron et al., 2023). Despite its leaderboard name, it is not a BERT-family model, and we added a disguise emoji to point to that. We trained it from scratch on the 100M-word BabyLM 2026 Strict corpus to test whether the official GPT-2 baseline accurately represents the performance attainable by a modern causal decoder under the same data restriction. Our goal with Chatty BERT was to provide a stronger practical baseline rather than to propose a novel architecture.

The model contains 127.6M parameters, 12 transformer layers, a hidden dimension of 896, an intermediate dimension of 2,304, 14 attention heads, and a maximum context length of 1,024 tokens. It uses a 16,384-entry byte-level BPE vocabulary with tied input and output embeddings.

We used the Hugging Face Llama patch provided by Liger Kernel (Hsu et al., 2024), enabling its optimized implementations of rotary positional embeddings, SwiGLU, and RMSNorm. We did not enable Liger’s cross-entropy or fused linear cross-entropy kernels. This implementation made the Llama–Muon configuration a practical 100M-word

baseline on a single GPU by reducing the memory and execution overhead associated with these operations.

The single largest and most reliable improvement over the AdamW-trained baselines came from the optimizer. We use Muon (Jordan et al., 2024), which orthogonalizes momentum updates for hidden weight matrices, and AdaMuon (Si et al., 2025), an adaptive variant, with a separate auxiliary learning rate for the non-matrix parameters (embeddings, norms, head). Track-level winners differed: at 10M words, AdaMuon with learning rate 0.02 and auxiliary rate 0.006 was best; at 100M words, Muon with learning rate 0.005 and auxiliary rate 0.007 was best, and the 100M scale showed a sharp cliff at higher rates (a 3×10^{-5} finetuning learning rate was already enough to partially degenerate some GLUE tasks on a sibling run). All final models use seed 42 and identical data ordering to their development runs.

2.2 Tail-weight averaging

For the strict-small entry we submit the average of the weights over the final 20% of training steps rather than the last checkpoint. On the twin development run this left zero-shot scores essentially unchanged (± 0.3 BLiMP) but improved finetuned SuperGLUE by +1.6 points (70.1 vs. 68.5 average), and the improvement was stable across finetuning learning rates and seeds. The 100M entry uses final weights, as tail averaging did not help at that scale in our sweeps.

2.3 Training configuration and checkpoints

MurphysLaw trains for 9,914 steps at 16,384 tokens per step (≈ 10 epochs of the official strict-small corpus); SidBert trains for 12,330 steps at 131,072 tokens per step on the official strict corpus. The Llama submission trains for 13,510 steps at 131,072 tokens per step (≈ 10 epochs of the official strict corpus). The models save the word-count milestone checkpoints required by the 2026 challenge: 19 for strict-small and 28 for strict. We publish these as revisions of the submission repositories, with the final weights on `main`, giving 20 and 29 published model states, respectively. We evaluate them with the official fast pipeline, and the same milestone checkpoints drive the Age-of-Acquisition analysis (Chang and Bergen, 2022; Frank et al., 2017). Finetuning for SuperGLUE follows the official pipeline; the best strict-small recipe was learning rate $8\text{--}9 \times 10^{-5}$ with seed 44,

while SidBert used $2-3 \times 10^{-5}$ with seed 42 and Llama used 3×10^{-5} with seed 42.

3 What did not work

Because the negative results consumed most of the project, we summarize them here; Table 2 gives representative numbers. All were evaluated on BLiMP (Warstadt et al., 2020) and BLiMP-Supplement under the official pipelines of their respective years.

Alternative and curriculum corpora. TinyStories (Eldan and Li, 2023) substantially underperformed the standard corpus at 10M words (47–51 BLiMP vs. ≈ 70). A hand-built Gutenberg corpus ordered by U.S. reading level (grades 1–4 $\rightarrow$ 5–8 $\rightarrow$ 9–adult), motivated by curriculum learning, performed no better, and grammar-textbook corpora were worse still. Reordering the corpus or looping epochs within grade bands did not change the result. This is consistent with previous BabyLM findings, in which curriculum learning was widely explored but generally did not improve aggregate performance over non-curriculum baselines (Oba et al., 2023; Hu et al., 2024; Charpentier et al., 2025).

