CARRYING A LENGTH AND COUNTING IT ARE DIFFERENT ABILITIES: A DOUBLE DISSOCIATION BETWEEN RECURRENCE AND ATTENTION IN CODE-FENCE GENERATION

Anonymous authors
Paper under double-blind review

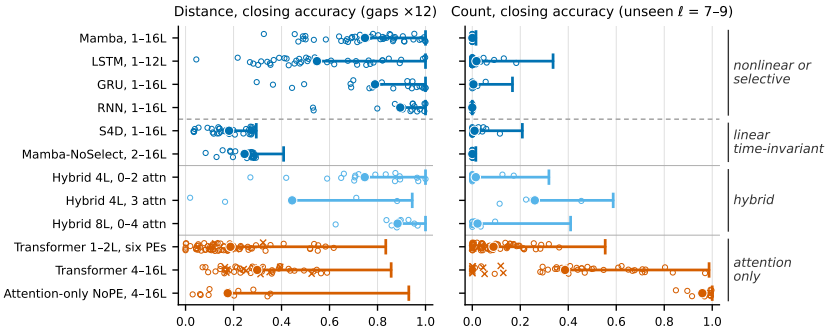


Figure 1: The distance-versus-count dissociation, 1,198 of the 1,216 gate-passing CLOSE runs in twelve families, the two linear time-invariant recurrences under a dashed rule. Left, closing accuracy on Distance (gaps $\times 12$). Right, on Count (unseen $\ell = 7$ to 9). Hollow points are cell means, crosses RoPE cells, the dot the family mean over runs, the bar end the best run.

ABSTRACT

A Markdown code fence closes at the first later run of backticks at least as long as its opener. A large language model (LLM) that wraps its answer must therefore open with a run longer than any inside, carry that length across the content and reproduce it. In LatentMD, a measurement of nine deployed Transformer LLMs, 38.0% of valid outputs were content-correct but boundary-broken, and inference alone cannot tell whether a model lacks the computation or does not use it. To isolate that computation, we built Minimal Boundary Transduction (MBT) and trained recurrent, state-space model (SSM), Transformer and hybrid architectures from scratch on its three synthetic tasks. Across 1,216 runs that pass an in-distribution gate, two extrapolation axes, distance (twelvefold gaps) and count (unseen fence lengths), split the typical solution of each family in opposite directions. At learning rates chosen on Distance, nonlinear or selective recurrence reaches closing accuracy 0.75 to 1.00 on Distance but at most 0.04 on Count. The two linear time-invariant recurrences stay below 0.30 on Distance. Deep attention stacks with learned or no positional encoding reach up to 0.99 on Count but fail on Distance once gaps stretch beyond a factor that depends on depth and encoding. RoPE stacks do not extrapolate count at any depth. In four-layer attention stacks each added recurrent layer lowers count extrapolation. Scored with the unchanged LatentMD protocol, eight SSM and hybrid LLMs show no ordering by attention or SSM share in how unclosed-fence rates grow with content length. They differ in how their unclosed fences end, which one failure rate conceals.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

A large language model (LLM) that answers in Markdown often wraps its whole answer in a fenced code block, and the answer may itself contain a code block. Under CommonMark a fence opened with a run of ℓ backticks (`) is closed by the first later line that begins with a bare run of at least ℓ backticks (MacFarlane et al., 2024). A three-backtick wrapper is therefore closed early by the three-backtick fence of the inner block. The wrapper is safe only when its run is longer than every run inside it, and a broken fence corrupts how the answer renders even when its content is correct. The same rule governs Rust raw strings, shell here-documents, PostgreSQL dollar quoting and MIME multipart boundaries (Appendix A). In each system the writer *chooses* an opener, *carries* it across content that must not collide with it, and *reproduces* it at the end.

Lee & Kang (2026) measured how often deployed models get this wrong. In its main grid of 4,179 prompts and nine language models, 38.0% of the valid outputs were content-correct but boundary-broken. All nine models were Transformers, so the measurement is silent on whether the failure depends on the architecture. The gap matters because synthetic studies have found recurrent and attention architectures to differ on two of the three requirements, *carrying* one value across irrelevant tokens and *reproducing* a count beyond the training range (Liu et al., 2023a; Sarrof et al., 2024; Chang & Bisk, 2025). Inference alone also cannot separate what a model can do from what it chooses to do. Newer models in the same study wrap on their own far less often and still fail when made to wrap, so their lower headline rates reflect behaviour and not capability (Lee & Kang, 2026). To ask which computation breaks, we train models from scratch on a task in which exposure and instruction following are under control.

We built Minimal Boundary Transduction (MBT), a suite of three synthetic tasks over abstract tokens. OPEN isolates the choice. It asks for an opening length that is safe for content the model has already seen. CLOSE isolates the carry and the reproduction. It asks a model to emit the closing run of an open fence after a stretch of filler and then to stop. VERIFY asks whether a completed block is safely closed, which tests the rule as a judgement without generation. Two evaluation-only splits stretch the two quantities that grow in a deployed answer, the content length and the outer length that must exceed every inner run. The Distance split stretches every filler gap twelvefold and the Count split asks for opening lengths that never appeared as answers in training.

Synthetic studies have tested state retention across irrelevant tokens, counting and hybrid designs in separate tasks (Liu et al., 2023a; Sarrof et al., 2024; Poli et al., 2024). MBT separates carrying a fence length over distance from producing a length outside the training range within one transduction. Across 1,216 runs that pass a learning gate, these two extrapolation axes split the typical solution of each family in opposite directions (Figure 1). Every family with a nonlinear or selective recurrent component carries the length and almost never closes a fence longer than the training maximum, and deep attention-only stacks do the reverse. We use the axes to name learned behaviour and do not claim an internal counter. Under the frozen LatentMD protocol, a held-out cohort of eight state-space model (SSM) and hybrid LLMs shows no ordering of unclosed-fence growth by attention or SSM share but large differences in the end state of those failures.

We make four contributions. **C1, the suite.** Three tasks under the CommonMark rule, six segment splits, closing accuracy on the items that require a closing run and a five-class error taxonomy aligned at family level with the LatentMD error types (Section 2). **C2, a double dissociation.** Nonlinear or selective recurrence extrapolates distance and not count. Attention-only stacks of eight or more layers do the reverse with learned or no positional encoding and not with RoPE. The split holds under learning-rate, depth and budget controls (Section 3). **C3, signatures and mechanism.** Segment splits, interventions and the attention-layer ladder place each side's fault. On the attention side the prefix collapse is a shortcut of the training distribution, the remaining limit of absolute encodings is an out-of-range position index and the site of far-distance failure depends on the positional encoding. Under-tuned recurrence loses only the record that a fence is already closed (Section 4). **C4, a reading against deployed models.** We score eight SSM and hybrid LLMs under the frozen LatentMD protocol and read them against its nine Transformers through a family-level correspondence of error types. The comparison gives a null on architecture share and a post hoc end-state observation for follow-up (Section 5).

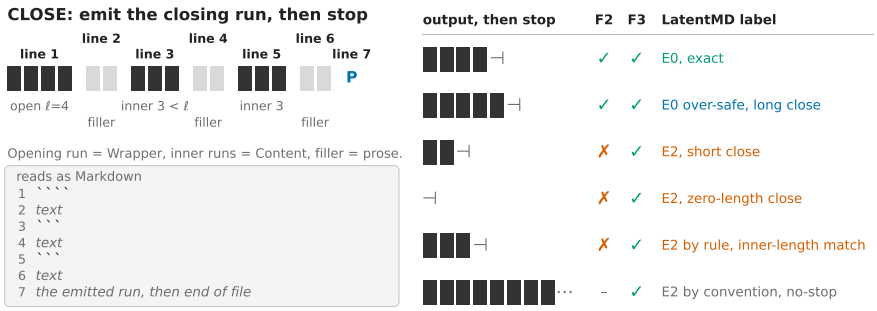


Figure 2: One worked CLOSE item ($\ell = 4$) read as a Markdown document, left, and its six output classes scored with F2 (Fence Balance) and F3 (Safe Length) and mapped to the LatentMD labels, right. Dark squares are fence tokens, light squares filler, P the probe, $\dashv$ the stop token and $\cdots$ a run that never stops. Appendix F.1 gives the same reading for VERIFY and OPEN.

2 THE MBT SUITE AND PROTOCOL

2.1 TASKS AND SPLITS

Every sequence is a stream of integer tokens. Filler tokens carry no information, and n repeated fence tokens form a run of length n. Runs are unary, so the count is explicit and tokenization is not a confound (Appendix J). The fence rule follows CommonMark. A run opens a fence when none is open. While a fence of length ℓ is open, a run of $n < \ell$ is literal content and a run of $n \geq \ell$ closes it. A run inside an open fence never opens a second one, so the task is not bracket matching. In CLOSE a fence of length ℓ is open or already closed in the input. After the PROBE token the model emits fence tokens one at a time and then STOP, or STOP alone if the fence is already closed. Figure 2 shows one such item with its output classes. Decoding is free-running and greedy, since teacher forcing would leak the count. In VERIFY the model judges a completed block safe or unsafe, safe meaning that every inner run is shorter than ℓ and the closing run $c \geq \ell$. In OPEN the model sees the inner runs, whose longest is m, and answers the opening length. The correct value is $m + 1$, and an answer of m or less is a safe-length failure.

The filler before the opening run is the prefix, and the filler between the runs and before the probe forms the gaps. Training draws ℓ from 3 to 6 and gaps of 2 to 8 tokens, with every range in Table 2, and half of the CLOSE samples are already closed. The ID split redraws this distribution with a fresh seed. The Distance split multiplies every gap by twelve, a factor fixed before training (Appendix E), so the opening-to-probe distance reaches about 157 tokens. The Count split draws ℓ from 7 to 9 with gaps in the training range. Six further evaluation-only splits stretch or pad the sequence. Prefix scales the prefix twelvefold, Inner the gaps between runs, Tail the gap before the probe, Gaps the inner and tail gaps together, which reproduces Distance, and Full all three. Pad instead front-pads the sequence to 1,000 tokens. Prefix and Pad are read against ID and the others against Distance, and Figure 5 in Appendix A draws every split on one instance. Task names are set in small capitals (CLOSE), held-out splits carry an initial capital (Distance), and the two abilities they probe are written in lowercase (the distance axis, the count axis).

2.2 METRICS, GATE AND ERROR CLASSES

Closing accuracy is exact accuracy over the items whose fence is still open at the probe, where the model must emit exactly ℓ fence tokens and then stop. The safe-close rate is the share of the same items closed with at least ℓ tokens, and answer-0 accuracy the share of already-closed items answered with STOP alone. About 42% of Distance items are already closed, so a policy that always stops at once scores about 0.42 over all items, and a rule accepting every long close would reward always emitting nine tokens and carrying nothing. Closing accuracy excludes both. Every capability sentence uses it, and answer-0 accuracy is reported separately. A run enters an out-of-distribution comparison only if its ID accuracy is at least 0.95. Wrong CLOSE outputs fall into the five classes of Figure 2 (Appendix F). Four of them leave the fence unclosed and are the counterpart of E2 (Unclosed Fence), and the fifth, the long close, is over-safe and the counterpart of E0 (Correct).

