

Halved CLM Exposure Mitigates Late-Training Degradation in Small Recurrent Language Models

Yash Maheshwari
Independent Researcher
yashmahe2018@gmail.com

Abstract

Small recurrent language models can degrade late in training under standard causal language modeling (CLM). On the BabyLM Strict-Small corpus, vanilla RWKV-7 peaks near 69% BLiMP early, then loses 2.57 to 3.85 pp by step 18k across all three seeds we ran. Halving how often CLM updates the weights mitigates this decline, via two mechanisms: alternating a backward MLM objective every other step (BiRWKV-7), or skipping every other gradient step (vanilla-skip). Vanilla-skip carries no auxiliary objectives yet beats BiRWKV-7 on peak BLiMP in every seed, so the auxiliaries are not necessary for stability. What halving changes is coupled: it reduces the CLM update count and doubles the learning-rate decay in update space. A constant-rate control still degrades at the two higher rates we tested, so a low terminal rate is not necessary for the decline; but at fixed update count, flooring the rate earlier recovers most of it. The update-space rate trajectory therefore also modulates the decline; it stays coupled to update count, and our design does not separate the two. On the official BabyLM 2026 pipeline our submission scores 68.70% BLiMP, above the GPT-2 baseline (65.23%, 98.4M) at about 28% as many parameters.

1 Introduction

The BabyLM challenge asks one central question in computational language acquisition: how much grammatical knowledge can a model extract from the quantity of data a human child encounters? We train on the official BabyLM 2026 Strict-Small corpus (Choshen et al., 2026; BabyLM Community, 2026). After applying our 32K BPE tokenizer, this gives 14.87M tokens drawn from children’s books and other text sources (Warstadt et al., 2023; Choshen et al., 2024). In this tight budget, sample efficiency matters more than raw compute.

Transformer architectures dominate large-scale language modeling but impose $O(T^2)$ attention costs at training time and KV-cache memory proportional to sequence length at inference. State space models (Gu et al., 2022; Gu and Dao, 2023), retention networks (Sun et al., 2023), and RWKV-style architectures (Peng et al., 2023, 2025) share $O(T)$ training complexity while offering fixed-size recurrent state at inference. Both training and evaluation ran on consumer hardware (Apple M-series, MLX framework (Hannun et al., 2023); details in Appendix A.1).

RWKV-7 (Peng et al., 2025) combines data-dependent decay with in-context correction for expressivity close to full attention while keeping recurrent inference fast. We extend it to BiRWKV-7: a variant that trains a forward ($L \rightarrow R$) causal pass and a backward ($R \rightarrow L$) masked reconstruction pass over the same shared block weights. A learnable fusion gate scales how much backward-pass signal enters the combined representation. N-gram embedding features and multi-token prediction (MTP) heads add further training objectives. At inference, only the forward pass is used.

During BiRWKV-7 training, we observed a consistent failure: vanilla RWKV-7 with full CLM exposure degrades by 2.57 to 3.85 pp across all three seeds, while BiRWKV-7 stays stable. We therefore asked: is that stability due to BiRWKV-7’s auxiliary objectives (MLM, MTP, n-grams), or just to the halved CLM update count that comes from alternating objectives? A vanilla-skip model (same RWKV-7 architecture, no auxiliary objectives, CLM updates halved by step-skipping) separates the two. It exceeds BiRWKV-7 on peak BLiMP across all three seeds (+0.83 pp seed 42, +2.58 pp seed 123, +1.81 pp seed 777) and drops only 0.33 to 0.83 pp from its peak.

Key finding. Full CLM training degrades BLiMP late in training for RWKV-7 across all

three seeds, while halving CLM updates, by objective alternation (BiRWKV-7) or step-skipping (vanilla-skip), mitigates it. Since the auxiliary-free vanilla-skip still outperforms BiRWKV-7, the auxiliary objectives are not necessary for the stability. Halving also makes the learning rate decay twice as fast in update space, so update count and that trajectory stay coupled in our design (Limitations).

Contributions. (1) We identify and characterize a consistent late-training BLiMP degradation under full CLM exposure in small recurrent LMs, across three seeds (Section 4). (2) We show that halving updates by either objective alternation (BiRWKV-7) or step-skipping (vanilla-skip) consistently mitigates the degradation across all three seeds; halving is sufficient for the stability, but since an early learning-rate floor at full exposure also mitigates the decline (seed 42), it is not necessary. Two controls probe the schedule. A constant-learning-rate control reproduces the decline with no annealing, and a fixed-update annealing probe shows the update-space learning-rate trajectory also affects the outcome. Reduced CLM update count and that trajectory are coupled in our design, so neither is isolated from the other and neither is established as the sole cause (Section 4). (3) We document that the best-validation-loss and peak-BLiMP checkpoints differ in all three BiRWKV-7 seeds; the BLiMP advantage is resolvable in seeds 42 and 123, while seed 777’s 0.17 pp gap is within the measured ~ 0.3 pp rerun tolerance (Section 4). (4) As a secondary result, n-gram features add +1.77 pp at inference. We also quantize recurrent KV states to int4; because the BLiMP scoring inputs fit in a single recurrence chunk, this reports infrastructure feasibility rather than compression quality (Appendix B.5).

2 Related Work

Prior BabyLM work has explored many strategies for extracting more grammatical knowledge from limited data. On the architecture side, ELC-BERT (Charpentier and Samuel, 2023) and LTG-BERT (Samuel et al., 2023) redesign the encoder, and modular expert paths add capacity along a different axis (Tapaninaho, 2025). On the objective side, GPT-BERT mixes masked and causal prediction (Charpentier and Samuel, 2024), adaptive masking decays the mask rate during training (Edman and Fraser, 2025), and curriculum methods reorder the

data. These approaches generally ask which architecture or objective improves sample efficiency. Recurrent state space models (S4 (Gu et al., 2022), S5 (Smith et al., 2023), Mamba (Gu and Dao, 2023), and RWKV variants (Peng et al., 2023, 2025)) share $O(T)$ training complexity and fixed-size recurrent state at inference. BLaLM (Haller et al., 2025), which uses the mLSTM cell from xLSTM (Beck et al., 2024), is the closest prior linear-time BabyLM entry we are aware of; on the 2025 Strict-Small track it reports 67.0% BLiMP and 33.7% Entity Tracking. Because that result is from a different year’s corpus and pipeline, we do not draw a head-to-head comparison and note it only to place linear-time models in the BabyLM setting. Our BLiMP evaluation protocol follows Warstadt et al. (2020).

Repeated-data overfitting under a fixed corpus is itself studied. Muennighoff et al. (2023) find that repeating data for up to about four epochs matches fresh data in the compute-optimal regime, with returns diminishing rapidly past that. Closer to our small-data scale, Chen et al. (2026) report that a 150M-parameter model on 75M tokens reaches its validation-loss minimum near epoch sixteen and then degrades monotonically, with training-time augmentation delaying the decline. Within BabyLM, Yamamoto and Miura (2025) counter over-training by repeating short learning-rate intervals with optimizer-state resets between them, and Roque and Velasco (2025) substitute simplified text rather than repeat it under a two-pass budget. Early optima and repetition-driven degradation are thus known. Our contribution is narrower: we document a broad late BLiMP degradation in small RWKV-7 and show that halving the CLM update count mitigates it without any auxiliary objectives, though update count stays entangled with the update-space learning-rate schedule (Section 4).