Synthetic data generation. Generating corpus text targeted toward evaluation-like linguistic phenomena, including minimal-pair-style sentences and prompt variations, reached at best 55.7 BLiMP, well below training on the standard corpus. A smaller 1% synthetic augmentation produced a more localized benefit: it improved BLiMP and BLiMP-Supplement to 72.5 and 70.3, respectively, but reduced entity tracking from 28.0 to 22.1 and SuperGLUE from 69.9 to 64.6, lowering the overall NLP average from 53.6 to 52.1. This result should not be interpreted as evidence that synthetic data is uniformly ineffective. Prior work indicates that its value depends strongly on the generation procedure, linguistic diversity, and proportion of natural text (Charpentier et al., 2025; Kang et al., 2025).

In our experiments, targeted synthetic augmentation improved the linguistic phenomena it most closely resembled, but these gains did not generalize across the broader evaluation suite.

Teacher-supervised objectives. A distillation-adjacent scheme in which a teacher reweights the student’s loss toward syntactically informative tokens plateaued at 66–67 BLiMP, a few points below

the plain GPT-BERT reconstruction, and hybridizing it with nanochat-style (Karpathy, 2025) training schedules did not close the gap (best 65.8/60.0).

Corpus mixing trade-offs. Dialogue-rich custom mixtures (our “Benchmark Target Mix” family) did improve zero-shot suites (+1.7 selective zero-shot average) but cost ≈ 3.5 points of finetuned SuperGLUE relative to the official corpus at 10M—a trade that loses under the challenge’s aggregate. This was the closest any data intervention came to helping, and it still lost to the official corpus.

4 Results

4.1 Strict-small (10M): candidate selection

Table 3 compares the checkpoint-eligible finalists from our development sweeps under MNTP evaluation. The submitted model wins on the challenge-relevant aggregate despite not winning every column: the Benchmark Target Mix-family models buy zero-shot supplement and EWok (Ivanova et al., 2024) points at the cost of SuperGLUE (Wang et al., 2019), while the official-corpus AdaMuon model has the best Global-PIQA and finetuning behavior. Its Age-of-Acquisition correlation with child CDI data is -0.151 ($p = 0.039$) over the 19 checkpoints.

4.2 GPT-BERT Strict (100M)

Table 4 reports the 100M sweep for GPT-BERT Mixed Focus. The submitted configuration (Muon, lr 0.005) dominates its siblings on every aggregate: +2 finetuned SuperGLUE over the next-best run and the best zero-shot scores of the entire campaign (80.2 BLiMP, 73.6 Supplement). The finetuning learning-rate cliff observed on the sibling run (partial task collapse at 3×10^{-5}) did not appear for this model, whose 2×10^{-5} and 3×10^{-5} recipes tie at 73.4–73.5. Table 4 also shows variant SidBert models: identically-configured, identically-seeded retrains with milestone checkpointing enabled; and Section 5 quantifies the reproduction fidelity we expect from such retrains.

4.3 A stronger causal baseline

The official leaderboard row for `gpt2-baseline-strict` was incomplete, with several missing evaluations represented as zero. We therefore reevaluated the official checkpoint using the BabyLM 2026 strict full-evaluation data, the causal backend, batch size 64, and the main model revision. Table 5 compares this

	MurphysLaw	SidBert	Llama
Track	strict-small	strict	strict
Words	10M (official)	100M (official)	100M (official)
Objective	mixed	mixed	causal
Optimizer	AdaMuon	Muon + AdamW	Muon + AdamW
LR / aux LR	0.02 / 0.006	0.005 / 0.007	0.01 / 1e-4
Steps $\times$ tok/step	$9,914 \times 16,384$	$12,330 \times 131,072$	$13,510 \times 131,072$
Final weights	tail-avg (20%)	last step	last step
Milestones	19 (1M–100M)	28 (1M–1000M)	28 (1M–1000M)
Eval mode	MNTP	MNTP	causal
FT recipe	8–9e-5, s44	2–3e-5, s42	3e-5, s42

Table 1: Submission configurations. Milestone counts exclude the final model state published on `main`.

