

What Survives the BabyLM Budget?

A Controlled Study of Sequence Mixers and Generative Objectives

Filipe Laitenberger Patrick Haller Jonas Golde Alan Akbik

Humboldt-Universität zu Berlin

filipe.laitenberger@hu-berlin.de

Abstract

BabyLM is a pressure test for language-modeling assumptions: which ideas still work when the data budget is closer to a child’s linguistic input than to web-scale pretraining? We test this question at 10M- and 100M-word budgets, holding the corpus, tokenizer, backbone size family, training schedule, and public evaluation wrapper fixed while varying two sequence mixers and six objectives. The retained scale sweep places autoregressive Transformers first in broad transfer at every tested scale and shows little gain beyond the 50M-parameter range. Corrected matched-base experiments give a more local picture: non-autoregressive systems approach the leaders on individual grammar, supplement, and entity metrics and lead controlled-probe slices, but no one system combines these gains. Paired candidate-ranking and generation tests expose the interface gap. Masked diffusion can exceed autoregression on WikiText completion or infilling, yet no tested decoder wins across LAMBADA and the two WikiText tasks, categorical flow remains unreliable across its tested decoder grid. Holding the model fixed and varying only the scoring interface bounds how much of the ranking that interface explains: supplying right context on an identical target span helps all ten non-AR cells and no AR cell, and swapping candidate ranking for open generation multiplies the between-objective spread more than tenfold – but the corrections are smaller than the transfer gap they would have to close. Small-data conclusions therefore depend on the objective, sequence mixer, decoder, and scoring interface together.

1 Introduction

BabyLM is built around the comparison that Children acquire language from a tiny fraction of the text used to train modern language models, yet many claims about new objectives or architectures

are tested only at much larger scale. The 2026 challenge keeps the 10M-word Strict-Small and 100M-word Strict English tracks, adds a multilingual track, and evaluates submitted models with shared zero-shot and fine-tuning tasks (Choshen et al., 2026). This lets us ask not only *what scores highest*, but which modeling assumptions remain useful when data, compute, and evaluation are all constrained.

Controlled comparison is already established BabyLM practice: previous entries contrast causal and masked training inside one model (Charpentier and Samuel, 2024), benchmark recurrent and linear-time mixers against attention (Haller et al., 2024, 2025), and re-audit elementary baselines under matched budgets (Păpușoi and Nisioi, 2025). Our contribution is therefore the coverage of the resulting grid. We cross the sequence-mixer axis with the *generative objective* axis, and we include the diffusion and flow families that have reshaped generative modeling at scale but have not been audited at BabyLM budgets. A *sequence mixer* is the part of each model block that exchanges information among token positions. We cross two sequence mixers, standard Transformer attention (Vaswani et al., 2017) and Gated DeltaNet (Yang et al., 2025), with six objectives: left-to-right autoregression, BERT-style masked language modeling (Devlin et al., 2019), three absorbing-mask diffusion variants (Sahoo et al., 2024, Nie et al., 2025, and Lou et al., 2024), and a non-distilled categorical flow-map endpoint objective over Gaussian-to-one-hot interpolants (Roos et al., 2026 and Davis et al., 2026). All cells share the same corpus, tokenizer, model family, training schedule, and public BabyLM evaluation wrapper, and every objective-mixer pair receives its own learning-rate sweep so that no family competes at another family’s optimizer setting.

The challenge fixes the corpus, so the design freedom that remains is how a block moves in-

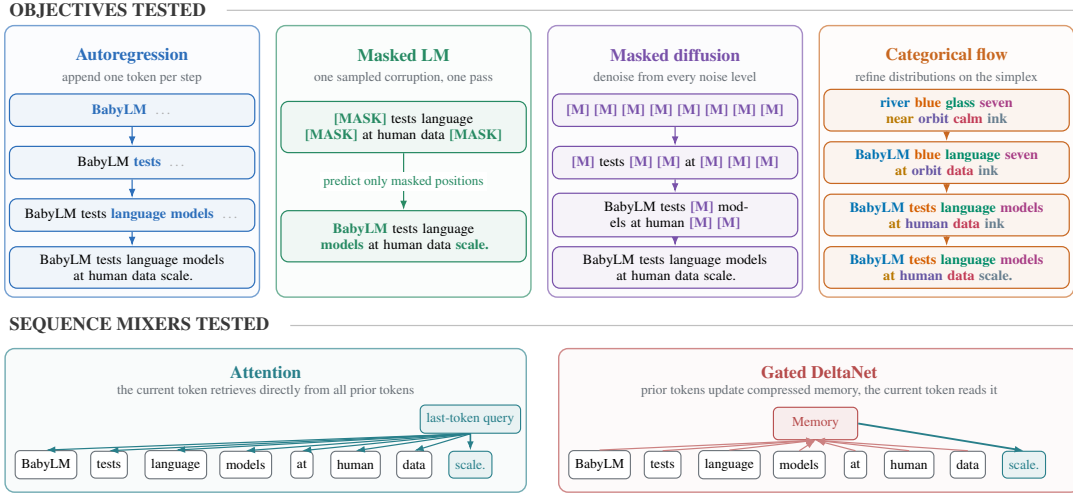


Figure 1: The two experimental axes. **Top:** AR appends tokens from left to right, MLM reconstructs one sampled mask pattern in a single pass, masked diffusion learns denoisers at all noise levels and chains them for generation, CatFlow moves continuous vocabulary states from Gaussian noise toward clean token endpoints (the displayed words are the most likely token per position). **Bottom:** the current position either attends directly to every preceding token or reads the compressed recurrent memory updated by Gated DeltaNet.

formation between positions and what prediction problem turns a sentence into gradient signal. Both axes carry an explicit small-data hypothesis. Gated DeltaNet represents the delta-rule family: its update overwrites the value stored at an addressed key rather than accumulating another association (Yang et al., 2024, 2025), and it is the mixer that has most recently replaced attention layers in a production-scale model (Merrill et al., 2026). Earlier BabyLM work found recurrent and linear-time mixers competitive at these budgets (Haller et al., 2024, 2025), Gated DeltaNet asks whether the current generation of that family still is. On the objective axis, masked diffusion is the natural “more supervision per sentence” hypothesis for scarce text: rather than one corruption level, each sentence yields a family of denoising problems indexed by noise (Austin et al., 2021; Sahoo et al., 2024). MDLM, LLaDA, and SEDD share that construction and differ in one decision each – loss weighting ($1/t$ versus uniform) and parameterization (token cross-entropy versus score entropy) – so together they separate the family from its implementation. MLM is their single-noise-level reference point, and categorical flow is the control that keeps bidirectional denoising but replaces the absorbing mask with a continuous Gaussian-to-one-hot path (Roos et al., 2026), separating what bidirectionality buys from what the mask construction buys.

Under the official public scorer, the autoregressive Transformer remains the strongest default: it

wins broad transfer throughout the retained scale sweep, while the corrected matched-base comparison finds local non-AR strengths rather than one consistently strong alternative. Diffusion and categorical flow occasionally improve token overlap on bidirectional WikiText infilling, but neither advantage becomes reliable prompt-conditioned generation across tasks, and Gated DeltaNet helps particular objective–benchmark pairs rather than supplying a uniform advantage. We therefore read the grid as a BabyLM lesson rather than a blanket ranking of model classes: small-data inductive bias helps only when it is aligned with the scoring interface and downstream use. The challenge for future BabyLM systems is not to replace attention with recurrence or autoregression with diffusion, but to design the objective, mixer, sampler, and scorer as one matched system.

2 Background

BabyLM and sample-efficient pretraining.

The BabyLM Challenge was introduced to make language-model pretraining experimentally accessible and cognitively motivated (Warstadt et al., 2023). The 2026 round keeps the Strict and Strict-Small tracks and frames the workshop around small-data training, evaluation, and architecture innovation (BabyLM Organizers, 2026). Previous strong BabyLM systems include hybrid causal and masked training recipes such as GPT-BERT (Charpentier and Samuel, 2024), as well as BabyLM-

specific recurrent and linear-time systems such as BabyHGRN and BLaLM (Haller et al., 2024, 2025). Several of these already report matched comparisons, and Păpușoi and Nisioi (2025) audit elementary baselines under a shared evaluation pipeline. What distinguishes the present study is that each of those works holds one of our two axes fixed: GPT-BERT varies the objective inside an attention backbone, BabyHGRN and BLaLM vary the mixer under a fixed training objective, and the elementary-baseline audit varies the model family under a fixed scorer. We vary both axes jointly, over a generative-objective set that these studies predate, and we do not tune a single submission recipe.

We vary two independent choices: how the sequence mixer exchanges information among positions and what prediction problem supplies the training signal (the *objective*). Figure 1 summarizes the resulting two-by-six grid and the generation process associated with each objective.

2.1 Sequence Mixing in Language Models

Transformer attention. Softmax attention lets each token retrieve a content-weighted combination of the other token representations (Vaswani et al., 2017). We use standard multi-head self-attention: its context is causal for autoregressive training and bidirectional for all denoising objectives.