3 Model and Training

Terminology. Throughout, “CLM exposure” means the number of optimizer updates driven by the CLM objective, not total token or batch exposure. BiRWKV-7 matches vanilla in data steps and optimizer-update events, but not necessarily in parameter-space movement: equal update counts do not imply equal movement when the update gradients differ.

Architecture. BiRWKV-7 uses shared RWKV-7 blocks with two training passes.

Forward pass (L→R): produces causal CLM logits, plus lightweight multi-token prediction (MTP) heads for positions $t+2$ and $t+3$. The MTP heads are delta projections that add multi-step prediction signal without touching the primary recurrence.

Backward pass (R→L): the token sequence is masked (15%, BERT-style (Devlin et al., 2019)) and reversed, then fed through the same block weights to produce masked language modeling (MLM) reconstruction logits. The forward and backward hidden states are combined through a learnable sigmoid gate that controls how much backward signal enters the joint representation. This backward pass is an auxiliary training objective only: its gradients flow through shared weights during training, but it is not used at inference. “Bidirectional” here means weight-sharing across two causal passes (L→R and R→L), not full masked bidirectionality where both directions are simultaneously visible at each position. All BLiMP scores in this paper use the forward CLM head only. Entity Tracking results using the fused representation are reported as a secondary finding in Section 4.

N-gram embedding projection: a $(2048 \times n_embd)$ linear projection blends character-level 2/3/4-gram features (MD5 hash, 2048 vocab) into token embeddings through a learnable scalar alpha (initialized at 0.10). Alpha rises to ~ 0.20 early in training and stabilizes near 0.13 across all three seeds.

Configuration. The shared RWKV-7 backbone has 27,369,984 parameters ($\sim 27.4M$), which is also the exact count of the vanilla and vanilla-skip models. BiRWKV-7’s auxiliary components (n-gram projection, MTP heads, fusion gate) add 787,585 parameters for 28,157,569 in total ($\sim 28.2M$); only the forward CLM path runs at inference. The model uses 8 layers, hidden dim 384, 6 heads (head size 64), 32K BPE vocabulary, and context length 512. Training uses a chunked WKV-7 kernel (chunk_size=64, diagonal-only intra-chunk approximation), which affects both training gradient flow and inference; exact-WKV effects are not measured and are noted in the Limitations section.

Evaluation protocol. BLiMP accuracy is computed as the fraction of minimal pairs where the

grammatical sentence has higher log-probability under the forward CLM head. Checkpoints are selected by validation loss (best-loss) and by BLiMP sweeps for peak-BLiMP, on a 2k-step grid (2k through 18k, nine checkpoints per seed) for vanilla and BiRWKV-7 and on a denser 500-step grid for the submitted vanilla-skip model (30 candidates for seed 42); the unequal grid inflates vanilla-skip’s peak margin over BiRWKV-7 and is quantified in Limitations. The trajectory BLiMP scores in Figure 1 and Tables B.1, 1, B.2, B.3, B.5, B.6, and B.7 use this internal MLX evaluator; the official BabyLM pipeline (Table 2) reads about 1.25 pp lower at the same checkpoint (vanilla-skip step 18k: 69.95% internal vs 68.70% official). Every within-model degradation comparison in this paper is computed on the internal evaluator alone, so this internal-to-official offset does not enter those comparisons; it bears only on the absolute official submission numbers. Repeated evaluation of an identical checkpoint varies by up to ~ 0.23 pp at a single point, from nondeterministic GPU reduction order in the MLX/Metal backend (measured by re-scoring a single near-peak seed-42 checkpoint three times: 68.08, 68.31, 68.21). We therefore treat ~ 0.23 pp as the observed single-point rerun variation, and ~ 0.3 pp as the corresponding tolerance on a quantity that differences two such measurements (a peak-to-final drop, or a checkpoint-to-checkpoint gap). This tolerance is measured at stable, near-peak checkpoints. The degraded step-18k vanilla checkpoints vary more: independent evaluation sweeps of the same checkpoint can differ by up to ~ 1.5 pp (for instance seed 777 reads 67.55% and 66.10% in two separate sweeps), so the vanilla peak-to-final drop magnitudes in Table 1 are accurate only to about that range. The direction of the vanilla decline is corroborated independently by rising held-out validation loss (Appendix B.9), so the degradation itself is robust even where its exact per-seed magnitude is not tightly pinned.

Training schedule. Training alternates between two objectives at each step:

- **CLM + MTP** (even steps): forward pass only. Cross-entropy on next-token targets (weight 0.60) + MTP delta heads for $t+2$, $t+3$ (weight 0.15).
- **MLM** (odd steps): masked forward + backward passes over shared weights. Reconstruct-

tion from 15%-masked context (weight 0.25).

Combined loss: $L = 0.60 \times L_{\text{CLM}} + 0.25 \times L_{\text{MLM}} + 0.15 \times L_{\text{MTP}}$. This alternation yields 9k effective CLM gradient updates over 18k total steps, exactly half of vanilla’s full exposure, with the remaining odd steps computing MLM reconstruction through both forward and backward passes.

Design principle. BiRWKV-7 differs from vanilla RWKV-7 in two ways: it adds auxiliary objectives (MLM, MTP, n-grams) and halves CLM exposure through objective alternation. These two factors are perfectly confounded inside BiRWKV-7. The vanilla-skip model separates them: it keeps the halved CLM update count while removing all auxiliary objectives, n-gram features, and the backward pass entirely. Its stability is therefore consistent with reduced CLM-objective updates rather than bidirectional auxiliary training, though the update-space learning-rate trajectory is coupled to that update count and is discussed in Section 4.

Baselines. *Vanilla RWKV-7*: same architecture and hyperparameters as BiRWKV-7, CLM-only (no MLM, MTP, n-grams, or backward pass), 18k CLM gradient updates over 18k steps. *Vanilla-skip*: same as vanilla but with CLM updates halved by step-skipping; odd steps advance the LR schedule with no gradient update, giving 9k effective CLM updates over 18k total steps.

Warm-start reference (non-competitive). A confounded warm-start run using Common Crawl initialization and an SGDR (stochastic gradient descent with warm restarts) schedule is documented in Appendix A. It does not qualify for the competitive track and is not used in any primary claims.

4 Results

4.1 Late-training degradation and its mitigation

Vanilla RWKV-7 (27,369,984 parameters, CLM-only over 18k steps) degrades late in training. It peaks between steps 4k and 6k, then loses 2.57 to 3.85 pp of BLiMP by step 18k across all three seeds (Figure 1; per-seed values in Table 1). Held-out validation loss rises alongside the decline (Figure 2), which argues against a BLiMP-specific scoring artifact. On seed 42, validation loss bottoms out near step 4k and then climbs about 1.3

nats to step 18k while training loss keeps falling, the classic overfitting signature.

Halving how often the CLM objective updates the weights mitigates the decline. BiRWKV-7 alternates CLM on even steps and MLM on odd steps, giving 9k effective CLM updates over 18k total, and stays stable through training. Its seed-42 trajectory holds near 69% from step 10k on, reaching 69.45% at step 16k (Appendix B.1, Table B.1), in the range of Baby Llama (58M parameters, 69.8% BLiMP; Timiryasov and Tastet, 2023, an earlier evaluation round, reported for context only, not as a controlled parameter-efficiency claim).