Intervention (10M)	BLiMP	Supp.
TinyStories corpus (20 ep)	47.3	45.3
Gutenberg curriculum corpus	≈ 66	≈ 55
Synthetic generation (best)	55.7	51.4
Teacher-supervised (best)	67.3	53.0
nanochat schedule (best)	57.0	41.0
Hybrid teacher+nanochat	65.8	60.0
GPT-BERT recon. + AdaMuon	71.6	63.9

Table 2: Representative development results for data- and objective-centric interventions versus the submitted optimizer-centric recipe (official corpus). Earlier rows use the evaluation pipeline of their year; the final row is the submitted strict-small model under the 2026 pipeline.

completed evaluation with the official Chatty BERT submission.

Chatty BERT improves the seven-suite NLP average by 3.39 points and outperforms GPT-2 on every benchmark category. The largest gains occur on BLiMP, BLiMP-Supplement, and SuperGLUE, suggesting that the modern training recipe particularly improves linguistic competence and downstream transfer. Improvements on EWoK and entity tracking are smaller, and the absolute entity-tracking scores remain low for both models. Thus, strengthening the causal baseline does not remove the broader difficulty of learning persistent entities and world structure from limited text.

The comparison should not be interpreted as isolating a uniquely Llama-specific architectural advantage. Chatty BERT jointly changes the model architecture, normalization and feed-forward design, positional encoding, optimizer, implementation stack, and tuned hyperparameters. Its role is instead to demonstrate that the official GPT-2 result is not a sufficient control for claims about the limits of causal modeling.

5 Evaluation pitfalls

Validation loss does not rank configurations well.

We selected hyperparameters by benchmark score, not held-out MLM loss: across a 10M learning-rate grid, validation loss correlates with BLiMP only through broken extremes (Spearman -0.91 overall, -0.40 among the four best), preferring lr 0.005 over the lr 0.02 recipe we submitted, and is uninformative among identically-configured replicas.

Identical retrains are not bit-identical. Our submission process required retraining models with checkpointing enabled, holding configuration, data order, and seed fixed. The retrain reproduced BLiMP to within 0.1 points, but the 82-item BLiMP-Supplement moved by 4.5 points and entity tracking (Kim and Schuster, 2023) by 1.9—consistent with GPU nondeterminism amplified by small suite sizes. Repeating this across 18 identical runs, the seven-suite NLP average spans 56.6–57.9, while evaluation itself is deterministic given fixed weights. Small-suite deltas under a few points should not be read as real effects at these scales, in our submission or anyone else’s.

Training and evaluation code, exact configuration files, and launch scripts for the submitted models are available in our public repositories. We note that this repository is designed to run from PyTorch, while the original GPT-BERT code is designed to run on Slurm, so this is another contribution that may be helpful for future competition participants.

¹ Because the repository also contains exploratory experiments conducted before and after model selection, we provide a paper-specific reproduction index that maps each reported model to its corresponding training configuration, evaluation recipe, and released weights.

¹<https://github.com/JorgeCMurillo/babylm-baby-steps>

GPT-BERT Mixed Focus Model (10M, MNTP)	BLiMP	Supp.	EWoK	Entity	PIQA	SG	Avg.
MurphysLaw (AdaMuon, official)	71.6	63.9	51.9	28.0	36.2	69.9 [†]	53.6
AdaMuon, MixBlend corpus	71.5	70.2	53.5	20.8	33.7	65.4	52.5
Muon, Benchmark Target Mix corpus	70.7	68.5	56.1	19.3	34.3	65.2	52.4
Muon, 1%-augmented corpus	72.5	70.3	51.2	22.1	31.7	64.6	52.1

Table 3: Strict-small finalists (development sweep, MNTP). SG = finetuned SuperGLUE; Avg. is the mean of the six columns, not including COMPS. [†]SuperGLUE for the submitted configuration measured on its identically-configured development twin (69.9–70.1 across the 8–9e-5 / seed-44 recipes); official-pipeline finetuning of the submitted weights uses the same recipe.

GPT-BERT Mixed Focus Model (100M)	BLiMP	Supp.	SG	NLP
Muon lr .005 (sub.)	80.2	73.6	73.5	57.3
Muon lr .005 (var. 1)	79.5	71.8	73.0	57.9
Muon lr .005 (var. 2)	79.4	70.9	71.6	56.0
Muon lr .01	76.6	62.4	70.3	53.8
AdaMuon lr .02	76.6	64.8	70.1	53.9

Table 4: Strict-track sweep (MNTP; SG = finetuned SuperGLUE average; NLP = six-suite mean). The AdaMuon recipe that wins at 10M loses at 100M, where lower-rate Muon dominates. Additionally, we show identical-training submission variants, displaying the GPU-based nondeterminism inherently present in training and evaluating at this scale (var 1. = identical training process, var. 2 = two-GPU distributed training).