Linear attention. Instead of retaining every key-value pair, linear attention compresses the prefix into a fixed-size fast-weight state S_t , making the same computation expressible as a recurrent update (Katharopoulos et al., 2020). In its additive form, $S_t = S_{t-1} + v_t k_t^\top$, so writing a new association can overwrite an old one only indirectly. DeltaNet adds an error-correcting write,

$$S_t = S_{t-1} + \beta_t(v_t - S_{t-1}k_t)k_t^\top,$$

which changes the value stored at the addressed key rather than simply adding another value (Yang et al., 2024).

Gated DeltaNet. Gated DeltaNet additionally learns how much of the previous state to retain and how strongly to write at each step (Yang et al., 2025). OLMo Hybrid uses these layers at 7B scale in place of OLMo 3’s sliding-window attention and reports better pretraining and mid-training evaluations (Merrill et al., 2026). We use Gated DeltaNet as one representative rather than sweep efficient mixers. It runs left to right for causal objectives,

for bidirectional objectives we average forward and reverse passes. This asks whether compressed, explicitly updated memory helps when sequences are short (256 tokens) but linguistic data are scarce.

2.2 Language Modeling Objectives

Autoregression. AR turns a sequence into a fixed series of next-token decisions,

$$L_{\text{AR}} = - \sum_{i=1}^n \log p_\theta(x_i \mid x_{<i}),$$

with a beginning-of-sequence symbol before x_1 . Generation repeats exactly the operation used in training: predict one token from the prefix, append it, and continue (Radford et al., 2019). This direct match also provides an exact left-to-right sequence score.

Masked language modeling. MLM presents the model with a partly observed sequence and asks it to recover a random 15% of tokens from both left and right context (Devlin et al., 2019):

$$L_{\text{MLM}} = - \sum_{i \in M} \log p_\theta(x_i \mid x_{\setminus M}).$$

It learns one denoising step at one fixed corruption level. Although this loss does not by itself specify a full-sequence generator, it matches the masked candidate scoring used in our non-AR evaluation.

Masked discrete diffusion. Masked diffusion extends the MLM denoising problem across the whole path from a clean sequence to an almost fully masked one. At training time, a sampled timestep t determines how many tokens are replaced by an absorbing mask, the model learns the reverse direction by recovering their clean identities. At generation time it begins from masks and repeats this recovery over decreasing noise levels, so early predictions provide context for later ones. Thus MLM is a useful single-noise-level reference point, while diffusion learns the family of denoisers needed for iterative unmasking (Austin et al., 2021; Sahoo et al., 2024).

MDLM, LLaDA, and SEDD are all instances of this masked diffusion picture. In our MDLM and LLaDA cells, $t \in [\epsilon, 1 - \epsilon]$ linearly controls a mask rate from 0.02 to 0.98, and the model predicts the original tokens at masked positions. MDLM weights their cross-entropy by $1/t$ (Sahoo et al., 2024), LLaDA uses uniform weighting (Nie et al.,

2025). SEDD uses the same absorbing-mask endpoints but changes the parameterization: with log-linear noise

$$\sigma(t) = -\log(1 - (1 - \epsilon)t),$$

it predicts log score ratios and optimizes score entropy rather than ordinary token cross-entropy (Lou et al., 2024). The reported SEDD results use the sigma-scaled, mean-token-reduced implementation, earlier uncorrected runs are excluded.

Categorical flow. Flow matching learns motion along a probability path from noise to data (Lipman et al., 2023). Our CatFlow model does not jump between mask tokens. It represents each position by an unnormalized Gaussian vocabulary state z_0 and interpolates directly to the one-hot target e_x ,

$$z_t = (1 - t)z_0 + te_x.$$

After timestep-dependent RMS normalization, a learned vocabulary projection maps the continuous state into the model hidden space. The model predicts the clean endpoint with cross-entropy, generation repeatedly uses those endpoint predictions to move the state toward a token. This is the non-distilled endpoint objective underlying Categorical Flow Maps (Roos et al., 2026; Davis et al., 2026), not a finite-state Discrete Flow Matching transition model (Campbell et al., 2024; Gat et al., 2024).

Evaluation. BLiMP tests English acceptability contrasts (Warstadt et al., 2020), COMPS robust property knowledge through conceptual minimal pairs (Misra et al., 2023), and Entity Tracking whether models maintain object-location state through discourse (Kim and Schuster, 2023), fine-tuning covers GLUE and SuperGLUE-style tasks (Wang et al., 2018, 2019). A central point of our results is that these benchmarks reward different inductive biases, so a single average hides a strong opposition between broad transfer and entity-state tracking.

3 Experimental Design

3.1 Shared Model and Data

All cells share the same language-model scaffold. Tokens are represented with a byte-level BPE vocabulary of 16,384 types over sequences of length 256. The retained scale study uses learned absolute position embeddings, the corrected base comparison uses sinusoidal positions so evaluation is not

bounded by the training length. Time-conditioned objectives add a learned timestep embedding. The backbone is a stack of pre-norm sequence blocks, each consisting of a sequence mixer followed by a feed-forward network. The language-model head is tied to the input embedding matrix.

We train on the official 2026 Strict-Small and Strict corpora. The Strict-Small track uses a 10M-word budget and the Strict track uses a 100M-word budget. Both are trained for at most ten epochs, following the challenge rule. The base configuration has hidden size 512, 8 layers, 8 heads, and intermediate size 2048. The scale grid keeps the training recipe fixed and evaluates six width- and depth-scaled configurations: hidden sizes 640, 960, and 1344 at eight layers, and 13, 28, and 60 layers at hidden size 512, giving approximate 50M, 100M, and 200M parameter bands (Appendix Table 4).

The final flow comparison is parameter matched within each mixer. Its added vocabulary projection changes the hidden-state budget, so we use hidden size 448 and eight layers, with feed-forward sizes 1414 for attention and 1552 for Gated Delta. Direct counts give 35.706M and 39.951M trainable parameters, within 0.008% and 0.004% of the corresponding AR/MLM backbones (Appendix Table 5).

Learning-rate selection. Fixing one learning rate across families would confound the objective comparison with optimizer sensitivity, so we do not. Each of the twelve objective-mixer cells in the matched base comparison receives its own sweep on Strict-Small, and the selected rate is frozen for both tracks and all three final runs of that cell. Existing cells select by seven-task fine-tuning macro, the rematched CatFlow rerun selects by held-out pretraining loss over seven rates from $2 \cdot 10^{-4}$ to $2 \cdot 10^{-3}$, with both optima interior. Selected rates (Appendix Table 6) span an order of magnitude, from $1 \cdot 10^{-4}$ for SEDD with attention to $1 \cdot 10^{-3}$ for CatFlow with Gated Delta, and no two objectives share the same pair. Every family is therefore compared at its own selected rate rather than at AR’s, so the ranking in Section 4.1 is not an artifact of a schedule chosen for autoregression. Two caveats remain: the retained AR/MLM scale study keeps its original fixed schedule, so that slice compares AR against MLM at a shared rate, and the sweep varies only the peak rate, not warmup, weight decay, batch size, or objective-specific settings such as the noise distribution (see Limitations).

We repeat every cell three times. The trustworthy scale slice contains

$$2 \text{ tracks} \times 2 \text{ objectives} \times 1 \text{ mixer} \\ \times 6 \text{ size families} \times 3 \text{ runs} = 72$$

completed runs.

3.2 Evaluation

We run the public BabyLM 2026 evaluation pipeline. Zero-shot evaluation includes BLiMP, BLiMP Supplement, Entity Tracking, COMPS, and reading metrics when available. For the main paper, Zero macro averages BLiMP, Supplement, Entity, and COMPS. Fine-tuning evaluation covers BoolQ, MultiRC, RTE, WSC, MRPC, QQP, and MNLI. FT macro averages the primary metric for these seven tasks, with F1 used for MRPC and QQP and accuracy for the others.

Autoregressive models are scored with causal likelihood. Non-AR objectives are scored by placing mask tokens at candidate positions and reading the logits at those positions. This evaluation choice is necessary for compatibility, and it can change the conclusion. A generative objective can be strong under its own sampler and still weak under a masked candidate-scoring interface.

We therefore add HellaSwag, PIQA, WinoGrande, and ARC-Easy in two forms (Zellers et al., 2019; Bisk et al., 2020; Sakaguchi et al., 2020; Clark et al., 2018). The standard form ranks supplied candidates by causal conditional likelihood for AR and leave-one-out masked pseudo-likelihood otherwise. The paired generative form removes the candidates and asks each model to produce the gold continuation, following the direct-answer motivation of ARC-DA (Bhakhavatsalam et al., 2021). Candidate ranking covers each complete validation split, generation uses the same deterministic 256 examples per task for every model and reports exact match, positional token accuracy, normalized token F1, and ROUGE-L under a gold-length oracle. LAMBADA tests last-word continuation directly (Paperno et al., 2016). A matched WikiText diagnostic (Merity et al., 2017) fixes the same 16-word target span and supplies either left context alone (completion) or both left and right context (infilling). The latter deliberately favors bidirectional objectives, the contrast is meant to expose interface alignment, not to define a neutral overall ranking.

AR decodes greedily. MLM has no native sampler, it uses deterministic left-to-right mask predic-

tion, one target per pass, as an evaluation heuristic. For diffusion we vary native reverse-process length, tokens committed per model call, and commitment order (confidence, left-to-right, right-to-left, random). For CatFlow we vary 1–128 integration steps, progressive versus terminal commitment, one to 16 tokens per event, and the same four orders. We compare a fixed decoder with one selected on LAMBADA/WikiText, the latter need not generalize to every task. The sweep is not inference-compute matched: every reverse or integration step is another full model evaluation, whereas AR advances one token at a time and caches its prefix. At 128 flow steps and a mean target length of 13.8 tokens this is roughly nine full-sequence evaluations per target token. The step budgets test whether extra denoising compute repairs generation, not whether the resulting sampler is efficient.