Auxiliary objectives and halved CLM exposure are confounded inside BiRWKV-7 (Section 3). The vanilla-skip model separates them. It shares the RWKV-7 backbone and the same per-scheduled-step LR schedule, carries no auxiliary objectives, and halves CLM exposure purely through step-skipping: odd steps advance the schedule but apply no gradient update, giving 9k effective CLM updates over 18k total. With no auxiliary objectives, it still exceeds BiRWKV-7 on peak BLiMP across all three seeds (+0.83, +2.58, +1.81 pp). On seed 42 it peaks at 70.28% (step 7,500) and drops only 0.33 pp by step 18k, about 1 pp above vanilla’s peak, with the dominant contrast in the post-peak trajectory (Figure 1a). Table 1 reports all three model types across all three seeds.

Vanilla-skip has no n-gram features, which add +1.77 pp at inference for BiRWKV-7 (Appendix B.3), yet it outperforms BiRWKV-7 on all three seeds. Halving CLM-objective updates is therefore enough to produce stability with no auxiliary objectives.

Reduced data exposure is not necessary for that stability either. Step-skipping halves how much data vanilla-skip sees (~ 5 epochs versus ~ 10 for vanilla; Table A.1), so less overfitting is a candidate explanation. BiRWKV-7 removes it with no extra run: it samples a fresh batch on every step, so it sees all 18k data steps (~ 10 epochs, as vanilla) and makes 18k optimizer updates (9k CLM plus 9k MLM), yet applies only 9k CLM-objective updates and stays stable. Data exposure alone therefore cannot explain the contrast. BiRWKV-7 matches vanilla in data steps and optimizer-update events, but the differing CLM and MLM gradients need not produce the same parameter-space movement. The two stable arms share one more thing

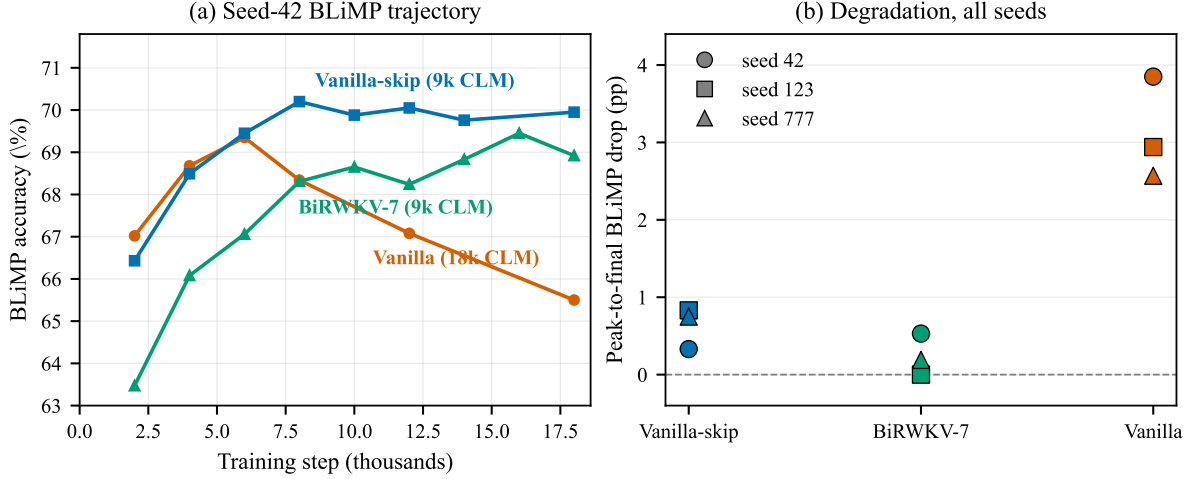


Figure 1: Halving CLM updates mitigates late-training BLiMP degradation. (a) Seed-42 BLiMP trajectory for vanilla RWKV-7 (18k CLM updates), vanilla-skip (9k, step-skipping), and BiRWKV-7 (9k, objective alternation). Vanilla peaks near step 6k and then declines; vanilla-skip peaks early and holds, while BiRWKV-7 rises and plateaus near 69% from step 10k on, broadly stable with only a small late decline. Each arm is drawn from a single evaluation sweep on the 2k-step grid; vanilla-skip’s true peak (70.28% at step 7,500) is located off this grid on a denser 500-step sweep. (b) Peak-to-final BLiMP drop for all three conditions across seeds 42, 123, and 777 (plotted as positive magnitudes; larger is worse). The two halved-update conditions sit near zero (0.00 to 0.83 pp); full-exposure vanilla drops 2.57 to 3.85 pp. Exact per-seed values are in Table 1.

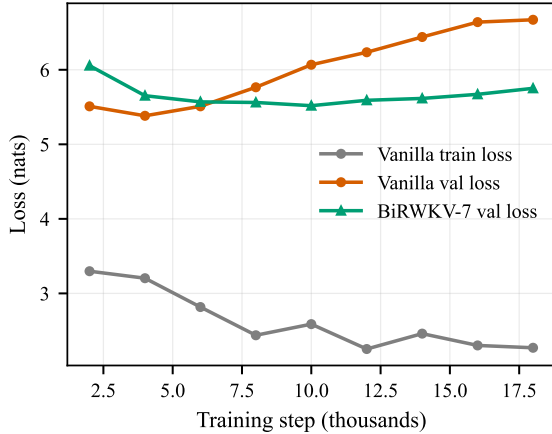


Figure 2: Seed-42 loss trajectory over the 2k-step grid. Vanilla RWKV-7’s validation loss reaches its minimum near step 4k and then rises about 1.3 nats while training loss keeps falling. BiRWKV-7, with half the CLM-objective updates, instead reaches its validation-loss minimum near step 10k and rises only about 0.2 nats thereafter. The per-category BLiMP breakdown of the same degradation is in Appendix B.9.

beyond halved CLM updates. Both spread their CLM updates across a cosine that reaches its floor in 9k CLM updates rather than 18k, so both incur a faster decay in CLM-update space, which the next subsection probes.

4.2 The learning-rate schedule confound

Halving CLM updates changes two things at once, and our design does not separate them. Skipping every other step leaves the per-scheduled-step learning rate unchanged: the schedule is indexed by scheduled step, so vanilla and vanilla-skip use the same rate at each scheduled step, and their mean per-update rate is essentially identical ($\sim 2.15e-4$ for both). What differs is the update-space trajectory. Because vanilla-skip traverses the cosine over 9k surviving updates rather than 18k, its rate decays about twice as fast per update, and at a matched update index it sees a lower rate than vanilla outside the brief warmup. Reduced CLM update count and the shape of the update-space schedule are thus entangled. Two probes bound what each contributes.