6 Discussion and Limitations

6.1 The importance of strong baselines

Our results support a broader but still qualified conclusion: BabyLM comparisons are highly sensitive to the strength of the baseline. Within GPT-BERT, optimizer choice and learning-rate tuning produced larger and more consistent gains than the multiple data interventions we tested. Across causal decoders, replacing the official GPT-2 reference with a modern Llama–Muon implementation raised the strict NLP average by 3.39 points and improved every individual benchmark.

These findings do not imply that data design or architectural innovation is unimportant. Rather, they show that gains attributed to novelty can be overstated when the comparison is made only against an outdated or undertuned reference implementation. A strong BabyLM control should ideally include both the official baseline and a tuned modern baseline trained under the same data and token restrictions.

6.2 Systems efficiency versus sample efficiency

It is also useful to distinguish statistical sample efficiency from systems efficiency. Liger kernels do not change the number of words observed by the model, and our experiment does not establish that they improve learning per token. Their benefit is practical: reduced memory pressure and improved throughput make Muon training and broader hy-

perparameter exploration easier under a fixed hardware budget. Because the Llama and GPT-2 systems differ in several respects, a clean architectural ablation would need to hold parameter count, tokenizer, data order, optimizer, token batch, learning-rate schedule, and total training compute constant.

6.3 Limitations

Our exploration of data interventions, though broad, was not exhaustive, and most sweeps used a single seed. The Llama–GPT-2 comparison is a comparison between practical training recipes rather than a controlled architectural ablation: architecture, optimizer, hyperparameters, and implementation differ simultaneously. Small evaluation suites also show substantial run-to-run variation, so differences of only a few points should be interpreted cautiously. We primarily optimized the learning rates. Finally, Age-of-Acquisition correlations are reported as produced by the official pipeline, and we do not attempt to interpret their sign.

Acknowledgments

We thank the BabyLM organizers for providing the evaluation pipeline and baseline models. We thank Zaishuo Xia for computation infra support. We also thank Shawheen Naderi and many other members of Yubei Chen’s self-supervised learning group at UC Davis for their helpful discussions and continued support.

Model	BLiMP	Supp.	EWoK	Entity	COMPS	PIQA	SG	NLP avg.
GPT-2 strict	74.73	65.00	54.37	16.91	55.85	36.62	66.06	52.79
Chatty BERT (Llama)	80.57	69.88	55.39	19.60	58.01	39.62	70.19	56.18
Difference	+5.84	+4.88	+1.02	+2.69	+2.16	+3.00	+4.13	+3.39

Table 5: Completed strict-track comparison between the official BabyLM strict GPT-2 checkpoint and Chatty BERT. NLP average is the unweighted mean of the seven displayed benchmark categories.