4 Experiments and Results

4.1 What Survives the Public Scorer?

Table 1 gives the system-level view. AR/attention has the best fine-tuning macro, while local winners differ: AR/Gated Delta leads Entity, MDLM/attention leads Supplement among the denoising systems, and SEDD/Gated Delta is the strongest non-AR BLiMP cell. Corrected CatFlow is no longer near chance – its attention cell reaches 64.70 BLiMP and 59.21 Supplement – but its 61-point fine-tuning macro stays below AR’s 68.

The external interface makes the broader advantage clearer (Table 2). AR/attention leads candidate ranking and common seven-task generation. Decoder tuning raises CatFlow/attention from 6.46 to 8.05 token F1, but not to AR’s 9.06. Its WikiText completion and infilling scores exceed AR, while its 3.52 LAMBADA score does not. Bidirectional reconstruction is therefore a real local strength, not a general replacement for causal continuation. Appendix Table 7 gives the complete official matrix for both tracks.

Individual LAMBADA cases go both ways – CatFlow exactly recovers final tokens that AR misses – but the aggregate is lopsided: AR is exact on 31 of 256 cases, versus 9 for tuned CatFlow/attention and 0 for CatFlow/Gated Delta. AR trains repeatedly on the same prefix-to-next-token decision LAMBADA evaluates, whereas denoisers spread signal across corruption levels and positions.

The ranking is also not a scale artifact. Fig-

Objective	Mixer	FT macro	BLiMP	Supp.	Entity	COMPS
AR	Attention	68.0 ±0.0	75.05±0.24	60.87 ±1.81	17.08±0.16	55.20 ±0.16
AR	Gated Delta	66.0±1.0	75.25 ±0.31	59.02±2.13	21.40 ±0.87	55.06±0.41
MLM	Attention	64.0±1.0	56.67±6.55	52.65±6.77	16.91±0.15	50.43±1.70
MLM	Gated Delta	64.0±0.0	71.33±3.57	58.99±0.94	21.23±1.03	52.89±0.43
CatFlow	Attention	61.0±1.0	64.70±0.56	59.21±0.65	20.79±0.89	52.81±0.52
CatFlow	Gated Delta	62.0±0.0	65.48±2.34	57.80±1.37	19.48±0.67	54.04±0.57
MDLM	Attention	64.0±0.0	64.44±2.37	59.53±2.33	18.07±0.40	52.62±0.35
MDLM	Gated Delta	64.0±0.0	66.69±4.33	58.76±3.44	19.66±1.26	52.50±0.47
LLaDA	Attention	64.0±1.0	63.11±2.91	58.13±1.77	18.75±1.32	52.28±0.62
LLaDA	Gated Delta	63.0±1.0	64.04±5.02	58.34±1.29	19.74±0.46	52.43±0.55
SEDD	Attention	59.0±1.0	53.13±3.14	49.82±2.30	16.67±0.41	50.64±0.67
SEDD	Gated Delta	60.0±0.0	67.40±2.75	58.05±1.08	19.56±0.37	52.57±0.57

Table 1: Corrected parameter-matched Strict results, mean and standard deviation over three seeds. Bold marks the best cell per column. Fine-tuning macro is shown as a percentage.

Objective	Mixer	Rank	Fixed	Tuned	LAMBADA	WT comp.	WT infill	Δ ctx
AR	Attention	40.24	9.06	—	12.50	8.77	8.77	0.00
AR	Gated Delta	39.70	7.34	—	5.08	7.86	7.86	0.00
MLM	Attention	38.84	2.05	—	0.39	2.67	6.08	+3.41
MLM	Gated Delta	38.34	5.47	—	0.39	7.61	8.67	+1.06
CatFlow	Attention	39.17	6.46	8.05	3.52	11.51	12.41	+0.90
CatFlow	Gated Delta	37.56	4.16	5.69	0.00	8.62	10.29	+1.67
MDLM	Attention	37.63	5.80	5.20	1.56	8.74	9.86	+1.12
MDLM	Gated Delta	38.14	3.88	4.34	0.00	6.58	8.02	+1.44
LLaDA	Attention	38.90	3.18	3.91	0.39	6.55	8.63	+2.08
LLaDA	Gated Delta	37.04	4.54	4.92	0.00	7.21	8.75	+1.54
SEDD	Attention	36.87	2.26	3.72	0.00	7.21	8.23	+1.02
SEDD	Gated Delta	38.16	3.34	5.85	0.00	8.82	9.69	+0.87

Table 2: External Strict evaluation (percent). Rank is normalized accuracy averaged over HellaSwag, PIQA, WinoGrande, and ARC-Easy. Fixed is seven-task generation token F1 under the common decoder, Tuned uses each non-AR family’s selected decoder. The WikiText columns use that selected decoder (the fixed decoder for AR/MLM). Δ ctx is the infilling minus completion difference: the same model and the same target span, with right context added. It is the one column that varies only the scoring interface, and it is positive for all ten non-AR cells and exactly zero for AR by construction. Bold marks the best value in each column.

ure 2 shows FT macro curves for the trustworthy AR/MLM attention slice. AR stays above MLM across width and depth. Every shown 50M model is within 1.2 points of the best larger member of its scaling family, and the 50M model is often best. The inconsistent gains from 100M and 200M models, especially under depth scaling, suggest that optimization and data are already binding near the bottom of this parameter range.

4.2 How Much of the Ranking Is the Interface?

Objective and scoring interface differ together between AR and non-AR cells, so the headline ranking cannot by itself attribute the gap to either. A fully crossed experiment is unavailable under a shared recipe – our AR models have no masked pseudo-likelihood head and our denoisers no exact causal factorization – but two manipulations in Table 2 vary the interface while holding the trained model fixed, and they bound how much of the ranking the interface can explain.

The first is the Δ ctx column. The WikiText diagnostic fixes the same 16-word target span and

supplies either left context alone or left and right context, nothing about the model, objective, or target changes. Adding right context raises token F1 for all ten non-AR cells, by +1.51 on average and by +3.41 for MLM with attention, and by construction leaves AR unchanged. The reordering is substantial: AR/attention falls from third of twelve to fifth, AR/Gated Delta from sixth to eleventh, and the number of non-AR cells above AR/attention doubles from two to four. A left-to-right interface therefore scores bidirectional objectives below their capability, by one to three token-F1 points.

The second is candidate ranking versus option-free generation. Asked only to compare supplied strings, the twelve cells span 36.87 to 40.24, a coefficient of variation of 0.03, asked to produce the continuation themselves, the same twelve checkpoints span 2.05 to 9.06, a coefficient of variation of 0.42. Almost all apparent separation between objectives lives in the generative interface, not the discriminative one.

Together these controls support a bounded claim rather than a strong one. Matching the interface

to the objective demonstrably lifts non-AR scores, and the choice of interface demonstrably sets how large the between-objective gap looks. But the corrections are one to three points, whereas AR leads MLM by four and SEDD by nine points of fine-tuning macro (Table 1) and still leads LAMBADA by 8.98 points over the best non-AR cell after decoder tuning. Interface mismatch is a measurable part of the gap, on this evidence it is not all of it. Settling the remainder needs each objective scored under its own native likelihood (Section 5).

4.3 What Does the Stateful Mixer Buy?

Averaging the six objectives in Table 1, replacing attention with Gated Delta changes fine-tuning macro by -0.17 points and Entity Tracking by $+2.13$. That is the trade-off in one line: at Strict scale the recurrent state buys entity-state accuracy and pays for it with a small loss of broad transfer.

Why the recurrent state suits entity tracking.

Read literally, the delta rule of Section 2.1 is a keyed overwrite: it replaces whatever value the state associates with k_t rather than adding a second association beside it, and the gate learns how much of the old state to discard. Entity Tracking is a sequence of exactly this operation – “the coin is in the cup,” then “the coin is in the box,” where the second mention must supersede the first at the same address. Attention has no overwrite primitive: both mentions stay in the key–value cache and the model must learn at read time to down-weight the stale one, a discrimination it acquires from data rather than inherits from the update rule. At a 100M-word budget that difference is worth points.

Two observations support this reading and bound it. The Entity gains concentrate in the two objectives with the sharpest per-position supervision, AR ($+4.32$) and MLM ($+4.32$), and shrink or reverse where the signal is spread over corruption levels (MDLM $+1.59$, LLaDA $+0.99$, CatFlow -1.31): a state only sometimes asked to be current is a weaker overwrite. The effect is also budget dependent. On Strict-Small (Appendix Table 7) the same averages are -0.24 on Entity and $+0.17$ on fine-tuning macro, so at 10M words the advantage disappears entirely.