Table 1: The double dissociation in one standard cell per family, closing accuracy on the Distance split (gaps $\times 12$) and the Count split (unseen $\ell = 7$ to 9) with the safe-close rate on each, mean $\pm$ standard deviation over five seeds. Table 20 pools all 1,216 runs by family, and Figure 1 draws the 1,198 of its twelve family rows.

model	layers	lr	Distance	Count	Distance (safe)	Count (safe)
Mamba	2	10^{-3}	0.95 ± 0.07	0.00 ± 0.00	0.96 ± 0.06	0.00 ± 0.00
LSTM	2	5×10^{-3}	0.75 ± 0.11	0.00 ± 0.00	0.77 ± 0.13	0.00 ± 0.00
GRU	2	5×10^{-3}	0.99 ± 0.02	0.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00
RNN	2	10^{-3}	1.00 ± 0.01	0.00 ± 0.00	1.00 ± 0.01	0.00 ± 0.00
Hybrid MAMA	4	3×10^{-4}	0.96 ± 0.03	0.00 ± 0.00	0.97 ± 0.03	0.00 ± 0.00
Hybrid MAMAMAMA	8	10^{-3}	0.97 ± 0.02	0.04 ± 0.06	0.98 ± 0.02	0.04 ± 0.06
S4D	2	10^{-2}	0.28 ± 0.01	0.00 ± 0.00	0.28 ± 0.01	0.00 ± 0.00
Mamba-NoSelect	2	3×10^{-4}	0.19 ± 0.10	0.00 ± 0.00	0.33 ± 0.19	0.00 ± 0.00
Transformer	2	5×10^{-3}	0.20 ± 0.03	0.09 ± 0.10	0.30 ± 0.09	0.11 ± 0.13
Transformer+ALiBi	2	10^{-4}	0.62 ± 0.29	0.11 ± 0.09	0.73 ± 0.35	0.11 ± 0.09
Transformer	12	10^{-3}	0.55 ± 0.18	0.72 ± 0.10	0.95 ± 0.03	0.73 ± 0.09
Attention-only NoPE	4	10^{-4}	0.10 ± 0.10	0.92 ± 0.05	0.15 ± 0.09	0.92 ± 0.05
Attention-only NoPE	8	10^{-2}	0.35 ± 0.34	0.99 ± 0.01	0.66 ± 0.28	0.99 ± 0.00

2.3 Roster and protocol

Standard configurations have two layers and width 128. Mamba follows the published algorithm (Gu & Dao, 2024). The Transformer is decoder-only and pre-LN with learned absolute positions (Vaswani et al., 2017; Radford et al., 2019), and Transformer without a qualifier means this model. Sinusoidal, RoPE (Su et al., 2024), ALiBi (Press et al., 2022), T5-style relative bias (Raffel et al., 2020) and NoPE (Kazemnejad et al., 2023) are its positional variants. LSTM (Hochreiter & Schmidhuber, 1997), GRU (Cho et al., 2014), tanh RNN (Elman, 1990) and S4D (Gu et al., 2022) use their published forms. Mamba–attention hybrids stack Mamba blocks and NoPE attention blocks in a named pattern such as MAMA (Lieber et al., 2024). Depths run from 1 to 16 layers. Every run trains with AdamW (Loshchilov & Hutter, 2019) for 30,000 steps at batch size 128 on 3.84 million fresh samples (Appendix B). The shared learning rate (lr) is 3×10^{-4}, the anchor lr. The cell at a family's best closing accuracy on Distance in a five-point scan from 10^{-4} to 10^{-2} (Table 6) is its lr-matched cell, and Table 1 reports this cell for each family. A cell fixes architecture and training settings and averages seeds, while a family row pools cells and reflects the sweep's composition. Each cell has three or five seeds (Appendix B), and 1,216 of the 1,353 fixed-budget CLOSE runs pass the gate (pass rates by family in Table 16, intervals in Appendix K).

3 Two extrapolation axes split the architectures

Figure 1 compares Distance and Count closing accuracy for the family rows drawn from the 1,216 fixed-budget CLOSE runs that pass the gate, and Table 1 reports one standard cell per family (pooled rows and notes in Appendix D).

3.1 The dissociation

Every family with a nonlinear or selective recurrent component reaches a best closing accuracy of 1.00 on the Distance split, and its lr-matched standard cell sits at 0.75 to 1.00 (Table 1). A recurrence is nonlinear when the next state is a nonlinear function of the previous state, as in the tanh RNN, the LSTM and the GRU. It is selective when the update is linear in the state with input-dependent coefficients, as in Mamba. S4D and Mamba-NoSelect instead have fixed-coefficient, linear time-invariant state updates, the pair below the dashed rule of Figure 1. Across their 59 gate-passing cells Distance stays below 0.30 and Count below 0.5 (Tables 6 and 19). These controls therefore separate the distance score of nonlinear or selective recurrence from recurrence with fixed coefficients, while the weak count score is shared. The table shows two-layer recurrent cells, since their count axis does not move with depth (Section 3.2), the attention depths at which the count axis appears, and

the alternating hybrid stacks of the attention-layer ladder (Appendix H). On Count the lr-matched standard cells of these families stay at or below 0.04 (Table 1). Deep learned-position and NoPE stacks score high on Count and lower on Distance. The standard twelve-layer learned-position cell scores 0.55 on Distance and 0.72 on Count, and the eight-layer NoPE cell 0.35 and 0.99 (Table 1). Under the safe-close rule only the attention side's distance score rises, because most of its distance errors are long closes. Attention therefore loses the exact transduction of ℓ, and recurrence loses count under either rule.

The dissociation holds run by run outside the single-Mamba stacks (Table 45). No cell with a recurrent block reaches 0.5 on Count, the stacks with a single Mamba layer coming closest at 0.11 to 0.45 and one eight-layer hybrid seed at 0.41 (Table 19). Seventeen attention-side runs sit at or above 0.5 on both axes, and one seed of the single-Mamba AAAM stack reaches 0.59 on Count with 0.02 on Distance (Table 37). Our claim is therefore about the typical solution each family finds under this protocol and not about what each architecture can represent.

On the attention side the count axis depends on depth and on the positional encoding. Every NoPE run and 36% of the deep Transformer runs reach 0.5 on Count, against 1% of the one- and two-layer Transformer runs (Table 45). At eight or more layers learned, sinusoidal, ALiBi and NoPE stacks reach 0.40 to 0.99, T5 only at lr 10^{-4} (Table 17), RoPE stays at 0.04 or below, and no recurrent row mean exceeds 0.03 (Figure 3). Nine of the 132 attention-only cells reach 0.5 on Distance alone, eight of them ALiBi (Table 19). The distance half is a statement at the twelvefold gap scale, which we fixed before any model was trained, and the attention side's distance score falls gradually with the gap scale. At fourfold gaps the deep learned-position cells carry both axes, at 0.78 to 0.98 on distance and 0.61 to 0.84 on count. The two-layer ALiBi cell holds 0.62 at twelvefold and 0.57 at 24-fold with seeds from 0.11 to 0.83, and loses the length under a longer budget (Table 7). At 24-fold the two-layer recurrent cells fall as well, Mamba from 0.95 to 0.79 and the LSTM from 0.76 to 0.50 (Figure 7, Table 21).

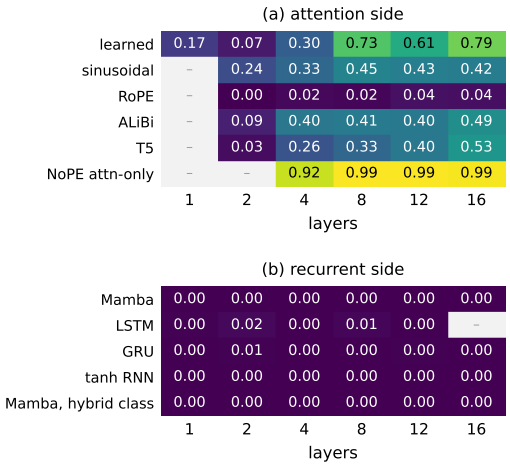


Figure 3: Count closing accuracy by depth, mean over gate-passing runs, (a) the attention side by positional encoding, (b) the recurrent side. A dash marks a depth with no gate-passing run.

3.2 Robustness

On associative recall the success of modern recurrent models is confined to a narrow window of learning rates (Okpekpe & Orvieto, 2025). Every row of Table 1 therefore has a five-rate scan, and its lr-matched cell is chosen on Distance and not on Count. Every cell with a recurrent block stays at or below 0.10 on Count at every rate of the scan (Table 6). The contrast persists across the tested learning-rate, depth, width and budget controls (Appendix C). Depth raises the count axis of the learned-position and NoPE attention stacks (Figure 3, Table 17), whereas every gate-passing Mamba, LSTM, GRU and tanh RNN cell of four to sixteen layers stays at 0.03 or below (Tables 11 and 18). The sixteen-layer LSTM never passes the gate at either of its rates (Table 16). At a tuned learning rate the four-layer Mamba keeps 0.95 on Gaps at 24-fold (Table 29).

Raising the training ceiling or training longer emissions separately does not remove the recurrent count failure. Mamba stays at 0.02 or below beyond training ceilings of 4, 5, 8 and 12, and all 18 recurrent emission-control runs stay at 0.05 or below against 0.88 for the twelve-layer Transformer (Tables 9 and 13). Parameter-count controls likewise keep the depth advantage of attention over width (Table 14). The failure concerns free generation. Recurrent cells score 0.57 to 0.82 on VERIFY, above chance at 0.5, against 0.94 to 1.00 for attention stacks of at least four layers (Table 23). On Count-Interp the LSTM reaches 0.50 and 0.78 while Mamba stays at 0.00 and 0.02 (Table 10). OPEN tests distance only, and on it the two-layer Mamba and LSTM reach 1.00 against 0.60 for the two-layer ALiBi Transformer (Table 26).

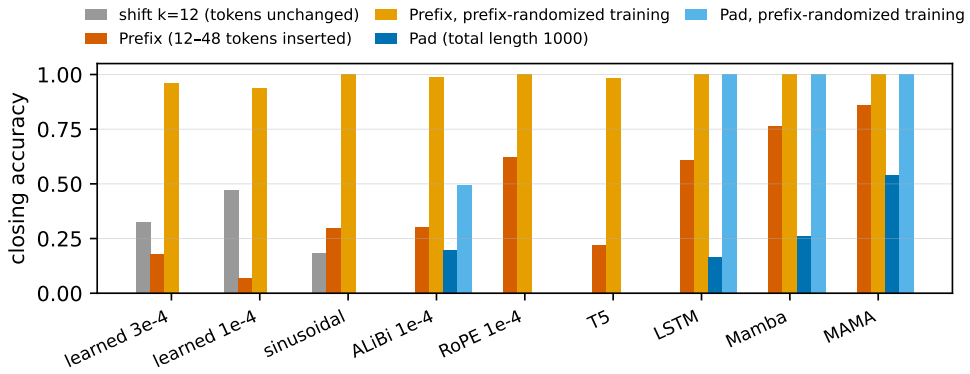


Figure 4: Closing accuracy after a position-index shift of 12 with the tokens unchanged (absolute encodings only), on Prefix and Pad, and on both after prefix-randomized training, two-layer cells and the alternating hybrid, three to five seeds. A missing bar is 0.00, and the grey bar is absent where no absolute index exists.