The first isolates the schedule’s terminal rate. We trained vanilla RWKV-7 at a constant learning rate (warmup, then held flat) for the full 18k updates, with no cosine decay, at three rates in the band where the cosine model is still declining ($1e-4$, $2e-4$, and $3e-4$; seed 42). At the two higher rates it peaks early and then sheds 3.0 to 4.0 pp by step 18k; at $1e-4$ it loses 1.2 pp. These rates sit above the cosine schedule’s terminal value, so the control never reaches the cosine model’s final low-rate regime. The narrow claim is that reaching a

Model Type	Seed	Peak BLiMP	Peak Step	Final (18k)	Degradation
Vanilla-skip (9k CLM, 0 aux)	42	70.28%	7,500	69.95%	−0.33 pp
Vanilla-skip (9k CLM, 0 aux)	123	70.55%	8,500	69.72%	−0.83 pp
Vanilla-skip (9k CLM, 0 aux)	777	69.96%	10,500	69.21%	−0.75 pp
BiRWKV-7 (9k CLM + 9k MLM)	42	69.45%	16,000	68.92%	−0.53 pp
BiRWKV-7 (9k CLM + 9k MLM)	123	67.97%	18,000	67.97%	0.00 pp
BiRWKV-7 (9k CLM + 9k MLM)	777	68.15%	14,000	67.96%	−0.19 pp
Vanilla (18k CLM, 0 aux)	42	69.35%	6,000	65.50%	− 3.85 pp
Vanilla (18k CLM, 0 aux)	123	68.78%	4,000	65.84%	− 2.94 pp
Vanilla (18k CLM, 0 aux)	777	68.67%	6,000	66.10%	− 2.57 pp

Table 1: All three model types across all three seeds. Halved-exposure models (vanilla-skip and BiRWKV-7) maintain stability in every case; full-exposure vanilla degrades 2.57 to 3.85 pp consistently. Vanilla-skip exceeds BiRWKV-7 on peak BLiMP by +0.83 pp (seed 42), +2.58 pp (seed 123), and +1.81 pp (seed 777); these peak margins are measured on an unequal checkpoint grid (denser for vanilla-skip) and are upper bounds, and the stability contrast is conservative for the same reason (Limitations).

low terminal rate is not necessary for the decline: sustained constant-rate CLM training degrades on its own. Repeating the 3e-4 control on seeds 123 and 777 gives peak-to-18k drops of 3.99, 2.66, and 3.56 pp for seeds 42, 123, and 777 (constant-LR figures, separate from the cosine degradation of Table 1), so the constant-rate degradation holds across all three seeds.

The second probe holds update count fixed and moves only the schedule. At full 18k-update exposure (seed 42) we vary when the cosine reaches its floor, over the first N updates (Figure 3). Flooring the rate earliest, over 4500 updates, ends at 68.43% BLiMP against 65.50% for the full vanilla schedule, a 2.9 pp gap at fixed update count. That gap is about twice the up-to-~1.5 pp same-checkpoint evaluation spread at degraded step-18k checkpoints (Section 3), so the extreme contrast is robust; the two intermediate horizons ($N=9000$, 13500) end only 0.51 pp apart, inside that spread, so we do not read their ordering as resolved (Appendix B.7). The update-space learning-rate trajectory therefore also modulates the decline at fixed update count.

Together the two probes place the effect within late-phase training dynamics. The two halved-CLM arms differ from vanilla in different subsets of architecture, auxiliary objectives, data exposure, and total optimizer updates, and no single one of those is shared by both, yet both stay stable. What both share with each other, and not with vanilla, is the halved CLM update count together with the faster CLM-update-space learning-rate decay that comes with it. The anneal probe shows the decline

also moves at fixed update count, so update count is not the sole lever; it stays coupled to the update-space schedule, and we do not decompose their individual shares. Vanilla-skip carries no MLM and still holds within 0.83 pp of its peak in every seed, so MLM gradient diversity is not necessary to prevent the decline. All arms share the RWKV-7 architecture, so these results speak to CLM exposure within one recurrent family, not to other architectures or transformers.

4.3 Official evaluation and secondary findings

On the official BabyLM 2026 pipeline, the GPT-2 baseline (GPT-2-base retrained on the Strict-Small data) scores 65.23% BLiMP on the shared leaderboard (BabyLM Organizers, 2026). Our primary submission (vanilla-skip seed 42, step 18k) scores 68.70% BLiMP, 3.47 pp above the baseline, at the submitted model’s 27,369,984 parameters against the baseline’s 98.4M, about 28% as many parameters. The comparison is uncontrolled and only situates our result in the BabyLM 2026 framework. The advantage is specific to BLiMP: on the rest of the suite the model is weak (Text Average 32.8 in Table 2), so we frame the central claim as one about grammatical acceptability under CLM training, not broad task competence.

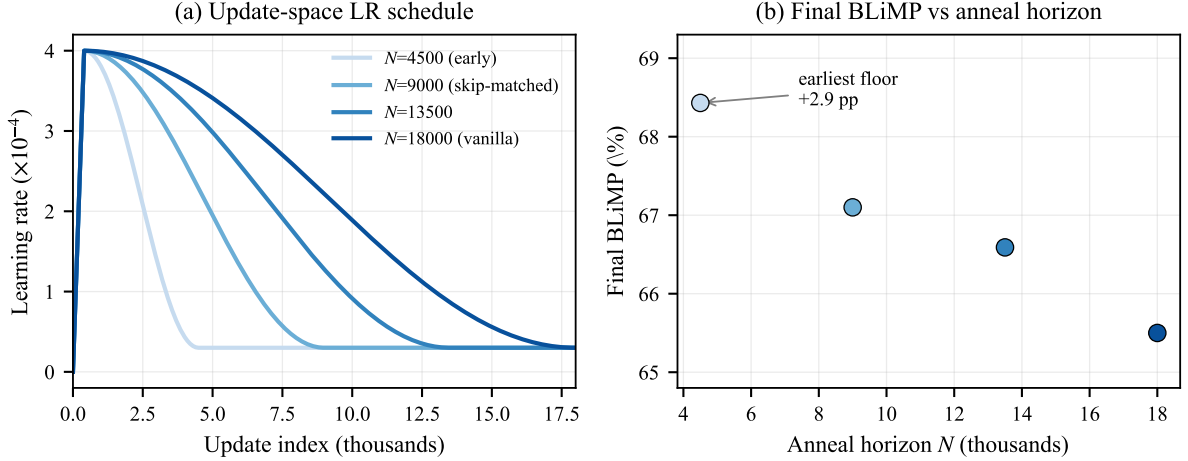


Figure 3: Anneal-timing probe (seed 42, full 18k-update exposure throughout). (a) Update-space learning-rate schedules: the cosine reaches its floor over the first N updates and is then held flat, for $N=4500$, 9000, 13500, and 18000; the $N=18000$ curve is the full vanilla schedule and the $N=9000$ curve reaches the floor at the same rate as vanilla-skip. All start at peak $4e-4$ and end at $3e-5$ (Table A.1). (b) Final BLiMP against anneal horizon N . Flooring earlier lifts final BLiMP: the earliest floor ($N=4500$) ends 2.9 pp above the full vanilla schedule ($N=18000$). The two intermediate horizons ($N=9000$, 13500) differ by only 0.51 pp, within the ~ 1.5 pp same-checkpoint evaluation spread at degraded step-18k checkpoints (Section 3), so their ordering is not resolved.

Task	Score
BLiMP (filtered)	68.70%
BLiMP Supplement	57.24%
COMPS	52.55%
Entity Tracking	19.98%
EWoK	1.53%
Reading (Eye-Track. 1.68, Self-Paced 0.05)	0.86
GLUE (avg, Appendix B.4)	59.03%
AoA	2.85
Text Average (8-col)	32.8
BoolQ (GLUE component)	66.06%

Table 2: Official BabyLM 2026 evaluation for vanilla-skip seed 42 at step 18k. Tasks are zero-shot except the GLUE fine-tuning average; the table reports the official task scores and the resulting Text Average. Evaluation and aggregation details are given in Appendix A.1.

Three secondary findings round out the evaluation, each detailed in the appendix. First, the best-validation-loss and peak-BLiMP checkpoints differ in all three BiRWKV-7 seeds; a resolvable BLiMP advantage appears in seeds 42 and 123, whereas seed 777’s 0.17 pp gap is within the measured ~ 0.3 pp rerun tolerance (Appendix B.2, Table B.2). Thus validation loss alone may miss peak grammatical performance in this setting. Second, disabling n-gram features at inference on seed 42 costs 1.77 pp of BLiMP (Appendix B.3, Table B.3), larger than either checkpoint-selection difference. Third, GLUE fine-tuning of the submitted model averages 59.03% across seven tasks

(Appendix B.4, Table B.4), below typical transformer baselines at this parameter count, as expected for causal recurrent models.