What the state costs. The complementary cost is retrieval: a fixed-size state must write down in advance whatever a later token needs, whereas attention retrieves any earlier key–value pair on demand. BLiMP contrasts and sentence-pair transfer

both need that arbitrary lookup. The averages need care here. Gated Delta raises BLiMP by $+5.52$ points on average, but that figure is carried by MLM ($+14.66$) and SEDD ($+14.27$), the two cells whose attention runs have by far the largest seed variance (± 6.55 and ± 3.14), across the four objectives whose attention cell trains stably the mean effect is $+1.04$. What looks like a grammatical advantage for recurrence is largely an optimization-stability effect – Gated Delta rescues configurations in which attention fails to train reliably at this budget rather than learning grammar better where attention already works. Mixer effects thus stay objective- and benchmark-specific: SEDD with Gated Delta reaches the best corrected non-AR BLiMP score (67.40), MDLM with attention the best Supplement score (59.53), and Strict Entity spans a narrow 16.4–19.7 range across the diffusion objectives.

4.4 Why the Non-AR Results Are Still Informative

The local non-AR wins above occur on different scoring interfaces and do not combine into one consistently strong system. Prompt-conditioned generation in particular remains weaker. Across 1–128 flow steps, one to 16 committed tokens per event, four orders, and progressive or terminal commitment, the selected CatFlow decoders reach 9.15% (attention) and 6.30% (Gated Delta) token F1 over LAMBADA and the two Wiki-Text tasks on the full 256-example set, still below AR/attention’s 10.01%. Appendix Figure 4 shows the non-monotonic curves: more integration is not a sufficient repair. The corrected samples are diverse and preserve every prompt, but local English often fails to form a coherent continuation.

Masked diffusion is better and responds differently to commitment size (Figure 3). Mean normalized token F1 rises from 3.99% with 16 tokens committed per call to 5.50% with one, and top-five token concentration falls from 80.3% to 41.4%, the best cell is MDLM with attention at 64 native steps (7.73%). Exact matches nevertheless remain almost absent, so this is a real diversity and overlap improvement rather than a recovery of reliable answers.

4.5 What the Overall Scores Miss

To see behavior hidden by the overall benchmark scores, we ran a 92-example controlled reasoning probe on a representative matched Strict slice, us-

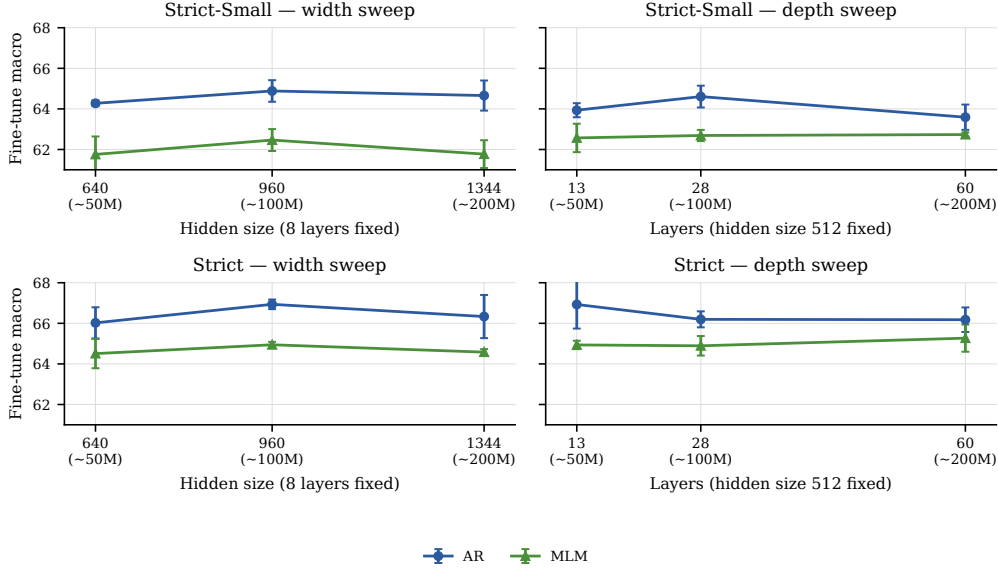


Figure 2: Fine-tuning macro for the trustworthy 72-run AR and MLM Transformer scale slice. The width panels vary hidden size over 640, 960, and 1344 while fixing depth at 8 layers, the depth panels vary 13, 28, and 60 layers while fixing hidden size at 512. Parenthetical labels give the approximate parameter band, error bars show standard deviation over three runs. Performance largely flattens after the 50M family.

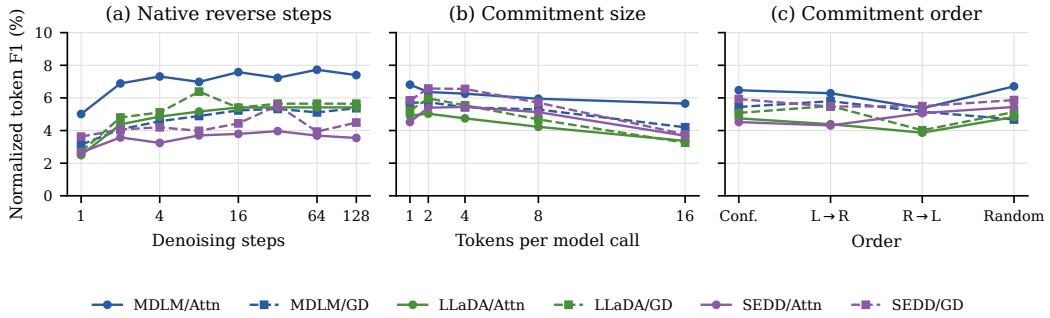


Figure 3: Masked-diffusion prompt-conditioned generation under decoder controls. Normalized token F1 is averaged over LAMBADA and paired WikiText completion and infilling. Panel (a) uses each objective’s native reverse process, panels (b) and (c) average the fixed-commit grid over the other displayed controls. More native steps help initially, smaller commitments improve overlap on average, while the best order depends on the objective and sequence mixer.

ing forced-choice items rather than task fine-tuning. A model is correct when the gold continuation has higher mean log probability than the foil, under causal likelihood for AR and masked pseudo-likelihood otherwise. Table 3 pairs probe with public scores for the same checkpoints. Appendix Table 8 summarizes the slices, Tables 9 and 10 give the full matrix and inventory, and Appendix D.1 lists the exact items.

The probe shows why a BabyLM paper should not rely on a single average across tasks. MDLM Gated Delta is strongest on counterfactual state rollback and causal counterfactuals, and CatFlow Gated Delta reaches 62.5% on both counterfac-

tual state and comparison cases even though its public fine-tuning score trails AR. The likely explanation is that the probe uses tightly matched forced choices, while the official suite additionally demands calibration, task transfer, and generation-compatible scoring.

Individual pairs also show how brittle an objective-level explanation would be. For “cutting the wire makes the alarm ring,” AR with attention prefers the correct *stay silent* once the cut is removed, whereas five of six diffusion cells prefer *ring*. On the parallel item “adding the clear drop makes the liquid turn red,” the same AR model prefers *red* after the drop is removed, while all

Model	Public FT	Public Entity	Probe all	CF state	Causal	WH	Role	NPI
MLM-T	64.0	16.91	0.576	0.438	0.625	0.750	0.700	0.833
AR-T	68.0	17.08	0.554	0.438	0.750	0.375	0.600	0.750
CatFlow-GD	62.0	19.48	0.533	0.625	0.500	0.500	0.300	0.500
MDLM-GD	64.0	19.66	0.533	0.688	0.875	1.000	0.700	0.083
LLaDA-GD	63.0	19.74	0.424	0.375	0.250	0.000	0.300	0.833

Table 3: Controlled probe scores beside corrected matched Strict results. Public columns are three-run BabyLM means, probe columns use representative matched checkpoints and report forced-choice accuracy over all 92 items or selected slices. CF state denotes counterfactual state rollback.

six diffusion cells correctly prefer *not red*. Such adjacent reversals point to lexical and calibration sensitivity rather than a reasoning ability cleanly supplied by one objective.

5 Discussion

Lesson 1: the scoring interface is an experimental variable. AR optimizes exactly the causal score BabyLM reads off, and the same model initializes sequence classifiers, bidirectional objectives reach the suite only through masked scores or fine-tuned heads. Section 4.2 shows the consequence is measurable and directional, but also that it does not explain the whole gap: the corrections we can measure are one to three points against a four- to nine-point fine-tuning deficit. Any BabyLM comparison across objective families thus reports an objective–scorer composite, and papers should say which part they intend to be reading. The control we leave undone is the fully crossed one – train each objective to admit the other family’s scorer (an any-order likelihood for denoisers, a masked pseudo-likelihood head for AR) and report the full objective $\times$ scorer matrix. That is a training-side change, which is why it does not fit inside a fixed shared recipe.

Lesson 2: state tracking is not broad language learning. Gated recurrent state fits Entity Tracking: at Strict scale Gated Delta is worth +2.13 Entity points on average and -0.17 points of fine-tuning macro, and the gain is largest exactly where supervision is sharpest per position. BLiMP and the fine-tuning suite reward retrieval of arbitrary earlier material, which a fixed-size state has to anticipate and attention does not. Statefulness does not imply broad language learning, and at the 10M-word budget it does not even buy the entity-state advantage.