4 Where and why each side breaks

4.1 Segment signatures and error classes

The six segment splits separate what Distance changes at once, the gaps between runs, the gap before the probe and the whole length (Figure 8 in Appendix E). Recurrent models keep the fence length across every segment, 0.82 to 1.00 in closing accuracy on Tail and Inner at the anchor lr, and lose over a long tail gap only the record that the fence is already closed. On the already-closed items of Tail they score 0.14 to 0.72 at the anchor lr and 0.90 to 1.00 at a higher rate (Table 22, every rate in Table 6). We read the loss as an under-tuned solution rather than a limit of recurrence, with one exception, since Mamba's lr-matched rate of 10^{-3} keeps the record at 0.67. Deep learned-position Transformers show the opposite profile, Tail at 0.57 to 0.88, Inner at 0.47 to 0.63 and a collapse to 0.06 to 0.14 under twelve to forty-eight irrelevant prefix tokens, and VERIFY shows the same directions. Recurrent models never fail to stop, and their wrong closes are short, at times at the length of another run. Two-layer Transformers fail by not stopping or by emitting nothing, and deep Transformers close long in 32% to 56% of items (Table 27).

4.2 Where the attention side loses the length

A prefix of twelve to forty-eight filler tokens changes the absolute position of every later token and the number of competitors in each softmax, and a position-shift intervention separates the two. With the tokens unchanged and the position indices started at 12 instead of 0, the two-layer learned-position model drops from 0.99 to 0.32 and 0.47 at its two learning rates and the sinusoidal model to 0.18, which reproduces most of the collapse under real insertion (Figure 4). For absolute encodings the prefix collapse is therefore largely a position effect. Relative encodings are invariant to the shift, and their collapse is competition from the inserted tokens (Appendix G). Training with prefixes of 1 to 48 tokens restores Prefix in all nine arms (Table 33). We read the prefix collapse as a shortcut opened by the fixed training prefix. Pad is restored only where a recurrent component exists. The LSTM and Mamba recover to 1.00 while the absolute-position Transformers stay at 0.00, which leaves an out-of-range index as the remaining limit of absolute encodings. Restoring one site's position under the shift does not help (Appendix G).

The far-distance loss has a site as well, and it differs by positional encoding. Readout attention patching replaces the last-layer attention output at the probe and answer positions with the output from the same item at training gaps (Table 28). RoPE recovers 0.92 and 0.75 of its lost accuracy at lr 3×10^{-4} and 10^{-4}, and ALiBi at 3×10^{-4} recovers 0.69. The learned-position model recovers nothing at that site and recovers 0.64 and 0.49 from the layer-0 residual at readout. Free-running rescoring reproduces every recovery within 0.01 at two layers and 0.06 at eight (Table 32).

4.3 MIXING THE COMPONENTS IN THE ATTENTION-LAYER LADDER

Four-layer hybrids with 0, 1, 2 and 4 NoPE attention layers share one class, initialization, data and budget at the anchor lr, where pure Mamba is under-tuned. Their closing accuracy on Distance is 0.42, 0.87, 0.96 and 0.06 over five seeds, an inverted U that an eight-layer replication reproduces (Figure 12). What attention adds at that rate is distance, since Gaps rises from 0.43 to 0.97 across the first three cells while Prefix and Pad are already high in pure Mamba. The gain is a statement about the anchor rate, because the four-layer Mamba class at its own rate of 10^{-2} reaches 1.00 on Distance (Table 18). The ladder does not show that attention adds distance to a well-tuned Mamba stack (Appendix H). Count stays at zero at every rate of the pure and alternating four-layer stacks (Table 36). Extrapolation on the count axis falls in steps with the number of recurrent layers, 0.90 in the four-layer NoPE stack, 0.11 to 0.45 with one Mamba layer and 0.00 to 0.17 with two or more (Table 37). At eight layers every cell with a Mamba block has a mean Count of 0.10 or below (Tables 1 and 36). A single Mamba layer carries distance only in the lower half of the stack, at 0.88 and 0.71 against 0.16 and 0.02.

4.4 DEPTH, LAYER ROLES AND THE RECURRENT SIDE

Depth moves the attention side's distance boundary and not its coordinate dependence. For learned positions the gap scale at which Gaps closing accuracy crosses one half moves from between 4 and 6 at two layers to between 8 and 16 at twelve and sixteen layers, and for RoPE it stays between 3 and 6 at every depth (Table 29).

In the layer-ablation map (Figure 11 in Appendix G) the first attention layer of the alternating hybrid carries distance and its first two layers carry Pad. In deep Transformers the critical layers sit in the bottom half and the redundant ones in the top half (Table 34). On the recurrent side the two-layer S4D and Mamba-NoSelect cells at width 128 stay at 0.08 to 0.28 on Distance at every learning rate, with their loss on Tail, and S4D stays there with its parameters matched to Mamba. Mamba reaches 0.58 to 0.95 and the tanh RNN 0.80 to 1.00 over the same rates (Table 6). Mamba-NoSelect keeps every gate and activation of the Mamba block, with 274,368 parameters against 298,880, and removes only the input dependence of the state update. The Mamba-NoSelect contrast confirms the design principle behind selection, the filtering of irrelevant tokens (Gu & Dao, 2024). Keeping every gate does not rescue Mamba-NoSelect, and removing the convolution keeps the two-layer Mamba at 0.72 on Distance against 0.81 with it (Table 11).

5 THE BRIDGE TO REAL LLM FAILURES

5.1 CORRESPONDENCE AND THE SAFE-LENGTH CHOICE

The LatentMD error labels and our taxonomy align at the family level, in a correspondence fixed before the comparison (Table 39 and Figure 13 in Appendix J). We separate retaining the opening length, detecting a collision and choosing a safe opener, which Lee & Kang (2026) treat as one missing discipline behind both error families, and we never write a synthetic quantity as E3 itself. Asked to wrap content containing a run of three to five backticks, eight of the nine LatentMD models chose a safe opener in less than 1% of cases. RoPE is the encoding of the six open-weight Transformers among the nine. It is also the one encoding of the suite that does not extrapolate count at any depth or learning rate we trained, at 0.13 or below at eight or more layers (Figure 3, Table 17). We read this as consistent with the LatentMD result and not as its cause, since the deep RoPE cells rest on three seeds each and the cohort adds no count measurement.

5.2 THE HELD-OUT COHORT

We scored the pure state-space model Falcon-Mamba-7B and seven instruction-tuned hybrids under the frozen LatentMD protocol with its released scoring tool unchanged, decoding greedily at temperature 0 with at most 7,168 output tokens. They are the open-weight models with a recurrent component that passed exclusion rules fixed before any generation, and none is among LatentMD's nine. Appendix J lists the models, the rules and every denominator. The E2 (Unclosed Fence) rates are taken over the valid generations that open at least one fence, 79.7% of Falcon-Mamba's generations and 95% to 100% for every other model. In the suite a nonlinear or selective recurrence

carries the fence length across twelvefold gaps (Section 3), so if that result transferred, the growth of E2 (Unclosed Fence) with content length would slow as the SSM share rises. It does not. Under the frozen protocol the growth is not ordered by the SSM share or the attention share, it is heterogeneous within every size tier, and counting truncated open fences as E2 (Unclosed Fence) changes the curves but not this absence of an ordering (Figure 14 in Appendix J). Hints separate models, not architectures. Falcon-H1-34B improves from 0.96 to 0.65 under the one-shot example, while all three hints worsen Falcon-Mamba over all prompts (Table 43).

A post hoc observation gives the clearest contrast. It concerns the end state in which a generation opens the outer fence and never writes another fence line (Figure 14d, Table 42). That end state takes at most 0.20 of the unclosed fences in any of the nine LatentMD Transformers, and it is highest in three of the eight held-out models, 0.74 in Nemotron-Nano-9B-v2 ($n = 39$ unclosed fences), 0.53 in Falcon-Mamba-7B ($n = 208$) and 0.49 in Granite-4.0-H-Tiny ($n = 112$), with the other five at 0.03 to 0.23. Missing content produces this end state by itself, so we restrict to the unclosed fences whose content passed the LatentMD content check. Falcon-Mamba's share halves to 0.27 ($n = 41$), the two hybrids keep theirs at 0.81 ($n = 32$) and 0.53 ($n = 45$), and the three small Transformers stay at or below 0.11. The end state is therefore not a property of the pure SSM, and it does not order along SSM share, since Falcon-H1-7B with attention in every block stays at 0.03 and the two hybrids with attention in about one layer in ten lie beside or above the model with none. We read it as compatible with an omitted closing action, and the cohort does not separate that from a lost opening length.

5.3 What does not transfer

The suite separates the terminal-obligation axis only in part. Its nearest analogue to a fence opened and never followed by another fence line, the zero-length close, appears on both sides of the dissociation on different axes, so the suite cannot tell an omitted close from a counter that never started. A post-training account competes with the architectural one. The cohort differs in training data, instruction following and size tier, its pure SSM class holds one model, and the synthetic hybrids differ from the deployed ones in scale, training and block design (Appendix J lists five confounds). These observations do not isolate an effect of architecture.

6 Related work

State retention and state tracking. Flip-Flop language modeling isolates the retention of one bit across irrelevant tokens, and on it Transformers show a long tail of errors, the attention glitches, that an LSTM does not (Liu et al., 2023a). The state-tracking limit of Merrill et al. (2024) concerns the composition of permutations and not the retention of one bounded value. Sarrof et al. (2024) show on Mamba that star-free state tracking, Flip-Flop included, generalizes in length and that counter languages are learned in distribution with weaker length generalization. This is the closest precedent for our two axes. CLOSE varies the dependency distance and the fence length inside one transduction, adds distractor runs and a fence that can close and reopen, and scores exact emission followed by STOP. Appendix L lists the closest settings beside ours.

Counting, recognition and stopping. The nearest neighbours of CLOSE remain $a^n b^n$ and its relatives (Weiss et al., 2018; Bhattamishra et al., 2020; Suzgun et al., 2019). Weiss et al. (2018) show that the LSTM cell state can implement a counter that generalizes past the training range and that the GRU and the tanh RNN cannot. Chang & Bisk (2025) report that the Elman RNN and the LSTM count inductively beyond the training range while modern recurrent models such as Mamba fall behind. Our count axis agrees with the last point for Mamba and with Weiss et al. (2018) on the GRU and the RNN. It disagrees with Chang & Bisk (2025) on the RNN and with both papers on the LSTM, since all three close short beyond a training range of four values. Those tasks train on longer count ranges, and recognition and free generation impose different demands (Butoi et al., 2025). Newman et al. (2020) show that training to predict an end token lowers length extrapolation even when the output length is given at test time, so we do not read a short close by itself as a lost count. Délétang et al. (2023) set the from-scratch methodology that we follow, and Liu et al. (2023b) show that Transformers learn shortcuts to automata in distribution, which motivates the segment splits and the prefix-randomized arm.