Entity Tracking is the one task where internal and official scores warrant separate comment. For the submitted model they agree closely: vanilla-skip at step 18k scores 18.56% internally and **19.98%** officially, a ~ 1.4 pp gap with the official score above the internal one (the reverse of the BLiMP offset for that checkpoint in Section 3, where the internal sweep read higher). Both are low in absolute terms and near chance for this task, so this is a task the model does not solve. The one large divergence is the fused BiRWKV-7 representation, which peaks at 40.20% internally but scores only **22.45%** officially; we treat the official value as authoritative and do not claim the internal peak (Appendix B.8).

5 Conclusion

In small recurrent language models on a Strict-Small budget, halving CLM exposure mitigates late-training BLiMP degradation, and auxiliary objectives are not necessary for it. Under full CLM exposure, vanilla RWKV-7 peaks early and then sheds several points of BLiMP by step 18k across seeds 42, 123, and 777. Halving that exposure mitigates the decline, whether by alternating a backward MLM objective (BiRWKV-7) or by skip-

ping every other gradient step (vanilla-skip); the auxiliary-free vanilla-skip stays within 0.83 pp of its peak in every seed, so the auxiliaries are not what buys the stability. Two effects stay coupled, though. A constant-rate control degrades with no annealing on all three seeds, so a decaying schedule is not necessary; yet both halving arms also decay twice as fast in CLM-update space, and an anneal-timing probe shows an earlier floor lifts final BLiMP from 65.50% to 68.43% (seed 42; Appendix B.7). Update count and the update-space schedule co-vary in every arm, so we bound rather than decompose their shares. A clean decoupling, for instance 9k CLM updates run over an 18k-update cosine so the rate ends mid-schedule rather than at the floor, is the key open experiment. Finally, the best-validation-loss and peak-BLiMP checkpoints differ in all three BiRWKV-7 seeds. The BLiMP advantage is resolvable in seeds 42 and 123, while seed 777’s 0.17 pp gap is within the measured ~ 0.3 pp rerun tolerance; validation loss alone may therefore miss peak grammatical performance in this setting.

Availability. Training and evaluation code, the submitted checkpoint, per-table reproduction commands, and result provenance are prepared for public release under an open-source license, available from the author and to be deposited in a public code and model repository. Each headline number reads from a single named checkpoint and evaluation command.

Limitations

Several constraints limit how far these findings generalize:

1. **Three seeds only.** Effect-size estimates are uncertain. Vanilla-skip is the better-separated claim: all three seeds show clear stability (degradation 0.33 to 0.83 pp, against vanilla’s 2.57 to 3.85 pp) and beat BiRWKV-7 on peak BLiMP (+0.83, +2.58, +1.81 pp). For BiRWKV-7 itself the evidence is thinner. Only seed 42 (-0.53 pp) shows a peak-to-final drop above the ~ 0.3 pp rerun tolerance, and even that is small; seed 123 (0.00 pp) and seed 777 (-0.19 pp) both sit within the tolerance, so they are better read as no measurable change than as exact stability. For seed 123 the peak is the terminal (18k) checkpoint, so its 0.00 pp drop is definitional and that trajectory may still be non-decreasing at 18k rather than shown to have plateaued. Replication across more seeds and larger model sizes would help.
2. **Multiple comparisons and evaluation asymmetry.** Searching 27 peak candidates ($9 \text{ checkpoints} \times 3 \text{ seeds}$) may inflate the apparent checkpoint-divergence gap. The peak-BLiMP checkpoint is also selected from a denser sweep for vanilla-skip (a 500-step grid, 30 candidates for seed 42) than for vanilla and BiRWKV-7 (a 2k-step grid, 6 to 9 candidates); a denser grid can only raise the recorded peak, which inflates vanilla-skip’s peak-BLiMP margin over BiRWKV-7 while making its peak-to-final stability contrast conservative, since a higher recorded peak can only widen the measured drop. The three arms’ trajectories are also read from separate evaluation sweeps that can differ by more than a point on an identical checkpoint, so cross-arm comparisons of absolute peak BLiMP (for instance vanilla-skip’s roughly 1 pp margin over vanilla’s peak) may carry a between-sweep offset; every peak-to-final degradation, by contrast, is computed within a single sweep and does not. At the degraded step-18k checkpoints, independent evaluation sweeps of the same checkpoint can differ by up to ~ 1.5 pp (Section 3), so the per-seed vanilla drop magnitudes carry that uncertainty, though the direction of the decline is corroborated by rising validation loss and is robust to it. Separately, validation loss is estimated from ~ 50 batches ($\sim 409\text{K}$ tokens) while BLiMP evaluates a 200-pair-per-paradigm subsample (about 13K of the roughly 65K pairs); this noisy-scalar versus larger-sample asymmetry does not, however, explain the consistent degradation pattern.
3. **Chunked WKV approximation.** The diagonal-only intra-chunk kernel affects both training gradient flow and inference. Exact-WKV effects are unmeasured for the from-scratch models in this paper.
4. **Mechanism of halving.** Under step-skipping, odd steps advance the LR schedule with no gradient update; under MLM alternation, odd steps update shared weights via MLM gradients. Both give $\sim 9\text{k}$ CLM up-

dates over 18k steps, but whether this mechanistic difference matters for long-run dynamics is unexplored. Reduced CLM update count is also not fully separated from the update-space learning-rate trajectory, since halving both shortens the cosine in CLM-update space; BiRWKV-7’s total optimizer-update schedule matches vanilla’s density, and only in CLM-update space does it share vanilla-skip’s faster decay. The two probes in Section 4 bound but do not decompose the two effects.

5. **One architecture.** All arms are RWKV-7 at 27 to 28M parameters (27,369,984 for the vanilla and vanilla-skip backbone; 28,157,569 for BiRWKV-7 with its auxiliary components), so the degradation and its mitigation are established within a single recurrent family at one small scale. Whether they transfer to other linear-time architectures (Mamba, S4, xLSTM) or to transformers is untested, so the title’s “recurrent language models” should be read as scoped to this family pending cross-architecture replication.

None of these constraints change the core observation: halving CLM exposure mitigates the degradation in every seed we ran, under either alternation or step-skipping. In practice, halving CLM update frequency reduced the late degradation we observed under cosine scheduling on this data-limited corpus. In this setting, validation loss alone may miss peak grammatical performance.

Acknowledgments

Generative AI assistants (large language models) were used throughout this work: for writing and debugging training and evaluation code, scaffolding experiment and evaluation harnesses, exploratory data analysis, and drafting and revising the manuscript text. The author designed the study, ran and supervised all experiments, and checked every reported number against its primary run artifact. The author is solely responsible for the scientific claims and for the final content of this paper.

References

- BabyLM Community. 2026. BabyLM-community/BabyLM-2026-Strict-Small: Official training corpus for the 2026 BabyLM

challenge (strict-small track). <https://huggingface.co/datasets/BabyLM-community/BabyLM-2026-Strict-Small>. Accessed: June 2026.