Lesson 3: what a matched non-AR BabyLM recipe would look like. Weak transfer through the BabyLM interface does not rule out diffu-

sion or flow, it says the surrounding recipe is borrowed from autoregression. Four concrete changes follow. (i) *Matched scoring*: report any-order or permutation-marginal likelihoods for denoisers alongside masked pseudo-likelihood, so ranking uses the model’s own factorization, Δ_{ctx} is the cheap version of this test and it already moves results. (ii) *Adaptive commitment*: our diffusion sweep improves monotonically as tokens committed per model call fall from 16 to one (3.99% to 5.50% token F1), pointing at entropy-thresholded or semi-autoregressive block commitment rather than a fixed schedule – reported against an inference-compute-matched AR baseline, since at 128 flow steps we spend roughly nine full-sequence evaluations per emitted token. (iii) *Hybrid training*: GPT-BERT shows one model can carry both a causal scorer and a bidirectional encoder (Charpentier and Samuel, 2024), and the same construction around a diffusion denoiser would keep the bidirectional training signal while retaining a native left-to-right score. (iv) *Objective-native fine-tuning*: attaching classifiers to a masked score discards the denoiser’s conditional structure, whereas conditioning the denoiser on the label would test transfer without routing it through an AR-shaped interface. Our results do not say diffusion cannot work at BabyLM scale, they say no part of the pipeline can be inherited unexamined.

Limitations

The comparison is a shared-recipe comparison.

Holding the training stack fixed is what makes the grid interpretable, but it also means each objective is measured under a recipe whose defaults come from autoregressive practice. We tune the peak learning rate independently for all twelve objective–mixer cells (Appendix Table 6) and sweep the decoder for every non-AR family, so the two hyperparameters most likely to be mis-set are not shared. Everything else is: warmup and weight decay, batch size, the 256-token sequence length, the

ten-epoch cap, the tokenizer, the depth of time conditioning, and the masking-rate and noise schedules, which we take from each objective’s published default rather than re-tuning at BabyLM scale. Diffusion and flow methods in particular are known to benefit from objective-specific schedules and architectural adaptations. Our numbers should therefore be read as characterizing these objectives *under a shared BabyLM recipe*, not at each objective’s own optimum, and a negative result here is evidence about the transferability of a recipe as much as about an objective.

The scorer control is partial. Section 4.2 varies the interface while holding the model fixed, but we do not report the fully crossed objective $\times$ scorer matrix, which would require training each family to admit the other’s likelihood. Claims about how much of AR’s advantage is interface rather than objective are correspondingly bounded rather than settled.

The probe is small and not frequency matched.

The controlled probe contains 92 hand-written forced-choice items, and its smallest slices contain eight. Its lexical material is not matched for frequency against the training corpora, so a preference could reflect familiarity rather than the structure the item targets. Adjacent items reverse (Section 4.5), which is what one expects at this sample size. The probe supports item-level diagnosis of where an inductive bias appears to help, it does not support rankings, and it does not license claims about what the public benchmarks over- or under-represent. The apparent non-AR advantage on NPI licensing rests on 12 items and motivates a larger frequency-controlled study rather than a conclusion.

The generative tasks are deliberately stringent.

The option-free counterparts to multiple-choice tasks are diagnostics, not replacement leaderboards. They supply one reference continuation even when other answers would be valid, and their gold-length oracle removes length prediction from the problem. They isolate content generation and interface alignment rather than measure unconstrained language quality. The decoder sweep is likewise not inference-compute matched, so it tests whether extra denoising compute repairs generation, not whether the resulting sampler is practical.

Scope. We report public rather than hidden BabyLM tasks, English rather than the 2026 multi-

lingual track, and two mixers rather than a survey of efficient architectures. Gated DeltaNet is one representative of the delta-rule family and results here should not be read as covering linear-attention mixers generally.

6 Conclusion

Autoregressive Transformers transfer best throughout the retained scale sweep, already near saturation at 50M parameters. Matched-base experiments nevertheless find local non-AR wins on official metrics, bidirectional infilling, and controlled probes, and varying only the interface shows that part of AR’s margin is the scorer, not the objective – though not all of it. Those wins do not combine, every tested diffusion or flow decoder stays behind AR on prompt-conditioned generation. Small-data progress requires testing objectives, mixers, decoders, and scorers together.

Acknowledgments

We thank all reviewers for their valuable comments. Alan Akbik and Patrick Haller are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Emmy Noether grant "Eidetic Representations of Natural Language" (project number 448414230). Alan Akbik is furthermore supported under Germany’s Excellence Strategy "Science of Intelligence" (EXC 2002/1, project number 390523135). Jonas Golde is supported by the German Federal Ministry of Economic Affairs and Climate Action (BMWK) as part of the project ENA (KK5148001LB0). The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) under the NHR project "Pre-Training SLMs via Knowledge Distillation". NHR funding is provided by federal and Bavarian state authorities.

References

- Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993.
- BabyLM Organizers. 2026. BabyLM challenge: Sample-efficient pretraining on a developmentally

- plausible corpus. <https://babylm.github.io/>. Accessed 2026-07-01.
- Sumithra Bhakthavatsalam, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, and Peter Clark. 2021. [Think you have solved direct-answer question answering? try ARC-DA, the direct-answer AI2 reasoning challenge](#). *Preprint*, arXiv:2102.03315.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. [PIQA: Reasoning about physical commonsense in natural language](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439.
- Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. 2024. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 5453–5512. PMLR.
- 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.
- 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.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try ARC, the AI2 reasoning challenge](#). *Preprint*, arXiv:1803.05457.
- Oscar Davis, Anastasiia Filippova, Pierre Ablin, Victor Turrisi, Amitis Shidani, Marco Cuturi, and Louis Béthune. 2026. [Scaling categorical flow maps](#). *Preprint*, arXiv:2605.07820.
- 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. 2024. [Discrete flow matching](#). *Advances in Neural Information Processing Systems*, 37:133345–133385.
- Patrick Haller, Jonas Golde, and Alan Akbik. 2024. [BabyHGRN: Exploring RNNs for sample-efficient language modeling](#). In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 82–94, Miami, FL, USA. Association for Computational Linguistics.
- 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.
- Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. 2020. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5156–5165. PMLR.
- Najoung Kim and Sebastian Schuster. 2023. [Entity tracking in language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.
- Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*.
- Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32819–32848. PMLR.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *The Fifth International Conference on Learning Representations*.
- William Merrill, Yanhong Li, Tyler Romero, Anej Svete, Caia Costello, Pradeep Dasigi, Dirk Groeneveld, David Heineman, Bailey Kuehl, Nathan Lambert, Jacob Morrison, Luca Soldaini, Finbarr Timbers, Pete Walsh, Noah A. Smith, Hannaneh Hajishirzi, and Ashish Sabharwal. 2026. [OLMo hybrid: From theory to practice and back](#). *Preprint*, arXiv:2604.03444.
- Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. [COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.
- Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong

- Wen, and Chongxuan Li. 2025. Large language diffusion models. *Advances in Neural Information Processing Systems*, 38:50608–50646.
- Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. [The LAMBADA dataset: Word prediction requiring a broad discourse context](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.
- Rareş Păpuşoi and Sergiu Nisioi. 2025. [A comparison of elementary baselines for BabyLM](#). In *Proceedings of the First BabyLM Workshop*, pages 218–225, Suzhou, China. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Daan Roos, Oscar Davis, Floor Eijkelboom, Michael Bronstein, Max Welling, İsmail Ilkan Ceylan, Luca Ambrogioni, and Jan-Willem van de Meent. 2026. [Categorical flow maps](#). In *Forty-third International Conference on Machine Learning*.
- Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. [Simple and effective masked diffusion language models](#). *Advances in Neural Information Processing Systems*, 37:130136–130184.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. [WinoGrande: An adversarial winograd schema challenge at scale](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008.
- 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. *Advances in Neural Information Processing Systems*, 32:3261–3275.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- 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-hananey, 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.
- Songlin Yang, Jan Kautz, and Ali Hatamizadeh. 2025. Gated delta networks: Improving Mamba2 with delta rule. In *The Thirteenth International Conference on Learning Representations*.
- Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. 2024. [Parallelizing linear transformers with the delta rule over sequence length](#). *Advances in Neural Information Processing Systems*, 37:115491–115522.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Configuration Details

Table 4 gives the controlled scale families used in the retained sweep and Table 5 the parameter matching for the corrected categorical-flow comparison.

B Selected Learning Rates

Table 6 lists the peak learning rate selected for each of the twelve objective–mixer cells in the matched base comparison. Each cell was swept independently on Strict-Small, the selected value was then frozen for both tracks and all three seeds of that cell. No two objectives share the same pair of rates, and the selected values span an order of magnitude.

C Objective and Mixer Interactions

Table 7 complements the Strict matrix in Table 1 with the corrected Strict-Small results.

D Controlled Probe Details

Table 8 summarizes the composition of the probe and Table 9 reports the full 12-model by 9-slice probe matrix. The main paper reports a selected subset of these columns to focus on the contrast between state tracking and broad transfer.

Table 10 gives the item inventory used to build the 92-case probe.

D.1 Probe Items

Two negation-scope items use commas here in place of the semicolons used by the scoring file.