Selective memory, recall and hybrids. Gu & Dao (2024) motivate selection by the filtering of irrelevant tokens and test it with selective copying. Copying and recall benchmarks test the retention of many content-dependent values. Jelassi et al. (2024) and Arora et al. (2024) attribute the recurrent gap there to the fixed-size state, and Okpekpe & Orvieto (2025) close part of that gap with a finer learning-rate scan. The Distance split instead asks for one bounded fence length across uninformative filler. The Mamba and Mamba-NoSelect contrast on Distance follows that principle, and what the suite adds is where the retained length shows, by segment and error class, and that retention does not extend to producing an unseen length (Sections 4.1 and 4.4). MAD scores hybrid designs on six synthetic capabilities and links them to scaling (Poli et al., 2024), Retrievit reports that Mamba-2 and attention hybrids outperform SSMs and match or exceed Transformers on n-gram retrieval and not on position retrieval (Pantazopoulos et al., 2026). Hartl et al. (2026) report that the nonlinear recurrent component of xLSTM gives perfect state tracking at every tested length and weaker counting extrapolation. Our hybrid results concern exact boundary generation under the tested placements and do not contradict those gains. Chen et al. (2025a) trace Mamba failures on designed synthetic tasks to its convolution, which Section 4.4 tests directly. Ben-Kish et al. (2025) report an effective receptive field bounded by the training length on a different task, whereas MBT stretches uninformative filler while the retained value stays fixed.

Positional encoding and structured output. Kazemnejad et al. (2023) and Bhattamishra et al. (2020) report that removing explicit position helps length generalization, whereas here the two-layer NoPE control never passes the gate and the NoPE attention-only stacks extrapolate count but not distance. Ruoss et al. (2023) randomize position indices in training, where our prefix arm inserts filler tokens instead and the position-shift intervention separates the index from the inserted tokens. Zhou et al. (2024) reach length generalization on counting and addition with Transformers whose success depends on the positional encoding and varies over seeds, and Anil et al. (2022) set the train-short, test-long framing. MDEval (Chen et al., 2025b) and StructEval (Yang et al., 2026) score structured output without a fence-level error typology, the E-types originate with Lee & Kang (2026), and constrained decoding (Willard & Louf, 2023; Geng et al., 2023) is the mitigation family. A decoder that knows the fence rule solves CLOSE by construction, so the suite cannot rank such mitigations.

7 LIMITATIONS, OPEN QUESTIONS AND CONCLUSION

The suite is one task family over unary runs, whereas deployed tokenizers write a fence as a short sequence of merged tokens (Table 44), so the tokenization axis is unverified. The count axis is measured under free-running generation, and in the judgment task the recurrent cells stay above chance (Section 3.2). Our count result therefore concerns this generation failure and not an absence of usable length information. Most cells have three seeds, and we report no inferential test beyond run-level intervals and one seed-level interaction test (Table 35). The two-layer mechanisms have no single site at eight layers, and the attention-layer ladder holds for one training distribution and budget. The bridge is directional evidence with the confounds of Section 5.3, and synthetic and real hybrids disagree at the opening step. The results concern the tested architectures and the solutions they learned. The synthetic Mamba follows the first published algorithm (Gu & Dao, 2024), the deployed hybrids build on it or on Mamba-2 blocks (Dao & Gu, 2024), and Mamba-3 solves parity and modular arithmetic where Mamba-2 does not (Lahoti et al., 2026). Mamba-2 and Mamba-3 are untested on MBT. Two gaps remain. A pair of models that differ only in architecture would isolate the effect of architecture, which the cohort does not, and a count measurement on deployed models would test the count half of the dissociation, which so far rests on the safe-length failure of LatentMD's nine Transformers. How the placement of a Mamba layer moves count extrapolation in these hybrids stays open.

Boundary-safe generation decomposes into two extrapolation axes, distance and count, and the typical solutions of nonlinear or selective recurrence and of attention split across them in opposite directions. Segment splits separate their error profiles, and two-layer interventions localize part of the attention side's loss. Depth or mixing moves the distance boundary, and in the tested four-layer hybrids count extrapolation falls as Mamba layers replace attention layers, with large variation across placements. The held-out cohort shows no ordering of unclosed-fence growth by attention or SSM share under either truncation rule, and a post hoc difference in the end state of unclosed fences that the aggregate rate conceals.

REPRODUCIBILITY STATEMENT

Appendix B lists the training protocol, the split definitions with their seed bases and every hyperparameter. The released repository (`https://huggingface.co/2026paper3/minimal-boundary-transduction`) contains the full cell inventory, 604 cells with seed means, standard deviations, gate counts, parameter counts and wall-clock times in one file, the per-seed values of every cell cited in the paper in a second file, together with the generators, the training loop, the test suite, the trained checkpoints, the run records and every analysis script that produced a number in this paper. The code is also browsable at `https://anonymous.4open.science/r/minimal-boundary-transduction-iclr-E180/`. Data generation is seeded and byte-identical across runs. The held-out cohort was scored with the released LatentMD scoring tool (Lee & Kang, 2026) unchanged, and its decoding settings and exclusion rules are listed in Appendix J.

AI USE STATEMENT

Following the ICLR 2027 policy, we disclose four uses of generative AI tools (Claude, Anthropic), the four items marked in the submission form. **To aid or polish writing.** The tools edited text for readability and for consistency of terms, and the authors reviewed every edited passage. **For retrieval and discovery.** The tools assisted the literature search and the collection of the delimiter systems of Appendix A, and every cited work was checked against the original. **For research ideation or execution.** Among the tasks with required disclosure, the tools gave feedback on the framing, the hypotheses and the experiment design, implemented the task generators and their tests, the models, the training loop and the evaluation and analysis scripts, cleaned and reformatted result tables, and assisted in interpreting results. They also wrote figure code. **To draft sections of the paper.** The tools drafted parts of this paper and suggested its structure. The synthetic datasets were produced by the seeded generators and not by a generative model, we have not used generative AI tools to formulate mathematical claims or to write proofs, and translation, the cleaning of external datasets and qualitative or thematic analysis are not applicable to this work. We have reviewed all AI-assisted work. The code is covered by a test suite that re-parses generated sequences against the fence rule, and every number in the paper was checked against the raw run logs and the cell inventory. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Assaf Ben-Kish, Itamar Zimerman, Shady Abu-Hussein, Nadav Cohen, Amir Globerson, Lior Wolf, and Raja Giryes. DeciMamba: Exploring the length extrapolation potential of Mamba. In *International Conference on Learning Representations (ICLR)*, 2025.

Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. On the ability and limitations of transformers to recognize formal languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7096–7116, 2020.

Aaron Blakeman, Aarti Basant, Abhinav Khattar, Adithya Renduchintala, Akhiad Bercovich, Aleksander Ficek, et al. Nemotron-H: A family of accurate and efficient hybrid Mamba-transformer models. *arXiv preprint arXiv:2504.03624*, 2025.

Tim Bray. The JavaScript object notation (JSON) data interchange format. Technical Report RFC 8259, Internet Engineering Task Force, 2017.

Alexandra Butoi, Ghazal Khalighinejad, Anej Svete, Josef Valvoda, Ryan Cotterell, and Brian DuSell. Training neural networks as recognizers of formal languages. In *International Conference on Learning Representations (ICLR)*, 2025.

Yingshan Chang and Yonatan Bisk. Language models need inductive biases to count inductively. In *International Conference on Learning Representations (ICLR)*, 2025.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. Extending context window of large language models via positional interpolation. *arXiv preprint arXiv:2306.15595*, 2023.

Tianyi Chen, Pengxiao Lin, Zhiwei Wang, and Zhi-Qin John Xu. Achilles' heel of Mamba: Essential difficulties of the Mamba architecture demonstrated by synthetic data. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Zhongpu Chen, Yinfeng Liu, Long Shi, Zhi-Jie Wang, Xingyan Chen, Yu Zhao, and Fuji Ren. MDEval: Evaluating and enhancing markdown awareness in large language models. In *Proceedings of the ACM on Web Conference 2025 (WWW '25)*, pp. 2981–2991, 2025b.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Grégoire Délétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A. Ortega. Neural networks and the Chomsky hierarchy. In *International Conference on Learning Representations (ICLR)*, 2023.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

Ned Freed and Nathaniel Borenstein. Multipurpose internet mail extensions (MIME) part two: Media types. Technical Report RFC 2046, Internet Engineering Task Force, 1996.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471, 2000.

GitHub Advisory Database. form-data uses unsafe random function in form-data for choosing boundary, GHSA-fjxv-7rqg-78g4 (CVE-2025-7783). `https://github.com/advisories/GHSA-fjxv-7rqg-78g4`, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024.

Albert Gu, Ankit Gupta, Karan Goel, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Anamaria-Roberta Hartl, Levente Zólyomi, David Stap, Pieter-Jan Hoedt, Niklas Schmidinger, Lukas Hauzenberger, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. On subquadratic architectures: From applications to principles. *arXiv preprint arXiv:2606.12364*, 2026.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

IBM Granite Team. Granite 4.0 language models. `https://github.com/ibm-granite/granite-4.0-language-models`, 2025. Model cards for granite-4.0-h-tiny and granite-4.0-h-small, released 2025-10-02.

IEEE and The Open Group. IEEE std 1003.1-2017 (POSIX.1-2017), shell command language, here-document. `https://pubs.opengroup.org/onlinepubs/9699919799/utilities/V3_chap02.html`, 2018.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Aakash Lahoti, Kevin Y. Li, Berlin Chen, Caitlin Wang, Aviv Bick, J. Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. In *International Conference on Learning Representations (ICLR)*, 2026.

Tom Lane. Re: BUG #4908: escaping and dollar quotes: "ERROR: unterminated string". pgsql-bugs mailing list, `https://www.postgresql.org/message-id/10216.1247087160@sss.pgh.pa.us`, 2009.

Sungjune Lee and Myungjoo Kang. LatentMD: Benchmarking Markdown boundary failures in LLM-generated text. *arXiv preprint arXiv:2609.06993*, 2026.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, et al. Jamba: A hybrid transformer-Mamba language model. *arXiv preprint arXiv:2403.19887*, 2024.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Exposing attention glitches with flip-flop language modeling. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *International Conference on Learning Representations (ICLR)*, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

John MacFarlane et al. CommonMark spec version 0.31.2. `https://spec.commonmark.org/0.31.2/`, 2024.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Benjamin Newman, John Hewitt, Percy Liang, and Christopher D. Manning. The EOS decision and length extrapolation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 276–291, 2020.

NVIDIA. NVIDIA Nemotron Nano 2: An accurate and efficient hybrid Mamba-transformer reasoning model. *arXiv preprint arXiv:2508.14444*, 2025.

Destiny Okpekpe and Antonio Orvieto. Revisiting associative recall in modern recurrent models. *arXiv preprint arXiv:2508.19029*, 2025.

OpenAI Developer Community. ChatGPT: Use longer Markdown fences so code blocks don't break. https://community.openai.com/t/chatgpt-use-longer-markdown-fences-so-code-blocks-don-t-break/1357890, 2025.