BabyLM Organizers. 2026. babylm-org/babylm-eval: Evaluation pipeline for the BabyLM 2026 challenge. <https://github.com/babylm-org/babylm-eval>. Accessed: June 2026.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. 2024. xLSTM: Extended long short-term memory. *arXiv preprint arXiv:2405.04517*.

Lucas Georges Gabriel Charpentier and David Samuel. 2023. Not all layers are equally as important: Every layer counts BERT. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*.

Lucas Georges Gabriel Charpentier and David Samuel. 2024. GPT or BERT: why not both? In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 262–283, Miami, FL, USA. Association for Computational Linguistics.

Michael K. Chen, Xikun Zhang, Fan Bai, Zhengding Hu, and Zhen Wang. 2026. Demystifying training-time augmentation for data-constrained language model pretraining. *ArXiv:2606.16246*.

Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Jaap Jumelet, Tal Linzen, Aaron Mueller, Suchir Salhan, Raj Sanjay Shah, Alex Warstadt, and Ethan Gotlieb Wilcox. 2026. BabyLM turns 4 and goes multilingual: Call for papers for the 2026 BabyLM workshop. *Preprint*, arXiv:2602.20092.

Leshem Choshen, Ryan Cotterell, Michael Y. Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. [call for papers] the 2nd BabyLM challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2404.06214*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Lukas Edman and Alexander Fraser. 2025. Mask and you shall receive: Optimizing masked language modeling for pretraining BabyLMs. In *Proceedings of the First BabyLM Workshop*.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

- Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently modeling long sequences with structured state spaces. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Patrick Haller, Jonas Golde, and Alan Akbik. 2025. Sample-efficient language modeling with linear attention and lightweight enhancements. In *Proceedings of the First BabyLM Workshop*, pages 175–191, Suzhou, China. Association for Computational Linguistics.
- Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. 2023. MLX: Efficient and flexible machine learning on Apple silicon. <https://github.com/ml-explore/mlx>.
- Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, volume 36, pages 50358–50376. Curran Associates, Inc. ArXiv:2305.16264.
- Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Koccon, Jiaming Kong, Bartłomiej Koptyra, and 13 others. 2023. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14048–14077, Singapore. Association for Computational Linguistics.
- Bo Peng, Ruichong Zhang, Daniel Goldstein, Eric Alcaide, Xingjian Du, Haowen Hou, Jiaju Lin, Jiaxing Liu, Janna Lu, William Merrill, Guangyu Song, Kaifeng Tan, Saiteja Utpala, Nathan Wilce, Johan S. Wind, Tianyi Wu, Daniel Wuttke, and Christian Zhou-Zheng. 2025. RWKV-7 “Goose” with expressive dynamic state evolution. *arXiv preprint arXiv:2503.14456*.
- Matthew Theodore Roque and Dan John Velasco. 2025. Beyond repetition: Text simplification and curriculum learning for data-constrained pretraining. In *Proceedings of the First BabyLM Workshop*, pages 246–255.
- David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*.
- Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. 2023. Simplified state space layers for sequence modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. Retentive network: A successor to Transformer for large language models. *arXiv preprint arXiv:2307.08621*.
- Joonas Tapaninaho. 2025. MoEP: Modular expert paths for sample-efficient language modeling. In *Proceedings of the First BabyLM Workshop*.
- 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*.
- 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.
- Rui Yamamoto and Keiji Miura. 2025. FORGETTER with forgetful hyperparameters and recurring sleeps can continue to learn beyond normal overfitting limits. In *Proceedings of the First BabyLM Workshop*, pages 91–99.
- Amir Zandieh, Majid Daliri, Majid Hadian, and Vahab Mirrokni. 2025. TurboQuant: Online vector quantization with near-optimal distortion rate. *Preprint*, arXiv:2504.19874.

A Training Configuration

A.1 Full Reproducibility Details

Data order. Training draws windows i.i.d. rather than shuffling epochs. At each step the loader samples `batch_size` start positions uniformly at random, with replacement, over the token memmap and takes the contiguous `ctx_len+1` window at each (`mlx_train.py`, `sample_batch`, `default schedule="random"`), so a single NumPy generator seeded by the run seed fixes the entire data order and it reproduces from the seed alone. Reported epoch counts are therefore total tokens seen divided by corpus size, not structured passes over a shuffled dataset; at matched steps all arms share the same expected per-token exposure, and MLM masking (BiRWKV-7) is resampled per step from the same generator.

Hyperparameter / Detail	Value
Parameters	27,369,984 backbone (vanilla, vanilla-skip); 28,157,569 for BiRWKV-7 with auxiliary components (<code>model.num_parameters()</code>), tied input/output embeddings)
Architecture	8 layers, hidden dim 384, head size 64 (6 WKV heads), attention-free
Optimizer	AdamW (beta1=0.9, beta2=0.999)
Weight decay	0.01
Gradient clipping	1.0 (max norm)
Learning rate schedule	Cosine; peak 4e-4, final 3e-5, 400-step linear warmup
Total steps	18,000 scheduled (vanilla-skip: 9,000 data steps $\approx$ 5 epochs over 14.87M tokens, since odd steps skip with no forward pass; full-CLM vanilla and BiRWKV-7: all 18,000 are data steps $\approx$ 10 epochs)
Batch size	16 sequences
Context length	512 tokens (sliding window on token memmap)
MLM mask probability	15% (BERT-style: independent replacement)
N-gram alpha init	0.10 (learned; stabilizes at $\sim$ 0.13)
Random seed (primary)	42 (seeds 123, 777 for replication)
Tokenizer	BPE, 32K vocab (trained on BabyLM Strict-Small)
Dataset format	uint16 token memmap (Strict-Small train split for training; separate held-out <code>val.bin</code> for validation)
Software versions	Python 3.11 (patch version not recorded); MLX 0.31.2; <code>mlx-metal</code> 0.31.2; <code>mlx-lm</code> 0.31.3
Training-code revision	65685aaf6ba719a990cb5235e0486ab57a8a660a (last committed revision of <code>scripts/mlx_train.py</code>)
Hardware and OS	Apple M5 Max, 128 GB unified memory; macOS 26.5.2 (build 25F84)

Table A.1: Full reproducibility details for the from-scratch models.

Validation construction. The training-time held-out estimate uses a separate uint16 token memmap `data/babylm/val/val.bin` (35,085,463 tokens), a distinct file from the Strict-Small training memmap used for optimization. It is built from two held-out shards of the BabyLM Strict text (about 21.9M whitespace words), tokenized with the same 32K BPE tokenizer as the training data; it serves only as a fixed held-out corpus for tracking generalization loss and never contributes a gradient update. At each validation event, the loader samples 50 batches of 16 contiguous 512-token windows with replacement; each window supplies the input sequence and its one-token-shifted next-token targets. Start positions are drawn by the same NumPy generator used by the run, so validation batches are resampled at each event rather than kept as a fixed panel. No optimizer update or MLM masking is applied to these batches.

Official evaluation reproducibility. Zero-shot scores use `babylm-org/babylm-eval` at commit 3bf5142; the GLUE fine-tuning average uses `babylm/evaluation-pipeline-2025` at commit bf55c11. The seven per-task GLUE scores use the official leaderboard primary metric (F1 for MRPC and QQP, accuracy for the other five), and the reported average is their unweighted mean, recomputed from the per-task pipeline artifacts, so it can differ slightly from the server-side aggregate. The GPT-2 baseline BLiMP

score (65.23%) is the published leaderboard value, read from the `babylm-eval` results table for the Baseline-GPT2-Strict-Small row; we did not re-run that baseline at commit 3bf5142. COMPS probes conceptual properties and EWoK probes world knowledge; the EWoK column reports the value returned by the official 2026 scorer without modification. GlobalPIQA was not part of the Strict and Strict-Small zero-shot suite at the frozen submission-time commit (3bf5142); it was added to the pipeline in a later revision, so it does not appear in this table. The Text Average is recomputed from the eight sourced per-task cells, so it can differ slightly from the server-side aggregate; WUG morphology tasks from earlier rounds are not scored in the 2026 pipeline.