1. NPI 1. Gold: Only the nurse has ever touched the red folder. Foil: Even the nurse has ever touched the red folder.
2. NPI 2. Gold: Only the pilot had ever opened the silver hatch. Foil: Even the pilot had ever opened the silver hatch.
3. NPI 3. Gold: No student has ever entered the blue room. Foil: Every student has ever entered the blue room.
4. NPI 4. Gold: No visitor has ever borrowed the green badge. Foil: Every visitor has ever borrowed the green badge.
5. NPI 5. Gold: Hardly any baker has ever used the cracked bowl. Foil: Many bakers have ever used the cracked bowl.
6. NPI 6. Gold: Hardly any dancer had ever missed the final cue. Foil: Many dancers had ever missed the final cue.
7. NPI 7. Gold: Did the artist ever paint the mural? Foil: The artist ever painted the mural.
8. NPI 8. Gold: Had the captain ever moved the compass? Foil: The captain had ever moved the compass.
9. NPI 9. Gold: The guard did not give anyone the key. Foil: The guard gave anyone the key.
10. NPI 10. Gold: The child did not see anything in the drawer. Foil: The child saw anything in the drawer.
11. NPI 11. Gold: Few reporters have ever met the senator. Foil: Most reporters have ever met the senator.
12. NPI 12. Gold: Only Mira would ever question the rule. Foil: Even Mira would ever question the rule.
13. WH 13. Gold: The detective revealed who the witness had accused. Foil: The detective revealed that the witness had accused.
14. WH 14. Gold: Maya remembered what the teacher had placed on the shelf. Foil: Maya remembered that the teacher had placed on the shelf.
15. WH 15. Gold: The clerk forgot which package the courier had delivered. Foil: The clerk forgot that package the courier had delivered.
16. WH 16. Gold: The singer announced which song the choir would perform. Foil: The singer announced that song the choir would perform.
17. WH 17. Gold: The judge asked who the lawyer had called. Foil: The judge asked that the lawyer had called.
18. WH 18. Gold: The parent knew what the toddler had hidden under the bed. Foil: The parent knew that the toddler had hidden under the bed.
19. WH 19. Gold: Which book did the editor praise? Foil: That book did the editor praise?
20. WH 20. Gold: Who did the reporter interview after lunch? Foil: That did the reporter interview after lunch?
21. State 21. Context: At first, the coin was in the cup. Nora moved the coin from the cup to the box. Question: Where is the coin now? Answer: Gold: the box. Foil: the cup.
22. State 22. Context: At first, the key was in the drawer. Liam moved the key from the drawer to the jar. Question: Where is the key now? Answer: Gold: the jar. Foil: the drawer.
23. State 23. Context: At first, the ticket was in the wallet. Ava moved the ticket from the wallet to the envelope. Question: Where is the ticket now? Answer: Gold: the envelope. Foil: the wallet.
24. State 24. Context: At first, the marble was in the basket. Omar moved the marble from the basket to the closet. Question: Where is the marble now? Answer: Gold: the closet. Foil: the basket.
25. State 25. Context: At first, the ring was in the shelf. Mina moved the ring from the shelf to the pouch. Question: Where is the ring now? Answer: Gold: the pouch. Foil: the shelf.
26. State 26. Context: At first, the button was in the tray. Ezra moved the button from the tray to the cabinet. Question: Where is the button now? Answer: Gold: the cabinet. Foil: the tray.

Scaling family	Configurations varied	Architectural coordinate held fixed
Width	hidden sizes 640 ($\sim 50M$), 960 ($\sim 100M$), 1344 ($\sim 200M$)	8 layers
Depth	13 ($\sim 50M$), 28 ($\sim 100M$), 60 ($\sim 200M$) layers	hidden size 512

Table 4: Controlled scale families, parenthetical labels are approximate parameter bands.

Mixer	Hidden	Layers	FFN	Flow params	Reference params
Attention	448	8	1414	35,705,712	35,708,416
Gated Delta	448	8	1552	39,951,072	39,952,768

Table 5: Parameter matching for the corrected categorical-flow comparison. References are the corresponding AR/MLM backbones, differences are 0.0076% for attention and 0.0042% for Gated Delta.

Objective	Attention	Gated Delta
AR	5	3
MLM	3	3
CatFlow	7	10
MDLM	3	2
LLaDA	5	3
SEDD	1	3

Table 6: Selected peak learning rates for the matched base comparison, in units of 10^{-4} . Existing cells select by seven-task fine-tuning macro, the rematched CatFlow rerun selects by held-out pretraining loss.

27. State 27. Context: At first, the card was in the folder. Iris moved the card from the folder to the tin. Question: Where is the card now? Answer: Gold: the tin. Foil: the folder.
28. State 28. Context: At first, the spoon was in the sink. Theo moved the spoon from the sink to the drawer. Question: Where is the spoon now? Answer: Gold: the drawer. Foil: the sink.
29. State 29. Context: At first, the map was in the desk. Jules moved the map from the desk to the box. Jules then moved the map from the box to the shelf. Question: Where is the map now? Answer: Gold: the shelf. Foil: the box.
30. State 30. Context: At first, the photo was in the album. Lea moved the photo from the album to the drawer. Lea then moved the photo from the drawer to the bag. Question: Where is the photo now? Answer: Gold: the bag. Foil: the drawer.
31. State 31. Context: At first, the shell was in the bowl. Noah moved the shell from the bowl to the basket. Noah then moved the shell from the basket to the cup. Question: Where is the shell now? Answer: Gold: the cup. Foil: the basket.
32. State 32. Context: At first, the note was in the book. Rosa moved the note from the book to the tin. Rosa then moved the note from the tin to the pocket. Question: Where is the note now? Answer: Gold: the pocket. Foil: the tin.
33. CF state 33. Context: At first, the map was in the desk. Jules moved the map from the desk to the box. Jules then moved the map from the box to the shelf. Question: If the first move had not happened, where would the map be? Answer: Gold: the desk. Foil: the shelf.
34. CF state 34. Context: At first, the photo was in the album. Lea moved the photo from the album to the drawer. Lea then moved the photo from the drawer to the bag. Question: If the first move had not happened, where would the photo be? Answer: Gold: the album. Foil: the bag.
35. CF state 35. Context: At first, the shell was in the bowl. Noah moved the shell from the bowl to the basket. Noah then moved the shell from the basket to the cup. Question: If the first move had not happened, where would the shell be? Answer: Gold: the bowl. Foil: the cup.
36. CF state 36. Context: At first, the note was in the book. Rosa moved the note from the book to the tin. Rosa then moved the note from the tin to the pocket. Question: If the first move had not happened, where would the note be? Answer: Gold: the book. Foil: the pocket.
37. CF state 37. Context: At first, the map was in the desk. Jules moved the map from the desk to the box. Jules then moved the map from the box to the shelf. Question: If the second move had not happened, where would the map be? Answer: Gold: the box. Foil: the shelf.
38. CF state 38. Context: At first, the photo was in the album. Lea moved the photo from the album to the drawer. Lea then moved the photo from the drawer to the bag. Question: If the second move had not happened, where would the photo be? Answer: Gold: the drawer. Foil: the bag.
39. CF state 39. Context: At first, the shell was in the bowl. Noah moved the shell from the bowl to the basket. Noah then moved the shell from the basket to the cup. Question: If the second move had not happened, where would the shell be? Answer: Gold: the basket. Foil: the cup.
40. CF state 40. Context: At first, the note was in the book. Rosa moved the note from the book to the tin. Rosa then moved the note from the tin to the pocket. Question: If the second move had not happened, where would the note be? Answer: Gold: the tin. Foil: the pocket.
41. CF state 41. Context: At first, the coin was in the cup. Nora moved the coin from the cup to the box. Question: If that move had not happened, where would the coin be? Answer: Gold: the cup. Foil: the box.
42. CF state 42. Context: At first, the key was in the drawer. Liam moved the key from the drawer to the jar. Question: If that move had not happened, where would the key be? Answer: Gold: the drawer. Foil: the jar.

Objective	Mixer	FT macro	BLiMP	Supp.	Entity	COMPS
AR	Attention	63.0±0.0	64.33±0.29	55.76±0.39	18.31±0.53	50.73±0.38
AR	Gated Delta	61.0±1.0	64.27±0.27	53.19±2.35	17.71±0.24	50.27±0.07
MLM	Attention	59.0±1.0	54.04±1.72	46.99±1.96	16.45±0.32	50.21±0.14
MLM	Gated Delta	61.0±1.0	54.10±0.52	51.10±2.71	16.77±0.38	49.63±0.40
CatFlow	Attention	61.0±0.0	55.82±0.49	48.12±1.20	17.20±0.39	49.80±0.23
CatFlow	Gated Delta	62.0±0.0	57.11±1.55	51.59±0.88	17.50±0.58	50.28±0.43
MDLM	Attention	59.0±1.0	54.32±0.09	45.86±1.69	17.22±0.60	49.87±0.06
MDLM	Gated Delta	59.0±1.0	55.14±1.13	46.10±0.74	16.40±0.29	49.83±0.21
LLaDA	Attention	59.0±1.0	55.12±1.00	45.59±1.39	17.26±0.16	50.11±0.47
LLaDA	Gated Delta	60.0±1.0	54.13±1.18	46.19±2.04	17.17±0.38	49.97±0.49
SEDD	Attention	59.0±1.0	48.34±3.93	48.60±1.78	17.15±0.27	50.43±0.18
SEDD	Gated Delta	58.0±1.0	54.86±1.80	45.47±0.72	16.62±0.34	50.09±0.40

Table 7: Corrected parameter-matched Strict-Small results, mean and standard deviation over three seeds. Fine-tuning macro is shown as a percentage.