Georgios Pantazopoulos, Malvina Nikandrou, Ioannis Konstas, and Alessandro Suglia. Retrievit: In-context retrieval capabilities of transformers, state space models, and hybrid architectures. *Transactions on Machine Learning Research*, 2026.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Michael Poli, Armin W. Thomas, Eric Nguyen, Pragaash Ponnusamy, Björn Deiseroth, Kristian Kersting, Taiji Suzuki, Brian Hie, Stefano Ermon, Christopher Ré, Ce Zhang, and Stefano Massaroli. Mechanistic design and scaling of hybrid architectures. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations (ICLR)*, 2022.

Python Software Foundation. The Python language reference, lexical analysis: String and bytes literals. https://docs.python.org/3/reference/lexical_analysis.html, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Anian Ruoss, Grégoire Délétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. Randomized positional encodings boost length generalization of transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL, Short Papers)*, 2023.

rust-lang/rust contributors. Raw string literal escape depth is technically limited, rust-lang/rust issue #50111. https://github.com/rust-lang/rust/issues/50111, 2018.

Yash Sarrof, Yana Veitsman, and Michael Hahn. The expressive capacity of state space models: A formal language perspective. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Mirac Suzgun, Yonatan Belinkov, Stuart Shieber, and Sebastian Gehrmann. LSTM networks can perform dynamic counting. In *Proceedings of the Workshop on Deep Learning and Formal Languages: Building Bridges*, pp. 44–54, 2019.

The PostgreSQL Global Development Group. PostgreSQL documentation, lexical structure: Dollar-quoted string constants. https://www.postgresql.org/docs/current/sql-syntax-lexical.html, 2025.

The Rust Project Developers. The Rust reference, tokens: Raw string literals. https://doc.rust-lang.org/reference/tokens.html#raw-string-literals, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Gail Weiss, Yoav Goldberg, and Eran Yahav. On the practical computational power of finite precision RNNs for language recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 740–745, 2018.

Brandon T. Willard and Rémi Louf. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jialin Yang, Dongfu Jiang, Tony He, Sherman Siu, Yuxuan Zhang, Disen Liao, Zhuofeng Li, Huaye Zeng, Yiming Jia, Haozhe Wang, Benjamin Schneider, Chi Ruan, Wentao Ma, Zhiheng Lyu, Yifei Wang, Yi Lu, Quy Duc Do, Ziyan Jiang, Ping Nie, and Wenhu Chen. StructEval: Benchmarking LLMs' capabilities to generate structural outputs. *Transactions on Machine Learning Research*, 2026.

Yongchao Zhou, Uri Alon, Xinyun Chen, Xuezhi Wang, Rishabh Agarwal, and Denny Zhou. Transformers can achieve length generalization but not robustly. *arXiv preprint arXiv:2402.09371*, 2024.

Jingwei Zuo, Maksim Velikanov, Dhia Eddine Rhaiem, Ilyas Chahed, Younes Belkada, Guillaume Kunsch, and Hakim Hacid. Falcon Mamba: The first competitive attention-free 7B language model. *arXiv preprint arXiv:2410.05355*, 2024.

Jingwei Zuo, Maksim Velikanov, Ilyas Chahed, Younes Belkada, Dhia Eddine Rhayem, Guillaume Kunsch, Hakim Hacid, et al. Falcon-H1: A family of hybrid-head language models redefining efficiency and performance. *arXiv preprint arXiv:2507.22448*, 2025.

A TASKS AND GENERATOR

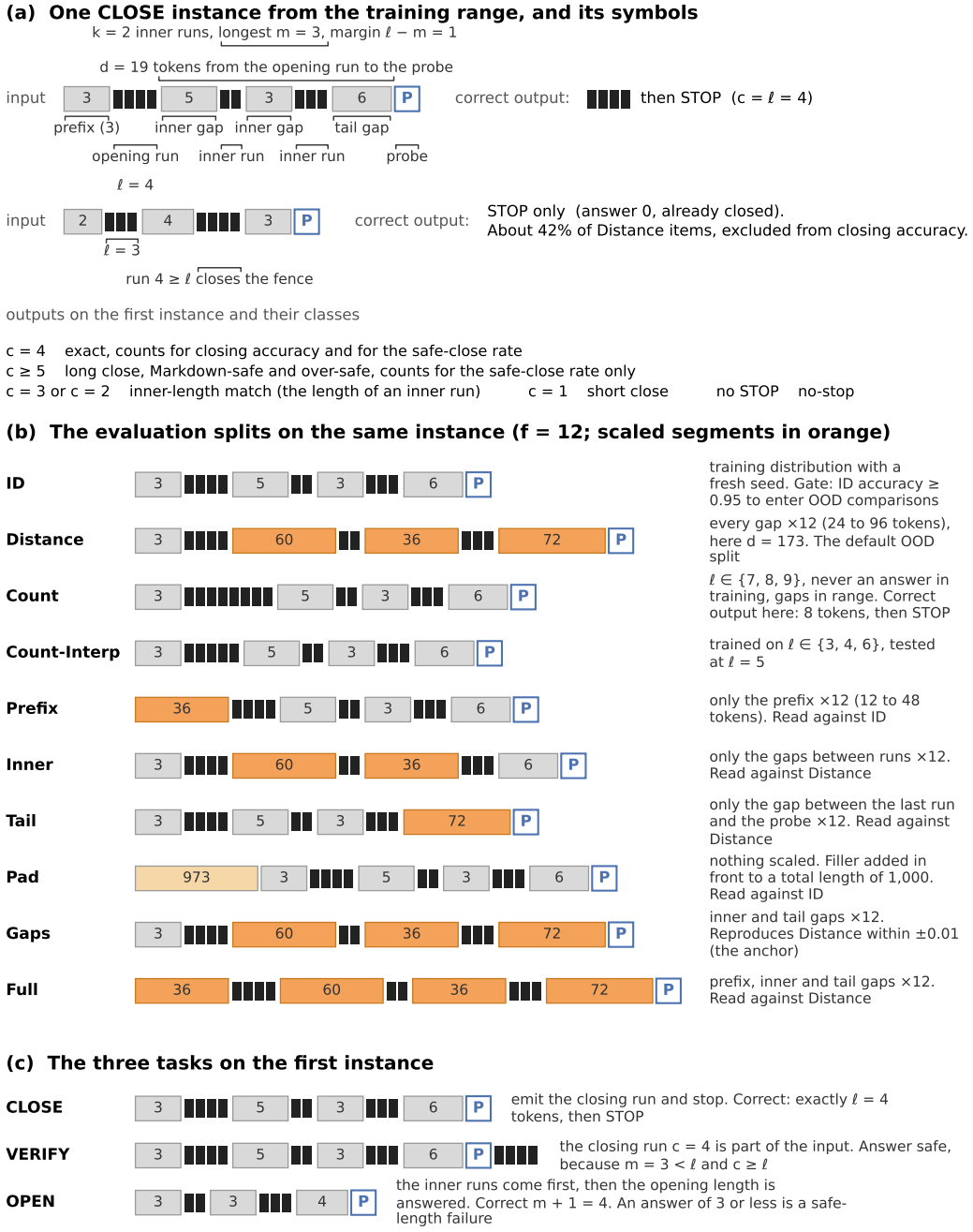


Figure 5: Every symbol and split on one worked instance. (a) A CLOSE input in the training range (prefix 3, $\ell = 4$, inner gaps 5 and 3, inner runs 2 and 3, so $m = 3$ and margin 1, tail gap 6, $d = 19$) and an input already closed by a run of 4, answer 0. (b) The ten evaluation splits built from the same instance, orange segments scaled by $f = 12$. (c) The three tasks on the first instance.

In Figure 5a the correct output of the first input is exactly ℓ fence tokens and then STOP, the second input is already closed, so its correct output is STOP alone, and the listed outputs name the error classes of Section 2.2. In panel (b) Prefix and Pad are read against ID and the other segment splits against Distance. In panel (c) VERIFY receives the closing run as input and OPEN receives the inner runs before it answers the opening length.

Fence rule. Runs of the fence token are unary, so a length is a count of identical tokens. A run of length n opens a fence when none is open. While a fence of length ℓ is open, a run of length $n < \ell$ is content and a run of length $n \geq \ell$ closes it. Runs inside an open fence never open a nested fence. The three tasks read this rule from three positions. CLOSE emits the closing run. VERIFY judges a finished block, which is safe when every inner run is shorter than ℓ and the closing run is at least ℓ. OPEN answers the smallest safe opening length $m + 1$ after content whose longest run is m.

Generator and vocabulary. Table 2 lists the generator ranges of the campaign, the token vocabulary and the seed of every split. Each split has one seeded random stream, each sequence is drawn independently from it, and every draw below is uniform over an inclusive integer range. A CLOSE sequence is drawn in the following order.

1. The opening length ℓ from 3 to 6.
2. The number of inner runs k from 0 to 2, then each inner length from 1 to $\ell - 1$. The runs keep the order drawn.
3. With probability one half the document is already closed. A closing run of length ℓ to 9 is then appended, and with probability 0.15 a fresh fence follows it, with a new opening length from 3 to 6 and zero to two inner runs shorter than it. Runs of length 7 to 9 therefore occur in CLOSE training inputs, but only as closing runs.
4. The prefix, 1 to 4 filler tokens. In the binding condition the designator token, the fence token of the item and 1 to 3 filler tokens follow the prefix.
5. The runs are written in order. Before every run after the first a gap of 2 to 8 filler tokens is drawn, after the last run a tail gap of 2 to 8, and the probe token ends the input.
6. The answer is computed by the fence rule from the run lengths, not from the generator's bookkeeping. It is the length of the open fence, or 0 when the document is closed. In the generative format the target is that many fence tokens and then STOP.

Every filler token is an independent draw from 10 to 99, so a gap of g tokens costs g draws.

A VERIFY sequence draws ℓ, a safe coin at one half and one to three inner runs of length 1 to $\ell - 1$. A safe item draws its closing run from ℓ to $\ell + 3$, capped at 9. An unsafe item takes one of its two causes at one half. Under the inner-collision cause one inner run, chosen at random, is redrawn from ℓ to 9 and the closing run is drawn as in a safe item. Under the short-close cause the closing run is drawn from 1 to $\ell - 1$. The layout is prefix, opening run, inner runs with gaps, tail gap, probe and closing run, so the closing run stands after the probe, and the label is recomputed from the fence rule and checked against the intended cause.

An OPEN sequence draws m from 1 to 6 and one to three inner runs whose longest is exactly m, the first equal to m and the others from 1 to m, in shuffled order. Its layout is prefix, designator pair, 1 to 3 filler tokens, inner runs with gaps of 2 to 8, tail gap and probe, and its answer is $m + 1$.

A segment split multiplies the range of one gap kind by f before the draw and changes nothing else, so Gaps at $f = 12$ draws every gap from 24 to 96 in the same order, which is why the generator at scale one reproduces the training generator byte for byte. The Count split replaces the range of ℓ by 7 to 9 and keeps every other range.