Kernel approximation. Chunked WKV-7 (`chunk_size=64`, diagonal-only intra-chunk) processes each length-512 sequence in chunks of 64 rather than step by step, which sharply reduces Metal kernel-dispatch overhead at T=512. Full recurrence correction is omitted within each chunk. This efficiency makes training recurrent models on consumer hardware feasible without a GPU cluster; exact-WKV effects are unmeasured for from-scratch models.

Code and checkpoints. Seed checkpoints (2k to 18k in 2k-step increments) are `.safetensors` weights with `metadata.json`. Reproducing any main finding requires only the seed number, one

evaluation command, and the checkpoint file.

A.2 Warm-Start Run (Non-Competitive)

A model pre-trained on ~ 49 M CC tokens, fine-tuned under SGDR ($T_0=3$ k), 14,250 total steps, reaching 70.8% BLiMP. Four factors differ from competitive runs (initialization, LR schedule, data volume, total steps), so no single factor can be credited for any BLiMP gap.

B Supplementary Material

All BLiMP scores in the main text use the forward CLM head only. This appendix collects the tables and diagnostics supporting the main results: the BiRWKV-7 seed-42 trajectory (Appendix B.1), the checkpoint-selection divergence (Appendix B.2), the n-gram inference ablation (Appendix B.3), GLUE fine-tuning detail (Appendix B.4), TurboQuant recurrent-state quantization (Appendix B.5), the vanilla-skip trajectories for seeds 123 and 777 (Appendix B.6), the anneal-timing probe (Appendix B.7), the entity-tracking reconciliation (Appendix B.8), and the overfitting diagnostics (Appendix B.9).

B.1 BiRWKV-7 Seed-42 Trajectory

Table B.1 reports the BiRWKV-7 seed-42 trajectory at key checkpoints, showing both the best-loss/peak-BLiMP divergence (detailed in Appendix B.2) and the absence of late degradation.

B.2 Checkpoint-Selection Divergence

Across all three BiRWKV-7 seeds, the best-validation-loss checkpoint is never the same as the peak-BLiMP checkpoint (Table B.2). The gaps are 0.80 pp (seed 42), 0.69 pp (seed 123), and 0.17 pp (seed 777). The first two exceed the measured ~ 0.3 pp rerun tolerance; the seed-777 gap is within it.

Seeds 42 and 123 show gaps of 0.80 and 0.69 pp, both above the ~ 0.3 pp tolerance for a difference of two BLiMP points (Section 3). Seed 777’s 0.17 pp gap sits within that tolerance, so the checkpoint-selection divergence rests on seeds 42 and 123. With only three seeds, parametric confidence intervals are not meaningful, so the pattern is reported directionally.

B.3 N-gram Contribution at Inference

To assess whether n-gram features contribute beyond their training effect, we disable them at inference on seed 42 ckpt16000 (Table B.3).

Checkpoint	Step	Val Loss	BLiMP (CLM)
ckpt0002000	2k	6.06	63.47%
ckpt0010000	10k	5.52	68.65%
ckpt0016000	16k	5.67	69.45%
ckpt0018000	18k	5.75	68.92%

Table B.1: BiRWKV-7 seed 42 trajectory (key checkpoints). The best-loss checkpoint (step 10k, val loss 5.52) and the peak-BLiMP checkpoint (step 16k, 69.45%) do not coincide; the 0.80 pp gap means loss-based early stopping would miss the peak-BLiMP checkpoint. Unlike vanilla RWKV-7 (Figure 1), the model holds stable rather than degrading.

Seed	Best-loss		Peak-BLiMP		BLiMP gap (pp)
	Step	Val	Step	Val	
42	10k	5.52	16k	5.67	0.80
123	6k	5.49	18k	5.66	0.69
777	8k	5.36	14k	5.52	0.17

Table B.2: Checkpoint-selection divergence across all three BiRWKV-7 seeds. The best-validation-loss and peak-BLiMP checkpoints differ in every seed. The BLiMP gaps in seeds 42 and 123 exceed the measured ~ 0.3 pp rerun tolerance, whereas seed 777’s 0.17 pp gap is within it. The peak-BLiMP checkpoint has a higher validation loss in each row. Values are the training-time held-out loss estimates from each seed’s trajectory summary.

B.4 GLUE Fine-Tuning

GLUE fine-tuning results for the submitted model (vanilla-skip seed 42, step 18k, last-token hidden state with left-padding) are in Table B.4.

B.5 TurboQuant Recurrent-State Quantization

Recurrent KV states from seed 42 (both ckpt16000 and ckpt18000) are reported in Table B.5.

B.6 Vanilla-Skip BLiMP Trajectories (Seeds 123 and 777)

Table B.6 reports the vanilla-skip trajectories for seeds 123 and 777; both peak above 69.9% and hold within 0.83 pp to step 18k, matching the seed-42 pattern of Figure 1a.

B.7 Anneal-Timing Probe (Early Learning-Rate Floor)

The constant-rate control (Section 4) shows the late decline does not require the cosine schedule’s terminal low rate. The complementary question is whether reaching the floor *earlier* mitigates the decline. We train vanilla RWKV-7 at full exposure

Config	BLiMP (CLM)
With n-grams (alpha learned)	~69.44%
Without n-grams (alpha = 0)	~67.67%
Contribution	+1.77 pp

Table B.3: N-gram inference-time ablation on seed 42 ckpt16000. The +1.77 pp contribution is larger than the gap between the best-loss and peak-BLiMP checkpoints (0.80 pp). A training ablation would be needed to establish how much comes from training vs. inference-time features.

Task	Metric	Score
BoolQ	Acc.	66.06%
MultiRC	Acc.	56.72%
RTE	Acc.	48.92%
WSC	Acc.	59.62%
MRPC	F1	81.60%
QQP	F1	57.46%
MNLI	Acc.	42.87%

Table B.4: GLUE fine-tuning results for vanilla-skip seed 42 (step 18k), Fine-tuned with LR=3e-5, batch size 32, 10 epochs (WSC: 30 epochs). Last-token hidden state with left-padding, appropriate for causal recurrent models. Each task reports its official BabyLM 2026 leaderboard primary metric: F1 for MRPC and QQP, accuracy for the other five (each results file stores accuracy, F1, and MCC, and the leaderboard selects one metric per task). The GLUE average (59.03%, Table 2) is the unweighted mean of these seven per-task metrics. These scores are below typical transformer baselines at this parameter count; GLUE classification is a known challenge for causal recurrent models.

(the same 18k CLM updates, no step-skipping) but anneal the cosine learning rate to its terminal value over the first N updates and then hold it flat, varying N over 4500, 9000, and 13500 (seed 42). This holds update count fixed and moves only the update-space learning-rate trajectory. Table B.7 reports the peak-to-final BLiMP drops for seed 42. Repeating the default early floor ($N=9000$) across all three seeds gives drops of 1.00, 1.79, and 0.71 pp (seeds 42, 123, 777; mean 1.17, sample standard deviation 0.56), each below vanilla’s 2.57 to 3.85 pp but with a spread that leaves the tighter question, whether an early floor fully eliminates the decline, unresolved at three seeds.