Slice	<i>n</i>	What it tests	Example gold / foil contrast
NPI licensing	12	Licensing of negative-polarity items such as <i>ever</i> and <i>anything</i> .	Only Mira would ever question the rule / Even Mira would ever question the rule.
WH filler-gap	8	Preference for a wh filler that licenses a gap over a local declarative foil.	who the witness had accused / that the witness had accused.
Ordinary state	12	Tracking an object’s actual location after one or two observed moves.	final location / previous location.
Counterfactual state	16	Rolling back an omitted move while ignoring the observed final location.	drawer / bag, when the second move to the bag is counterfactually removed.
Causal counterfactual	8	Removing an observed effect when its cause is counterfactually absent.	invalid / valid, when stamping a card is counterfactually removed.
Comparison	8	Transitive ordering and reversed-order counterfactuals.	Ava / Cara, after comparing a three-person height order.
Negation scope	8	Resolving exceptives and contrastive negation against local distractors.	the bear / the train, after all toys except the bear are packed.
Semantic role	10	Binding agents, patients, and recipients across active and passive forms.	the student / the teacher, after the teacher gives the student a pencil.
Novel binding	10	Binding novel words to newly introduced properties and negated properties.	the dax / the wug, after dax is defined as the tool that cuts paper.

Table 8: Composition of the controlled probe. Each item is a matched forced-choice comparison scored by preference for the gold continuation over the foil. Examples are abridged only for table length.

Model	All	State	CF state	Causal	Comp.	Neg.	NPI	WH	Role	Novel
AR-Attn	55.4	41.7	43.8	75.0	37.5	75.0	75.0	37.5	60.0	60.0
AR-GD	46.7	41.7	56.2	0.0	25.0	62.5	58.3	75.0	50.0	40.0
MLM-Attn	57.6	58.3	43.8	62.5	50.0	62.5	83.3	75.0	70.0	20.0
MLM-GD	44.6	50.0	37.5	25.0	37.5	62.5	25.0	25.0	70.0	70.0
SEDD-Attn	44.6	41.7	50.0	25.0	50.0	62.5	41.7	12.5	50.0	60.0
SEDD-GD	46.7	50.0	50.0	25.0	62.5	50.0	58.3	25.0	50.0	40.0
LLaDA-Attn	40.2	50.0	56.2	25.0	50.0	62.5	25.0	0.0	20.0	60.0
LLaDA-GD	42.4	41.7	37.5	25.0	50.0	62.5	83.3	0.0	30.0	40.0
MDLM-Attn	42.4	41.7	62.5	25.0	50.0	75.0	41.7	0.0	20.0	50.0
MDLM-GD	53.3	25.0	68.8	87.5	37.5	25.0	8.3	100.0	70.0	70.0
CatFlow-Attn	52.2	41.7	62.5	75.0	37.5	62.5	75.0	25.0	40.0	40.0
CatFlow-GD	53.3	50.0	62.5	50.0	62.5	62.5	50.0	50.0	30.0	60.0

Table 9: Full controlled-probe matrix. Values are forced-choice accuracies in percent for representative matched Strict checkpoints. State is ordinary state tracking, CF state is counterfactual state rollback, Novel is novel-word compositional binding.

43. CF state 43. Context: At first, the ticket was in the wallet. Ava moved the ticket from the wallet to the envelope. Question: If that move had not happened, where would the ticket be? Answer: Gold: the wallet. Foil: the envelope.
44. CF state 44. Context: At first, the marble was in the basket. Omar moved the marble from the basket to the closet. Question: If that move had not happened, where would the marble be? Answer: Gold: the basket. Foil: the closet.
45. CF state 45. Context: At first, the ring was in the shelf. Mina moved the ring from the shelf to the pouch. Question: If that move had not happened, where would the ring be? Answer: Gold: the shelf. Foil: the pouch.
46. CF state 46. Context: At first, the button was in the tray. Ezra moved the button from the tray to the cabinet. Question: If that move had not happened, where would the button be? Answer: Gold: the tray. Foil: the cabinet.
47. CF state 47. Context: At first, the card was in the folder. Iris moved the card from the folder to the tin. Question: If that move had not happened, where would the card be? Answer: Gold: the folder. Foil: the tin.
48. CF state 48. Context: At first, the spoon was in the sink. Theo moved the spoon from the sink to the drawer.

Slice	Subcategory	<i>n</i>	Construction
NPI licensing	licensor present	12	Licensed <i>ever</i> , <i>anyone</i> , or <i>anything</i> under negation, questions, <i>only</i> , <i>few</i> , or related downward-entailing contexts, paired with an unlicensed foil.
WH filler-gap	wh versus that	8	Wh fillers such as <i>who</i> , <i>what</i> , and <i>which</i> license a gap, while the foil replaces the filler with a local declarative form.
Ordinary state	one move	8	An object starts in one container and is moved once. The gold answer is the post-move location.
Ordinary state	two moves	4	An object is moved twice. The gold answer is the final location rather than the intermediate location.
Counterfactual state	undo single move	8	A one-step move is counterfactually removed. The gold answer is the initial location.
Counterfactual state	undo second move	4	A two-step story removes only the second move. The gold answer is the intermediate location.
Counterfactual state	undo first move	4	A two-step story removes the first move. The gold answer is the initial location.
Causal counterfactual	prevent effect	8	A local rule maps an observed cause to an observed effect. The counterfactual question removes the cause and the gold answer removes the effect.
Comparison	transitive order	6	Two pairwise comparisons define a three-item order. The question asks for an endpoint of the order.
Comparison	reversed order	2	The same ordering format asks what would happen if the order were reversed.
Negation scope	exception or contrast	8	Exceptives and contrastive negation override a nearby distractor answer.
Semantic role	agent-patient binding	10	Active, passive, and ditransitive sentences test whether the model binds agents, patients, and recipients.
Novel binding	positive property	7	Novel words are assigned new properties. The question asks which novel item has a target property.
Novel binding	negated property	3	Novel words are assigned contrastive positive and negated properties. The question asks for the item satisfying the negated relation.

Table 10: Full item inventory for the controlled probe. Counts sum to 92 matched forced-choice examples.

- Question: If that move had not happened, where would the spoon be? Answer: Gold: the sink. Foil: the drawer.
49. Causal 49. Context: In this game, pressing the blue switch opens the gate. Sam pressed the blue switch. Question: If Sam had not pressed the switch, what would happen to the gate? Answer: Gold: it would stay closed. Foil: it would open.
50. Causal 50. Context: In this lab, adding the clear drop makes the liquid turn red. Dana added the clear drop. Question: If Dana had not added the drop, what color would the liquid be? Answer: Gold: not red. Foil: red.
51. Causal 51. Context: In this machine, cutting the wire makes the alarm ring. The worker cut the wire. Question: If the worker had not cut the wire, what would the alarm do? Answer: Gold: stay silent. Foil: ring.
52. Causal 52. Context: In this game, placing the token on star makes the door lock. Priya placed the token on star. Question: If Priya had not placed the token on star, what would the door do? Answer: Gold: remain unlocked. Foil: lock.
53. Causal 53. Context: In this kitchen, turning the knob starts the fan. Alex turned the knob. Question: If Alex had not turned the knob, what would the fan do? Answer: Gold: stay off. Foil: start.
54. Causal 54. Context: In this puzzle, pulling the rope lowers the bridge. Kim pulled the rope. Question: If Kim had not pulled the rope, what would the bridge do? Answer: Gold: stay raised. Foil: lower.
55. Causal 55. Context: In this rule system, stamping the card marks it valid. The clerk stamped the card. Question: If the clerk had not stamped it, how would the card be marked? Answer: Gold: invalid. Foil: valid.
56. Causal 56. Context: In this device, tapping the screen saves the file. The user tapped the screen. Question: If the user had not tapped the screen, what would happen to the file? Answer: Gold: it would not be saved. Foil: it would be saved.
57. Comp 57. Context: Ava is taller than Ben. Ben is taller than Cara. Question: If the order were reversed, who would be tallest? Answer: Gold: Cara. Foil: Ava.
58. Comp 58. Context: Milo arrived before Vera. Vera arrived before June. Question: If the arrival order were reversed, who would arrive last? Answer: Gold: Milo. Foil: June.
59. Comp 59. Context: Ava is taller than Ben. Ben is taller than Cara. Question: Who is tallest? Answer: Gold: Ava. Foil: Cara.
60. Comp 60. Context: The red box is heavier than the green box. The green box is heavier than the blue box. Question: Which box is lightest? Answer: Gold: the blue box. Foil: the red box.
61. Comp 61. Context: Milo arrived before Vera. Vera arrived before June. Question: Who arrived last? Answer: Gold: June. Foil: Milo.
62. Comp 62. Context: The oak tree is older than the pine tree. The pine tree is older than the birch tree. Question: Which tree is oldest? Answer: Gold: the oak tree. Foil: the birch tree.
63. Comp 63. Context: The silver key is smaller than the brass key. The brass key is smaller than the black key. Question: Which key is largest? Answer: Gold: the black key. Foil: the silver key.
64. Comp 64. Context: Lena scored more points than Kai. Kai scored more points than Omar. Question: Who scored the fewest points? Answer: Gold: Omar. Foil: Lena.
65. Neg 65. Context: Nina packed every toy except the bear. Question: Which toy was not packed? Answer: Gold: the bear. Foil: the train.
66. Neg 66. Context: All cups except the red cup are clean. Question: Is the red cup clean? Answer: Gold: no. Foil: yes.