Table 2: Generator, vocabulary and split seeds of the campaign. Ranges are inclusive and uniform. The seed of a split is its base plus the run seed s, so seeds 0, 1 and 2 draw disjoint sequences and every run with the same seed sees the same data.

setting	value
Sequence	
opening length ℓ	3 to 6; the Count split draws 7 to 9 and Count-Interp trains on three of the four values
inner runs	0 to 2 in CLOSE, 1 to 3 in VERIFY and OPEN, each of length 1 to $\ell - 1$
filler gaps	2 to 8 tokens between runs and before the probe, 24 to 96 on Distance
prefix	1 to 4 filler tokens before the opening run, 1 to 48 in prefix-randomized training
already closed	half of the CLOSE items, closed by a run of length ℓ to 9, and 15% of those reopen a fresh fence with 0 to 2 inner runs
VERIFY block	closing run of length ℓ to $\ell + 3$, half of the items safe, half of the unsafe items unsafe through an inner collision and the rest through a short close
Vocabulary	
probe and STOP	0
VERIFY answers	0 for safe and 1 for unsafe
filler	10 to 99, every token an independent draw
fence token	100, or one of 100 to 149 named by the designator token 255 in the binding condition and in OPEN
content identifiers	210 to 239, OPEN-K only
Splits and seeds	
training	3,840,000 sequences, seed $1000 + s$
ID, Distance, Count, Count-Interp, depth	10,000 items each, seeds $2000 + s$, $3000 + s$, $4000 + s$, $4500 + s$ and $5000 + s$
segment splits	4,000 items each, scale factor 12 unless stated, Pad to a total of 1,000 tokens

OPEN-K. The knowledge-conditioned variant asks for the opening length before any content is seen. Thirty content identifiers, tokens 210 to 239, each carry a fixed longest inner run m between 1 and 6. In CLOSE and VERIFY training sequences an identifier precedes the content and the longest run of that content equals its m, so the pairing can be learned only through the next-token auxiliary loss with weight λ. An OPEN-K item shows the identifier and the probe and asks for $m + 1$, and a knowledge probe asks for m. Eighteen identifiers may appear in OPEN-K training items and twelve appear only inside CLOSE and VERIFY content, so the held-out scores of Table 38 are zero-shot. The joint arm trains all four item types together and the isolated arm drops the content sequences, so its held-out score is the memorization baseline. The pilot trains for the standard budget on a mixture of 40% CLOSE, 40% VERIFY, 10% OPEN-K and 10% probe items.

Real systems. Table 3 lists the delimiter systems named in Section 1, the choice a writer makes in each and the failure a collision produces. Each rule is documented in its specification, and the collision it describes is reported in the field. Rust caps the hash count at 255 and rejects a raw string whose content reaches the closing delimiter (rust-lang/rust contributors, 2018). A PostgreSQL function body that needs a second dollar quote must use another tag (Lane, 2009). A multipart library whose boundary could be predicted let a field value contain the boundary and inject a further part (GitHub Advisory Database, 2025). The here-document rule has only its specification behind it here. On Markdown, Lee & Kang (2026) collect the reports, and one user report against ChatGPT proposes the rule of OPEN, a fence one longer than the longest run inside (OpenAI Developer Community, 2025).

Python triple quotes and JSON are the two contrasts. A triple quote has a fixed length, so the writer can only escape it or switch to the other quote family, and JSON opens and closes with different characters, so no run can be mistaken for a closer. Lee & Kang (2026) measured both, with parse validity between 31.7% and 97.8% by model tier for Python docstrings against 100% for JSON. HTML preformatted blocks, YAML block scalars and LaTeX verbatim are a fixed tag with escaping, an indentation state and a single delimiter character, and none of them asks the writer for a length.

Table 3: Delimiter systems with a length or a string to choose, the safe choice, and the failure a collision produces. Labels follow LatentMD.

system	what the writer chooses	safe choice	failure under a collision	source
CommonMark fenced code block	number of backticks	more than the longest inner run, or the tilde family	an inner bare fence closes the outer block (E1 (Premature Closure)) and the intended close opens a block that runs to the end (E2 (Unclosed Fence))	MacFarlane et al., 2024
Rust raw string	number of # signs	more than any "# run inside	the inner "# closes the string early and the rest is parsed as code (E1 (Premature Closure))	The Rust Project Developers, 2025; rust-lang/rust contributors, 2018
POSIX here-document	terminator word	a word absent from the content	a content line equal to the word ends the document early and the remaining lines run as commands (E1 (Premature Closure))	IEEE and The Open Group, 2018
PostgreSQL dollar quoting	tag between the dollar signs	a tag absent from the body	an inner $$ closes the string (E1 (Premature Closure)) and the next tag opens one that never closes (E2 (Unclosed Fence))	The PostgreSQL Global Development Group, 2025; Lane, 2009
MIME multipart	boundary string, committed in the header before the content	a random string	a body line that starts with the boundary ends the part early and creates a part without headers (E1 (Premature Closure))	Freed & Borenstein, 1996; GitHub Advisory Database, 2025
Python triple-quoted string	nothing, the delimiter is fixed	escape, or the other quote family	an inner triple quote closes the string (E1 (Premature Closure)) and the next one opens a string that never terminates (E2 (Unclosed Fence))	Python Software Foundation, 2025
JSON	nothing, opener and closer differ	no choice exists	no collision is possible, only an unbalanced brace (E2 (Unclosed Fence))	Bray, 2017

B PROTOCOL

Ninety CLOSE cells have five seeds instead of three. The sixty cells of the S4D and Mamba-NoSelect scans, one, two, four, eight, twelve and sixteen layers at five rates, were trained at five seeds from the start. Of the other thirty, two seeds were added to cells whose first three seeds disagreed, namely the GRU and RNN anchors, the one-layer Mamba at lr 10^{-3}, the four-layer LSTM at lr 5×10^{-3}, the sinusoidal and the sixteen-head two-layer Transformers and the three cells of the filler control, to the eight-layer Mamba cell at lr 10^{-2} to test whether a converging seed exists, and to the four cells of the four-layer attention-layer ladder to check the inverted U.

Two more seeds went to the standard cells so that the family comparison does not rest on three seeds, namely the lr-matched cell of every other family, with both candidate rates for Mamba, the RNN, the twelve-layer Transformer, the eight-layer NoPE stack and the eight-layer alternating hybrid, the two-layer Transformer at lr 10^{-4} and the two-layer ALiBi cell. All seeds enter every mean, count and best value, and each table prints its number of runs. The 300,000-step arm of the ladder has one seed per cell, and three for the alternating cell, and is marked where it appears.

Training. Table 4 lists every training setting. The generator ranges and the split seeds are in Table 2.

Table 4: Training settings shared by every run. Cells that depart from them are listed under exceptions.

setting	value
optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.01
learning rate	anchor lr 3×10^{-4}; scan over 10^{-4}, 3×10^{-4}, 10^{-3}, 5×10^{-3} and 10^{-2}, with 10^{-5} and 3×10^{-5} added for the Transformer
schedule	cosine to zero after 200 warm-up steps; constant in the schedule control
gradient clipping	global norm 1.0
steps and batch	30,000 steps of 128 sequences, one pass over 3,840,000 sequences
loss	teacher-forced cross-entropy on every answer position
evaluation	every 500 steps on 10,000 items per split, greedy free-running decoding for CLOSE, final checkpoint reported, no early stopping
width, depth, heads	128, 2, 4, no dropout
position table	1,024 positions for learned positions and for the residual position table of Mamba
Mamba	state 16, convolution width 4, expansion 2, sequential scan
S4D	S4D-Lin initialization and zero-order hold after the authors' reference; the state-space parameters are trained without weight decay at lr 10^{-3} as in that reference
seeds	0, 1 and 2, with five seeds in the twenty-seven cells named below

Roster and fidelity. Table 5 records where each implementation comes from, which published setting it keeps and how it was checked. A suite of 152 tests re-parses generated sequences against the fence rule and checks the models.

Table 5: Implementation source, published setting kept and checks for every model of the roster.

model	implementation	published setting kept	checks
Transformer	pre-LN decoder on the PyTorch multi-head attention module, GPT-2 initialization	each positional encoding in its published form	equivalence under weight transfer, causality, shift invariance of the relative encodings, gradient flow
LSTM	PyTorch LSTM	forget-gate bias of 1 (Gers et al., 2000)	shape, padding invariance, overfit on a tiny set
GRU	PyTorch GRU	default initialization, since the GRU paper gives no bias recipe	the same
RNN	PyTorch tanh RNN	default initialization	the same
Mamba	reimplementation of the reference path of the authors' code with a sequential scan, so that the block interior can be modified	state 16, convolution 4, expansion 2, softplus Δ with the published projection initialization, exponential discretization of A initialized at 1 to N, the D skip and the SiLU gate branch	forward equality of the four variants, padding invariance, overfit on a tiny set; no numerical parity test against the fused CUDA kernel
S4D	reimplementation after the authors' S4D module	S4D-Lin initialization, zero-order hold, optimizer groups	kernel output equal to the authors' kernel on random inputs within 10^{-6}, initialization values checked
Mamba-NoSelect	Mamba with Δ, B and C fixed as input-independent learned parameters	the time-invariant ablation of the Mamba paper	absence of the input projection and learning of the fixed parameters checked
Hybrid	Mamba blocks and NoPE attention blocks in a named order	as above	forward and shape tests

Parameters and compute. Parameter counts of every cell are in the cell inventory (Appendix K). The campaign used about 3,100 GPU-hours over 1,987 runs, one GPU per run on NVIDIA RTX 3090, A100 40 GB, A100 80 GB and H100 80 GB cards. A two-layer cell takes 0.4 hours per run for the Transformer and 1.8 hours for Mamba, whose scan is sequential. Mamba-NoSelect is time invariant, so its runs use an FFT form of the same scan, the kernel of its fixed coefficients convolved with the input by rfft and irfft as in S4D, which brings a four-layer run from 3.5 hours to 0.4 hours. On three trained four-layer checkpoints the two forms agree to a relative error below 10^{-5} in the logits and below 2×10^{-5} in the gradients, with the same argmax at every position. The longest cells are the four-layer Mamba of width 512 at 15 hours and the length-matched 512-token Mamba at 63 hours per run.

Sampled decoding. Every CLOSE score of the paper uses greedy decoding. Scoring the standard cells of Table 1 again with the next token drawn from the softmax at temperature 0.5 and 1.0, on 2,000 items per split, moves no cell by more than 0.06 in closing accuracy on Distance and no Count value by more than 0.01, and the safe-close rate does not rise, so the two readings of Section 3 do not depend on the decoding rule.

Exceptions. Five cells or arms depart from the shared settings. The length-matched 512-token cells evaluate every 2,500 steps on 2,500 items with a position table of 8,192, because every evaluation pass runs over 6,144-token sequences; their final metrics are defined as everywhere else. The budget ladder trains for 100,000 or 300,000 steps on the same 3,840,000 sequences, so it repeats data. The training-ceiling control moves the upper end of the ℓ range to 4, 5, 8 or 12. The OPEN-K pilot trains on the mixture described in Appendix A. S4D trains its state-space parameters in their own optimizer group.

Roster and campaign (detail for Section 2). Ten thousand items score the ID, Distance and Count splits and 4,000 items each segment split, so the item-level standard error of a single run is below 0.8 points. Of the 1,987 runs, 1,353 are CLOSE runs at the fixed budget and 1,216 of them pass the gate, 36 are CLOSE runs of the training-budget ladder at 100,000 or 300,000 steps, 55 are CLOSE runs of the training-ceiling control, 21 are the emission-control runs of Section 3.2, 246 are VERIFY runs, 165 are OPEN runs, and the remaining 111 are nested variants. Parameter counts and wall-clock times are listed per cell in the cell inventory of the released repository.