References

- BabyLM Organizing Committee. 2026. The BabyLM challenge 2026: Call for papers. <https://babylm.github.io>. Accessed July 2026.
- Tyler A. Chang and Benjamin K. Bergen. 2022. **Word acquisition in neural language models**. *Transactions of the Association for Computational Linguistics*, 10:1–16.
- Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams. 2025. **Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data**. In *Proceedings of the First BabyLM Workshop*, pages 399–420, Suzhou, China. Association for Computational Linguistics.
- Lucas Georges Gabriel Charpentier and David Samuel. 2024. **GPT or BERT: why not both?** In *Proceedings of the BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*. Association for Computational Linguistics.
- Ronen Eldan and Yuanzhi Li. 2023. **TinyStories: How small can language models be and still speak coherent English?** *Preprint*, arXiv:2305.07759.
- Michael C. Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A. Marchman. 2017. **Wordbank: An open repository for developmental vocabulary data**. *Journal of Child Language*, 44(3):677–694.
- Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, Siyu Zhu, Steven Shimizu, Shivam Sahni, Haowen Ning, and Yanning Chen. 2024. **Liger Kernel: Efficient Triton kernels for LLM training**. *arXiv preprint arXiv:2410.10989*.
- 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 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**. *Preprint*, arXiv:2405.09605.
- Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. 2024. **Muon: An optimizer for hidden layers in neural networks**. <https://kellerjordan.github.io/posts/muon/>. Blog post; see also the NanoGPT speedrun records.
- Feiyang Kang, Newsha Ardalani, Michael Kuchnik, Youssef Emad, Mostafa Elhoushi, Shubhabrata Sen-gupta, Shang-Wen Li, Ramya Raghavendra, Ruoxi Jia, and Carole-Jean Wu. 2025. **Demystifying synthetic data in LLM pre-training: A systematic study of scaling laws, benefits, and pitfalls**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10739–10758, Suzhou, China. Association for Computational Linguistics.
- Andrej Karpathy. 2025. **nanochat: The best ChatGPT that \$100 can buy**. <https://github.com/karpathy/nanochat>. GitHub repository.
- Najoung Kim and Sebastian Schuster. 2023. **Entity tracking in language models**. *Preprint*, arXiv:2305.02363.
- Miyu Oba, Akari Haga, Akiyo Fukatsu, and Yohei Oseki. 2023. **BabyLM challenge: Curriculum learning based on sentence complexity approximating language acquisition**. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 290–297, Singapore. Association for Computational Linguistics.
- Chongjie Si, Debing Zhang, and Wei Shen. 2025. **AdaMuon: Adaptive muon optimizer**. *Preprint*, arXiv:2507.11005. ArXiv preprint.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. **LLaMA: Open and efficient foundation language models**. *arXiv preprint arXiv:2302.13971*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. [SuperGLUE: A stickier benchmark for general-purpose language understanding systems](#). In *Advances in Neural Information Processing Systems* 32.

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*. 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.

A Reproduction details

All models were trained on NVIDIA RTX 5090 GPUs. MurphysLaw and Chatty BERT were trained on a single GPU, while SidBert was trained on two GPUs using data parallelism. Submission repositories publish the final weights on `main` and each word-count milestone as a branch named `check_{N}M`, as required for the challenge’s fast checkpoint evaluation and AoA analysis. Finetuning used the official evaluation pipeline’s scripts with per-task defaults, batch size 4 at 100M scale to fit memory alongside concurrent jobs, and the learning rates and seeds given in Table 1.

Objective focus and evaluation mode. The sweeps reveal an interaction between training objective, evaluation mode, and scale rather than a uniformly best objective family. At strict-small scale, mixed-focus models evaluated with MNTP generally provided the strongest balance across the zero-shot suites, although the best causal-focus configuration was competitive. At strict scale, the result was less decisive: causal-focus models evaluated with MNTP performed particularly well on EWoK, entity tracking, and COMPS, whereas mixed-focus models evaluated with MNTP attained the strongest BLiMP and BLiMP-Supplement results and performed well under the broader submission evaluation. We therefore selected mixed-focus training with MNTP evaluation for both submitted models. This choice should be interpreted as a submission-level model-selection decision rather

than evidence that mixed-focus training uniformly dominates causal-focus training.