B.8 Entity Tracking Reconciliation

The submitted model’s official ET (19.98%) is deterministic, reproducing across three independent runs and across first-index versus random tiebreaks applied to the same captured log-probs.

Mode	s42 Peak	s42 Final
none (FP32)	69.44%	68.92%
fp16	69.44%	68.92%
bf16	69.44%	68.92%
uniform_int8	69.44%	68.92%
uniform_int4	69.44%	68.92%
turbo_int4	69.44%	68.92%

Table B.5: TurboQuant (Zandieh et al., 2025) recurrent-state quantization (seed 42 ckpt16000 and ckpt18000). BLiMP is identical to the reported precision across all six modes, from FP32 through fp16/bf16 to int8 and per-row int4 (the FP32 peak reads 69.44% here versus 69.45% in Table B.1, a 0.01 pp difference between separate scoring runs). This is expected rather than evidence of lossless compression: BLiMP scoring inputs fit in a single recurrence chunk, so they never read a state that was compressed and carried across a chunk boundary. The table therefore reports that quantization does not disturb the single-chunk forward pass, not that the recurrent-state compression is lossless. We report BLiMP quality only and make no compression-ratio claim; the realized ratio depends on bit packing and per-row scale overhead that our feasibility path does not measure consistently.

Step	Seed 123	Seed 777
2,500	66.98%	65.86%
8,500	70.55% (peak)	69.45%
10,500	69.90%	69.96% (peak)
18,000	69.72%	69.21%

Table B.6: Vanilla-skip seeds 123 and 777 trajectories (key checkpoints). Seed 123 peaks at 70.55% (step 8,500) and ends at 69.72% at step 18k (−0.83 pp). Seed 777 peaks at 69.96% (step 10,500) and ends at 69.21% at step 18k (−0.75 pp). Both show small drops (0.83 and 0.75 pp), well below vanilla’s 2.57 to 3.85 pp under halved CLM exposure.

The internal MLX script (loading weights directly) and the official HuggingFace-converted pipeline therefore differ by only ~1.4 pp for this model, with the official score above the internal one; this is the reverse of the BLiMP offset for the submitted checkpoint in Section 3, where the internal sweep read ~1.25 pp higher than the official pipeline.

The exception is the fused BiRWKV-7 representation: 40.20% internally (step 14k) against 22.45% officially. Both official scores are low in absolute terms, in the near-chance range for this task. The source of the fused gap, likely numerical precision, batching, or HuggingFace conversion, is unresolved, and we treat the official scores as authoritative; the qualitative pattern of fused ET

Anneal N	Peak (@step)	Final (18k)	Drop
4,500 (early)	68.97% (12k)	68.43%	0.54
9,000 (skip-matched)	68.10% (6k)	67.10%	1.00
13,500 (late)	69.68% (6k)	66.59%	3.09
18,000 (vanilla)	69.35% (6k)	65.50%	3.85

Table B.7: Anneal-timing probe, seed 42, full 18k-update exposure throughout; only the learning-rate anneal horizon N varies. The 18,000 row is the vanilla cosine schedule (Table 1); the skip-matched row ($N=9000$) reaches the floor in update space at the same rate as vanilla-skip. The robust contrast is between the extremes: the earliest floor ($N=4500$) ends at 68.43% against the vanilla schedule’s 65.50%, a 2.9 pp gap, roughly twice the up-to- ~ 1.5 pp spread seen at degraded step-18k checkpoints (Section 3). The two intermediate horizons ($N=9000$, 13500) end between these extremes, but their finals differ by 0.51 pp from each other and the $N=13500$ final (66.59%) sits only 1.09 pp above the vanilla endpoint, inside that degraded-checkpoint spread, so we do not read the intermediate ordering as resolved. The $N=9000$ peak (68.10%) is also the lowest in the family, so the individual peak-to-final drops are not directly comparable across rows. This is a single-seed probe over widely separated conditions, not a powered dose-response; the three-seed replication at $N=9000$ (Section text above) has a sample standard deviation of 0.56 pp.

rising while CLM ET degrades may be real but cannot be confirmed officially.

The high early-checkpoint ET is an artifact of a barely-trained model on a near-chance task. Internally, vanilla-skip CLM peaks at 41.72% at step 500 and degrades to 18.56% by step 18k, and BiRWKV-7 CLM peaks at 36.77% then degrades to 20.43%; the step-500 peak re-confirms through the official pipeline (41.33%), so it is not a capability the converged model retains. Finally, BiRWKV-7 seed 42 shows a three-way checkpoint split, best-loss (10k), peak-BLiMP (16k), and peak-internal-fused-ET (14k), extending Table B.2 to a third axis. The fusion gate grows over training (mean sigmoid ~ 0.19 at step 2k to ~ 0.42 at step 18k), so the backward pass contributes more to the joint representation as training progresses.

B.9 Overfitting Diagnostics

The late-training degradation carries a held-out overfitting signature (Figure 2). On seed 42, over the 2k-step checkpoint grid shared with BiRWKV-7, vanilla RWKV-7’s validation loss reaches its minimum early, near step 4k (5.38 nats), then

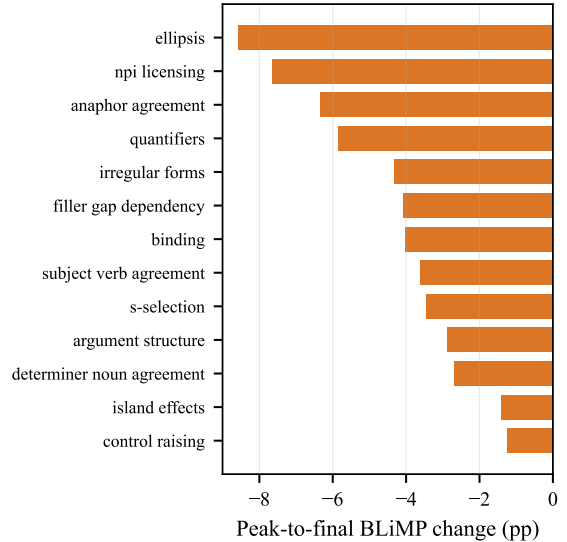


Figure B.1: Peak-to-final BLiMP change (step 6k to 18k) for all 13 high-level linguistic categories, vanilla RWKV-7 seed 42. Every category declines, from about 1 pp to about 9 pp, so the degradation is broad rather than concentrated in one fragile paradigm.

rises about 1.3 to step 18k (6.67) while its training loss keeps falling, tracking the BLiMP decline of Figure 1. BiRWKV-7, trained on the same ~ 10 epochs of data but with half the CLM-objective updates, instead reaches its validation-loss minimum near step 10k and rises only about 0.2 to step 18k, with BLiMP stable. Both use the training-time held-out estimate, so the contrast is between comparable quantities.

The degradation is broad rather than concentrated (Figure B.1). Grouping the BLiMP paradigms into their high-level linguistic categories, all 13 categories decline from peak (step 6k) to step 18k for vanilla seed 42, ranging from about 1 pp to about 9 pp. (These are within-run diagnostic re-evaluations and carry the same ~ 0.23 pp single-point rerun variation as the main tables; we read only the peak-to-final direction, which is negative in every category.) A single fragile paradigm would show one large drop against a flat remainder; instead the decline is spread across phenomena, consistent with overfitting that erodes grammatical acceptability broadly.