C CONTROLS

Table 6 reports the learning-rate scan of every row of Table 1 and of the S4D and Mamba-NoSelect stacks of one to sixteen layers in closing accuracy on Distance and on Count, and Table 7 the budget ladder in closing accuracy. The lr-matched cell of a family is its best rate in closing accuracy, the metric of every capability sentence. This metric does not score the already-closed items, so it can favour a policy that always emits, and the already-closed accuracy of the segment tables (Appendix E) guards that reading.

The two-layer families were scanned first, with the seed count of each cell in brackets where it is not three, the deeper rows of Table 1 at three seeds per rate, and the S4D and Mamba-NoSelect stacks of one, four, eight, twelve and sixteen layers at five seeds per rate. The best rate leaves the anchor lr for four of the deeper rows, the twelve-layer Transformer, the two NoPE stacks and the eight-layer alternating hybrid, and Table 1 reports those cells. The one-layer Mamba-NoSelect cells fail the gate at every rate, and the two families stay below 0.30 on Distance at every depth and rate.

Table 6: Learning-rate scan of every row of Table 1 and of the S4D and Mamba-NoSelect stacks, closing accuracy on Distance and on Count, mean over gate-passing seeds, three per cell unless bracketed, plus the already-closed accuracy on Distance for the recurrent two-layer rows. Bold marks the best rate, which sets the standard cell, a dagger the rate that Section 4 names, and the last row is S4D at width 210, Distance only.

family	metric	10^{-4}	3×10^{-4}	10^{-3}	5×10^{-3}	10^{-2}
Mamba	closing accuracy	0.58	0.81	**0.95** (5)	0.69	0.85† (5)
	already closed, Distance	0.24	0.73	0.69	0.83	0.87
	Count, closing accuracy	0.00	0.00	0.00	0.00	0.00
LSTM	closing accuracy	0.59	0.72	0.61	**0.75**† (5)	0.71
	already closed, Distance	0.38	0.12	0.30	0.88	0.73
	Count, closing accuracy	0.01	0.00	0.09	0.00	0.00
GRU	closing accuracy	0.69	0.82 (5)	0.87	**0.99**† (5)	0.95
	already closed, Distance	0.24	0.49	0.65	1.00	0.92
	Count, closing accuracy	0.00	0.00	0.06	0.00	0.00
RNN	closing accuracy	0.80	0.95 (5)	**1.00** (5)	0.98	0.95† (5)
	already closed, Distance	0.38	0.61	1.00	0.86	0.99
	Count, closing accuracy	0.00	0.00	0.00	0.00	0.00
S4D	closing accuracy	0.16 (5)	0.14 (5)	0.14† (5)	0.18 (5)	**0.28** (5)
	Count, closing accuracy	0.01	0.03	0.01	0.00	0.00
Mamba-NoSelect	closing accuracy	0.19† (5)	**0.19** (5)	0.13 (5)	0.15 (5)	0.08 (5)
	Count, closing accuracy	0.00	0.00	0.00	0.00	0.00
Transformer 2L, learned positions	closing accuracy	0.11† (5)	0.13	0.10	**0.20** (5)	0.11
	Count, closing accuracy	0.22	0.04	0.06	0.09	0.09
Transformer 12L, learned positions	closing accuracy	0.39	0.45† (5)	**0.55** (5)	0.17	0.17
	Count, closing accuracy	0.61	0.84	0.72	0.29	0.39
Transformer+ALiBi 2L	closing accuracy	**0.62**† (5)	0.32	0.19	0.54	0.27
	Count, closing accuracy	0.11	0.05	0.07	0.10	0.00
Attention-only NoPE 4L	closing accuracy	**0.10** (5)	0.06† (5)	0.10	0.06	0.03
	Count, closing accuracy	0.92	0.90	0.97	0.96	0.85
Attention-only NoPE 8L	closing accuracy	0.34	0.20† (5)	0.17	0.22	**0.35** (5)
	Count, closing accuracy	0.99	0.99	1.00	0.99	0.99
Hybrid `MAMA` 4L	closing accuracy	0.83	**0.96**† (5)	0.89	0.70	0.74
	Count, closing accuracy	0.00	0.00	0.00	0.00	0.00
Hybrid `MAMAMAMA` 8L	closing accuracy	0.90	0.95† (5)	**0.97** (5)	0.89	0.89
	Count, closing accuracy	0.01	0.10	0.04	0.00	0.00
S4D 1L	closing accuracy	gate	**0.05** (5)	0.04 (5)	0.04 (5)	0.03 (5)
	Count, closing accuracy	gate	0.00	0.00	0.00	0.00
S4D 4L	closing accuracy	0.12 (5)	0.09 (5)	0.17 (5)	0.27 (5)	**0.28** (5)
	Count, closing accuracy	0.04	0.01	0.00	0.00	0.00
S4D 8L	closing accuracy	0.13 (5)	0.20 (5)	0.24 (5)	0.27 (5)	**0.27** (5)
	Count, closing accuracy	0.02	0.00	0.00	0.00	0.00
S4D 12L	closing accuracy	0.14 (5)	0.19 (5)	0.25 (5)	0.27 (5)	**0.28** (5)
	Count, closing accuracy	0.06	0.00	0.00	0.00	0.00
S4D 16L	closing accuracy	0.15 (5)	0.21 (5)	0.26 (5)	0.28 (5)	**0.28** (5)
	Count, closing accuracy	0.01	0.00	0.00	0.00	0.00
Mamba-NoSelect 1L	closing accuracy	gate	gate	gate	gate	gate
	Count, closing accuracy	gate	gate	gate	gate	gate
Mamba-NoSelect 4L	closing accuracy	0.26 (5)	0.26 (5)	0.21 (5)	0.20 (5)	**0.28** (5)
	Count, closing accuracy	0.00	0.00	0.00	0.00	0.00
Mamba-NoSelect 8L	closing accuracy	0.29 (5)	**0.29** (5)	0.27 (5)	0.28 (5)	0.28 (5)
	Count, closing accuracy	0.00	0.00	0.00	0.00	0.00
Mamba-NoSelect 12L	closing accuracy	0.27 (5)	0.28 (5)	**0.28** (5)	0.28 (5)	0.28 (5)
	Count, closing accuracy	0.00	0.00	0.00	0.00	0.00
Mamba-NoSelect 16L	closing accuracy	0.27 (5)	0.28 (5)	**0.29** (5)	0.28 (5)	0.28 (5)
	Count, closing accuracy	0.00	0.00	0.00	0.00	0.00
S4D 2L, width 210	closing accuracy	0.14	0.03	0.11	0.11	0.27
	Count, closing accuracy	–	–	–	–	–

Table 7: Training-budget ladder. Closing accuracy on the Distance split / on the Count split, mean over gate-passing seeds (three per cell; the count in brackets where a seed fails the gate). The training set is fixed at 3.84 million sequences, so 100,000 steps is 3.3 epochs and 300,000 is 10 epochs. The 30,000-step column is the corresponding cell of the fixed-budget campaign.

cell	30,000 steps	100,000 steps	300,000 steps
Transformer+ALiBi 2L, lr 10^{-4}, cosine	0.62 / 0.11	0.32 / 0.12	0.36 / 0.05
Transformer+ALiBi 2L, lr 10^{-4}, constant	–	0.37 / 0.05	0.35 / 0.11
Transformer 2L, lr 10^{-4}, cosine	0.11 / 0.22	–	0.18 / 0.05
Transformer+RoPE 2L, lr 10^{-4}, cosine	0.11 / 0.00	–	0.20 / 0.00
Transformer NoPE 2L, lr 10^{-4}, cosine	gate	–	gate
Transformer+ALiBi 8L, lr 10^{-4}, cosine	0.59 / 0.47	0.50 / 0.49	–
Mamba 2L, anchor lr, cosine	0.81 / 0.00	0.68 / 0.00	0.60 / 0.00

Table 8: The four-layer attention-layer ladder after ten times the training budget. Closing accuracy on the Distance and Count splits, 30,000 steps (mean of five seeds) against 300,000 steps (seed 0, and the mean of three seeds for M A M A, whose Count values are 0.23, 0.00 and 0.27). The order of the four cells is unchanged.

	Distance		Count	
cell	30k	300k	30k	300k
M M M M	0.42	0.33	0.00	0.00
M M A M	0.87	0.61	0.00	0.03
M A M A	0.96	0.84	0.00	0.17
A A A A	0.06	0.03	0.90	0.99

Table 9: Training-ceiling control for the Count axis. The training range of ℓ ends at 4, 5, 8 or 12 instead of 6, and the Count split asks for every length from the ceiling plus one to 9, or 13 to 16 from ceiling 12. Entries are closing accuracy on the unseen lengths, mean over seeds, with the per-length values in brackets and the per-seed means of the LSTM in parentheses where one seed extrapolates.

cell	ceiling 4 (test 5 to 9)	ceiling 5 (test 6 to 9)	ceiling 8 (test 9)	ceiling 12 (test 13 to 16)
Mamba 2L, lr 10^{-2}	0.00 [0.00 0.00 0.00 0.00 0.00]	0.00 [0.00 0.00 0.00 0.00]	0.00 (5 seeds)	0.02 [0.06 0.01 0.00 0.00]
LSTM 2L, lr 5×10^{-3}	0.00 [0.00 0.00 0.00 0.00 0.00]	0.00 [0.00 0.00 0.00 0.00]	0.21 (1.00/0.05/0/0/0)	0.13 [0.37 0.13 0.00 0.00] (0.35/0.03/0)
Transformer 2L, anchor lr	0.00 [0.00 0.00 0.01 0.00 0.00]	0.00 [0.00 0.00 0.00 0.00]	0.59	0.23 [0.63 0.23 0.06 0.00]
Transformer 12L, anchor lr	0.34 [0.96 0.56 0.14 0.02 0.01]	0.46 [0.92 0.64 0.18 0.08]	–	–
Attention-only NoPE 4L, anchor lr	0.40 [0.94 0.58 0.25 0.13 0.08]	0.80 [0.97 0.89 0.74 0.58]	0.99	–

Table 10: Count-Interp, the interpolation control of Section 3, two layers at the anchor lr. Closing accuracy on the held-out length per seed, its mean, and accuracy on the already-closed items. The two hold-out patterns differ in which length is withheld.

cell	seeds	closing accuracy per seed	mean	ans-0
train $\ell \in \{3, 4, 6\}$, *test* $\ell = 5$				
LSTM	3	0.57, 0.59, 0.34	0.50	0.86
Mamba	3	0.07, 0.00, 0.00	0.02	0.87
Transformer	3	0.21, 0.02, 0.41	0.22	0.95
train $\ell \in \{3, 5, 6\}$, *test* $\ell = 4$				
LSTM	3	0.49, 0.89, 0.96	0.78	0.94
Mamba	3	0.00, 0.00, 0.00	0.00	0.89
Transformer	3	0.04, 0.06, 0.02	0.04	0.91

Table 11 varies width, depth, the Mamba components, the state capacity and the schedule at the anchor lr. Width raises Mamba's distance score from 0.71 at 64 to 0.91 at 256 and 512 and lowers the LSTM's to 0.35 at 512, and no width moves the count axis of either. Depth beyond two layers costs both recurrent families at this learning rate and raises the Transformer's count axis, from 0.04 at two layers to 0.68 at eight. Removing the causal convolution moves Mamba from 0.81 to 0.72, a residual position table changes nothing, the symmetric convolution reads future tokens and cannot generate, and Mamba-NoSelect keeps 0.19. At matched state capacity a wide model with a small

state beats a narrow model with a large state. A constant schedule helps Mamba at both learning rates, 0.81 to 0.91 and 0.85 to 0.99, and hurts the LSTM at the anchor, 0.72 to 0.54.