B Hyperparameter sweeps for 100M track

Model	Family	LR	Mode	Blimp	Blimp SUPP	Ewok	Entity Tracking	Comps	SelfPace R	Eye Tracking
muon-10m-CAUSAL-auxlr0.001		0.001	causal	71.64	61.97	52.43	33.92	53.33	3.92	8.84
muon-10m-CAUSAL-auxlr0.001		0.001	mntp	71.50	64.15	52.99	41.38	53.08	4.40	9.55
muon-10m-CAUSAL-auxlr0.002		0.002	causal	72.20	62.52	51.84	19.57	53.03	3.52	9.06
muon-10m-CAUSAL-auxlr0.002		0.002	mntp	72.01	65.02	53.80	30.80	53.28	3.86	9.72
muon-10m-CAUSAL-auxlr0.003		0.003	causal	71.55	60.93	52.76	24.15	53.38	3.66	9.39
muon-10m-CAUSAL-auxlr0.003		0.003	mntp	70.40	63.07	53.70	30.72	53.28	4.03	10.03
muon-10m-CAUSAL-auxlr0.005		0.005	causal	73.73	59.91	52.39	14.63	52.89	3.53	9.52
muon-10m-CAUSAL-auxlr0.005		0.005	mntp	71.50	62.88	50.79	26.67	52.85	3.67	9.43
muon-10m-CAUSAL-auxlr0.01		0.01	causal	73.51	63.47	52.11	17.34	52.66	4.05	9.82
muon-10m-CAUSAL-auxlr0.01		0.01	mntp	71.57	61.71	51.85	32.77	53.05	4.80	10.26
muon-10m-mixed-auxlr0.001		0.001	causal	69.85	59.51	51.25	27.45	52.58	4.03	9.66
muon-10m-mixed-auxlr0.001		0.001	mntp	72.22	60.74	50.14	40.49	52.99	4.38	10.21
muon-10m-mixed-auxlr0.002		0.002	causal	70.50	56.06	50.93	27.89	53.14	4.38	9.21
muon-10m-mixed-auxlr0.002		0.002	mntp	71.59	59.38	52.08	36.88	53.00	4.43	9.74
muon-10m-mixed-auxlr0.003		0.003	causal	69.82	57.89	51.51	29.10	53.25	3.95	9.74
muon-10m-mixed-auxlr0.003		0.003	mntp	71.72	61.52	52.99	38.29	53.34	4.51	10.49
muon-10m-mixed-auxlr0.005		0.005	causal	69.31	58.59	50.68	29.77	52.78	4.25	9.84
muon-10m-mixed-auxlr0.005		0.005	mntp	72.15	62.75	51.89	43.03	53.77	4.60	10.54
muon-10m-mixed-auxlr0.01		0.01	causal	70.04	60.19	51.46	33.35	52.56	3.68	9.74
muon-10m-mixed-auxlr0.01		0.01	mntp	71.33	62.84	52.86	35.27	52.93	4.01	10.50

Table 6: In the Strict Small 10M track, Hyperparameter sweep across different learning rates, before entity tracking was updated to remove the non-nothing , and the addition of hidden task GlobalPiqa. We retain the original results because they were the measurements available when the learning rates and model configurations were selected.

LR	BLiMP	BLiMP Supp.	EWoK	Entity Tracking	COMPS	SelfPace R	Eye Tracking
Model focus: Causal eval=causal							
0.00025	75.53	62.69	53.21	31.98	55.92	3.52	10.17
0.0025	78.13	70.22	56.75	32.98	58.14	3.39	9.08
0.005	78.74	69.62	56.60	29.38	58.15	3.18	8.69
0.01	79.95	68.75	57.89	34.82	58.65	3.01	8.16
0.02	79.45	67.88	56.45	30.92	58.39	3.36	8.70
Model focus: Mixed eval=causal							
0.00025	73.41	58.84	52.47	40.22	54.91	3.04	10.06
0.0025	77.43	65.98	53.25	28.34	56.68	3.54	9.33
0.005	78.62	67.38	56.24	33.08	56.92	3.12	9.35
0.01	78.52	63.82	55.59	31.88	57.51	3.26	9.27
0.02	78.06	63.95	54.36	27.99	56.83	3.26	8.83
Model focus: Causal eval=mntp							
0.00025	73.25	62.25	53.62	30.36	55.29	3.52	9.86
0.0025	77.67	69.13	56.31	41.57	58.28	3.99	9.81
0.005	78.76	70.45	57.27	36.46	58.44	3.23	9.23
0.01	79.43	70.65	59.09	44.13	59.47	3.47	9.11
0.02	78.77	71.26	57.38	42.39	59.01	3.62	9.58
Model focus: Mixed eval=mntp							
0.00025	75.32	62.93	52.75	37.00	54.73	3.17	10.18
0.0025	79.50	73.26	54.72	40.81	57.52	3.27	9.43
0.005	80.23	73.62	56.25	41.46	58.72	3.26	9.45
0.01	79.90	73.63	57.48	41.00	59.01	3.14	9.63
0.02	79.16	71.33	55.26	44.59	59.40	3.61	10.33

Table 7: GPT-BERT Learning-rate sweep results for the BabyLM 100M track, before the entity tracking update and addition of the Global PIQA task. We retain the original results because they were the measurements available when the learning rates and model configurations were selected.