67. Neg 67. Context: The key is not in the left drawer, it is in the right drawer. Question: Where is the key? Answer: Gold: the right drawer. Foil: the left drawer.
68. Neg 68. Context: The note was not under the mat, it was behind the lamp. Question: Where was the note? Answer: Gold: behind the lamp. Foil: under the mat.
69. Neg 69. Context: Every child but Mira received a sticker. Question: Who did not receive a sticker? Answer: Gold: Mira. Foil: every child.
70. Neg 70. Context: The dog chased the cat, not the rabbit. Question: What did the dog chase? Answer: Gold: the cat. Foil: the rabbit.
71. Neg 71. Context: The red button starts the motor, but the green button does not. Question: Which button starts the motor? Answer: Gold: the red button. Foil: the green button.
72. Neg 72. Context: The square is inside the circle, not outside it. Question: Where is the square? Answer: Gold: inside the circle. Foil: outside the circle.
73. Role 73. Context: The doctor read the book to the child. Question: Who read? Answer: Gold: the doctor. Foil: the book.
74. Role 74. Context: The teacher gave the student a pencil. Question: Who received the pencil? Answer: Gold: the student. Foil: the teacher.
75. Role 75. Context: The chef was praised by the waiter. Question: Who was praised? Answer: Gold: the chef. Foil: the waiter.
76. Role 76. Context: The musician was followed by the reporter. Question: Who followed someone? Answer: Gold: the reporter. Foil: the musician.
77. Role 77. Context: The letter was written by the sailor. Question: Who wrote? Answer: Gold: the sailor. Foil: the letter.
78. Role 78. Context: The artist showed the critic a painting. Question: What was shown? Answer: Gold: a painting. Foil: the critic.
79. Role 79. Context: The nurse called the patient after dinner. Question: Who was called? Answer: Gold: the patient. Foil: the nurse.
80. Role 80. Context: The captain was warned by the guard. Question: Who gave the warning? Answer: Gold: the guard. Foil: the captain.
81. Role 81. Context: The child carried the basket to the porch. Question: What was carried? Answer: Gold: the basket. Foil: the porch.
82. Role 82. Context: The driver thanked the mechanic. Question: Who was thanked? Answer: Gold: the mechanic. Foil: the driver.
83. Novel 83. Context: A dax is a tool that cuts paper. A wug is a tool that stirs soup. Question: Which tool cuts paper? Answer: Gold: the dax. Foil: the wug.
84. Novel 84. Context: A blicket is a fruit that tastes sour. A toma is a fruit that tastes sweet. Question: Which fruit tastes sweet? Answer: Gold: the toma. Foil: the blicket.
85. Novel 85. Context: A fep is a machine that prints labels. A lorp is a machine that folds boxes. Question: Which machine folds boxes? Answer: Gold: the lorp. Foil: the fep.
86. Novel 86. Context: A zup is a bird that sings at night. A niff is a bird that sleeps at night. Question: Which bird sings at night? Answer: Gold: the zup. Foil: the niff.
87. Novel 87. Context: A mip is a stone that glows in water. A tane is a stone that floats in water. Question: Which stone floats in water? Answer: Gold: the tane. Foil: the mip.
88. Novel 88. Context: A vorn is a plant that needs shade. A pilk is a plant that needs sun. Question: Which plant needs shade? Answer: Gold: the vorn. Foil: the pilk.
89. Novel 89. Context: A blicket is sour but not sweet. A toma is sweet but not sour. Question: Which item is not sour? Answer: Gold: the toma. Foil: the blicket.
90. Novel 90. Context: A dax cuts paper but does not stir soup. A wug stirs soup but does not cut paper. Question: Which item stirs soup? Answer: Gold: the wug. Foil: the dax.
91. Novel 91. Context: A fep prints labels but does not fold boxes. A lorp folds boxes but does not print labels. Question: Which item does not print labels? Answer: Gold: the lorp. Foil: the fep.
92. Novel 92. Context: A zup sings at night but does not sleep at night. A niff sleeps at night but does not sing at night. Question: Which item does not sleep at night? Answer: Gold: the zup. Foil: the niff.

E Categorical-Flow Decoder Sweep

Figure 4 shows the full categorical-flow decoder grid summarized in Section 4.4.

F Generation Diagnostics

Table 11 reports the generated-sample diagnostics used in Section 4.4. AR models use continuation sampling. Masked and diffusion objectives use full-unmask and prompt-unmask samplers. CatFlow also uses unconditional flow sampling.

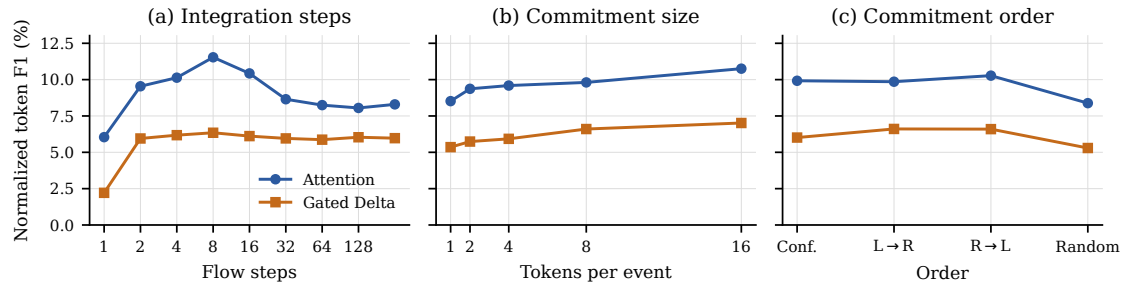


Figure 4: Categorical-flow prompt-conditioned generation under decoder controls. Normalized token F1 is averaged over LAMBADA and paired WikiText completion and infilling. Panel (a) holds progressive left-to-right commitment fixed, panels (b) and (c) average the fixed-commit grid over the other displayed controls. More integration steps, fewer committed tokens, and alternative orders do not materially repair generation for either sequence mixer.

Obj.	Mix.	Mode	Tokens	Dist-2	Rep-4
AR	GD	continuation	28.3	0.824	0.048
AR	Attn	continuation	25.3	0.807	0.071
CatFlow	GD	flow-unconditional	59.3	0.705	0.019
CatFlow	GD	full-unmask	58.8	0.717	0.014
CatFlow	GD	prompt-unmask	62.3	0.681	0.020
CatFlow	Attn	flow-unconditional	57.0	0.613	0.169
CatFlow	Attn	full-unmask	57.2	0.625	0.159
CatFlow	Attn	prompt-unmask	60.9	0.670	0.071
LLaDA	GD	full-unmask	31.0	0.121	0.696
LLaDA	GD	prompt-unmask	50.1	0.334	0.444
LLaDA	Attn	full-unmask	12.0	0.024	0.975
LLaDA	Attn	prompt-unmask	51.5	0.102	0.883
MDLM	GD	full-unmask	54.0	0.127	0.852
MDLM	GD	prompt-unmask	54.0	0.285	0.606
MDLM	Attn	full-unmask	1.0	0.000	0.000
MDLM	Attn	prompt-unmask	43.4	0.116	0.871
MLM	GD	full-unmask	32.0	0.194	0.450
MLM	GD	prompt-unmask	55.1	0.330	0.479
MLM	Attn	full-unmask	33.0	0.016	0.984
MLM	Attn	prompt-unmask	39.5	0.342	0.591
SEDD	GD	full-unmask	12.0	0.238	0.474
SEDD	GD	prompt-unmask	49.2	0.287	0.597
SEDD	Attn	full-unmask	1.0	1.000	0.000
SEDD	Attn	prompt-unmask	12.2	0.884	0.112

Table 11: Generated-sample diagnostics on representative matched Strict models. Tokens is mean decoded length over 64 samples, Dist-2 is distinct bigrams and Rep-4 is the repeated four-gram rate. Corrected CatFlow rows replace the legacy softmax-input runs.

Model	Mode	Representative output
AR-Attn	continuation	Prompt: The child found. Output: The child found it hard to stop.
AR-GD	continuation	Prompt: The child found. Output: The child found a way to get into the little girl's arms and hold her tight.
LLaDA-Attn	prompt-unmask	Prompt: A scientist explained that. Output: A scientist explained that.....
MDLM-GD	prompt-unmask	Prompt: Yesterday the teacher said. Output: Yesterday the teacher said, "I didn't know you." I didn't. "I didn't know."
MDLM-Attn	full-unmask	Output: .
MLM-Attn	prompt-unmask	Prompt: Yesterday the teacher said. Output: Yesterday the teacher said that.
SEDD-Attn	prompt-unmask	Prompt: The child found. Output: The child found.
CatFlow-GD	prompt-unmask	Prompt: A scientist explained that. Output: A scientist explained that, problem the it there may be a key Earth with medicine.
CatFlow-Attn	prompt-unmask	Prompt: In the box there was. Output: In the box there was no sign to show them the ones they had the clubs out for dinner.

Table 12: Representative generated outputs. Examples are selected from the same sample-analysis run as Table 11 and are shortened only for space.