Table 6 holds the learning-rate ladders of the recurrent families, and the Count closing accuracy of every recurrent cell in it is at most 0.09. RNN and GRU reach 1.00 and 0.99 at their own rate, the LSTM reaches 0.75 at width 128 and 1.00 at width 512 (Table 18), and S4D and Mamba-NoSelect stay below 0.3 at every rate and every depth from one to sixteen layers, their best cells at 0.27 to 0.29 from four layers on.

Table 12 lists the task-condition and vocabulary controls. Naming the fence role by a designator token lowers every family. Removing the inner runs, the pure-carry filler condition, lowers the LSTM to 0.04 while Mamba keeps 0.50, and a higher learning rate recovers part of both. Paired inner runs raise every family, a narrowed vocabulary lowers the LSTM most, and length-matched training halves the LSTM, leaves Mamba unchanged and removes the Transformer's distance score entirely, because the twelvefold sequences leave its position table. The two Count-Interp rows are the interpolation control of Section 3 under both hold-out patterns, and Table 10 gives their per-seed values.

Table 11: Architecture axes at the anchor lr on CLOSE. Runs and, in brackets, gate-passing runs; ID is accuracy over all items; Distance and Count are closing accuracy on the two splits and safe is the safe-close rate on Distance, means over gate-passing runs. Width 128 and two layers are the standard cell of Section 2.

cell	runs	ID	Distance	Count	safe
LSTM, width at two layers					
width 64	3 (3)	1.00	0.51	0.05	0.53
width 128	3 (3)	1.00	0.72	0.00	0.75
width 256	3 (3)	1.00	0.68	0.06	0.68
width 512	3 (3)	1.00	0.35	0.00	0.39
Mamba, width at two layers					
width 64	3 (3)	1.00	0.71	0.00	0.73
width 128	3 (3)	1.00	0.81	0.00	0.85
width 256	3 (3)	1.00	0.91	0.00	0.96
width 512	3 (3)	1.00	0.91	0.00	0.93
Transformer, width at two layers					
width 64	3 (3)	1.00	0.10	0.16	0.12
width 128	3 (3)	1.00	0.13	0.04	0.19
width 256	3 (3)	1.00	0.18	0.00	0.21
width 512	3 (3)	1.00	0.21	0.01	0.25
LSTM, depth at width 128					
1L	3 (3)	1.00	0.48	0.01	0.55
2L	3 (3)	1.00	0.72	0.00	0.75
4L	3 (3)	1.00	0.47	0.00	0.48
8L	3 (3)	1.00	0.50	0.00	0.52
Mamba, depth at width 128					
1L	3 (3)	0.96	0.45	0.00	0.56
2L	3 (3)	1.00	0.81	0.00	0.85
4L	3 (3)	1.00	0.50	0.00	0.52
8L	3 (3)	1.00	0.49	0.00	0.51
Transformer, depth at width 128					
1L	3 (3)	0.96	0.05	0.17	0.10
2L	3 (3)	1.00	0.13	0.04	0.19
4L	3 (3)	1.00	0.16	0.28	0.31
8L	3 (3)	1.00	0.28	0.68	0.74
Mamba components, two layers					
state size 8	3 (3)	1.00	0.65	0.00	0.67
state size 32	3 (3)	1.00	0.80	0.00	0.90
conv width 2	3 (3)	1.00	0.85	0.01	0.91
conv width 8	3 (3)	1.00	0.98	0.00	0.98
no conv	3 (3)	1.00	0.72	0.03	0.75
residual position table	3 (3)	1.00	0.80	0.00	0.80
symmetric conv	3 (0)	0.00	–	–	–
Mamba-NoSelect	5 (5)	0.99	0.19	0.00	0.33
Mamba at matched state capacity					
width 64, state 32	3 (3)	1.00	0.67	0.00	0.69
width 256, state 8	3 (3)	1.00	0.95	0.00	0.96
Constant learning-rate schedule					
LSTM, anchor lr	3 (3)	1.00	0.54	0.00	0.55
LSTM, lr 5×10^{-3}	3 (3)	1.00	0.82	0.00	0.83
Mamba, anchor lr	3 (3)	1.00	0.91	0.00	0.92
Mamba, lr 10^{-2}	3 (3)	1.00	0.99	0.00	1.00
Transformer+ALiBi, anchor lr	3 (3)	0.98	0.29	0.04	0.46
Transformer+ALiBi, step drop at 6k	3 (3)	0.99	0.28	0.07	0.52
Transformer attention heads, two layers					
4 heads	3 (3)	1.00	0.13	0.04	0.19
8 heads	3 (3)	1.00	0.18	0.00	0.22
16 heads	5 (5)	1.00	0.18	0.04	0.24

Table 12: Task-condition and vocabulary controls at the anchor lr, two layers, three seeds unless stated. Each entry is closing accuracy on Distance with ID accuracy over all items in brackets, mean over gate-passing runs. The Count-Interp rows train on three of the four lengths and report closing accuracy on the held-out middle length.

condition	LSTM	Mamba	Transformer
Distance closing accuracy (ID accuracy in brackets)			
standard cell	0.72 (1.00)	0.81 (1.00)	0.13 (1.00)
binding, designator token names the fence role	0.42 (1.00)	0.33 (1.00)	0.11 (1.00)
filler, pure carry with no inner runs	0.04 (1.00), $n = 5$	0.50 (1.00), $n = 5$	0.07 (1.00), $n = 5$
pairing, inner runs come as open and close pairs	0.76 (1.00)	0.87 (1.00)	0.22 (1.00)
vocabulary, both noise and fence pools narrowed	0.27 (1.00)	0.69 (1.00)	0.10 (1.00)
vocabulary, noise pool narrowed	0.43 (1.00)	0.49 (1.00)	0.09 (1.00)
vocabulary, fence pool narrowed	0.30 (1.00)	0.84 (1.00)	0.08 (1.00)
length-matched to 128 tokens	0.37 (1.00)	0.72 (1.00)	0.00 (1.00)
length-matched to 256 tokens	0.39 (1.00)	0.68 (1.00)	0.04 (0.99)
length-matched to 512 tokens	0.35 (1.00)	0.82 (1.00)	0.00 (0.99)
filler at a higher learning rate			
filler, lr 5×10^{-3}	0.22 (1.00)	–	–
filler, lr 10^{-2}	0.34 (1.00)	0.73 (1.00)	–
Count-Interp closing accuracy on the held-out middle length			
train $\ell \in \{3, 4, 6\}$, test $\ell = 5$	0.50	0.02	0.22
train $\ell \in \{3, 5, 6\}$, test $\ell = 4$	0.78	0.00	0.04

C.1 Robustness of the dissociation in full (detail for Section 3)

Learning rate. The table already contains the lr-matched cell of every family, so the split is not a learning-rate artifact. At the anchor lr the recurrent families are under-tuned. Their Distance closing accuracy rises from 0.72 to 0.75 for the LSTM and from 0.81 to 0.95 for Mamba at their lr-matched rates, and RNN, GRU and LSTM all reach 0.99 or more in their best cells (Table 6).

Depth. Depth raises the attention side's count axis and does not raise the recurrent side's distance axis. Deep learned-position Transformers reach Count closing accuracy of 0.68, 0.84 and 0.67 at 8, 12 and 16 layers while their Distance means stay at 0.28 to 0.53, and a lower learning rate moves the two axes in opposite directions for these models. At their matched learning rate the Mamba stacks of one to eight layers keep the distance axis at 0.85 or above, and the four-layer stack stays at 1.00 on Gaps through a 16-fold gap scale and at 0.95 at 24-fold (Table 29). At 12 and 16 layers the learning rate that trains shallow Mamba fails the gate on every seed, and the next lower rate reaches closing accuracy of 0.76 and 0.64 while it answers almost no already-closed item correctly, the under-tuned signature of Section 4 in its strongest form. The count axis stays at zero at every depth and learning rate.

A reader may take the deep learned-position cells for a trend rather than a boundary, an attention side that is a superset waiting on depth and data. The data allow that reading for distance, where depth moves the boundary. For count the recurrent side does not move with depth, width, learning rate or a tenfold budget. No recurrent-only cell of the campaign exceeds 0.12 on Count and no single run 0.34, over 166 cells and 700 runs. What moves it is the removal of recurrent layers. The four-layer stacks with a single Mamba layer reach 0.11 to 0.45 and two of the six stacks with two Mamba layers 0.13 and 0.17 (Table 37), one seed of the eight-layer alternating hybrid reaches 0.41, the four-layer alternating hybrid reaches 0.23 and 0.27 in two of three seeds after 300,000 steps, and two LSTM seeds extrapolate in the ceiling control (Table 9).

Budget. Ten times the training budget lowers both sides in the repeated-data regime. After 300,000 steps the two-layer ALiBi Transformer falls from 0.62 to 0.36 and the anchor Mamba from 0.81 to 0.60 in closing accuracy, and the ordering does not change (Table 7). For the hybrid ladder the tenfold budget rests on one seed, and on three for the alternating cell. Under-training does not explain the attention side's distance failure.

Gate. The two-layer NoPE Transformer never passes the gate at any learning rate or budget. Its numbers are those of a model that did not learn the task. The sixteen-layer LSTM fails it on every seed at both of its rates (Table 16), so the LSTM row of Figure 1 ends at twelve layers. The one-layer Mamba-NoSelect cells fail it on every seed at every rate, with ID accuracy of 0.64 to 0.91, so the Mamba-NoSelect row starts at two layers, and the one-layer S4D cell at lr 10^{-4} fails it on every seed.

Count ceiling. The count ceiling is not a definitional artifact. Deep and NoPE attention stacks extrapolate the same counts in the same task, and the generative format has no unseen answer token. A control that moves the training ceiling to 4, 5 and 8 sharpens both sides (Table 9, three seeds per cell and five for the LSTM and Mamba cells at ceiling 8, all of which pass the ID gate). Deep and NoPE attention stacks extrapolate from every ceiling with a graded decay. Two-layer learned-position attention extrapolates from ceilings 8 and 12, at 0.59 and 0.23, and from ceilings 4 and 5 not at all, so count extrapolation on the attention side is a property of depth and of the NoPE stacks. On the recurrent side one LSTM seed of five extrapolates perfectly from ceiling 8 to length 9, the one recurrent run of the campaign that closes an unseen length, while the other four LSTM seeds sit at 0.05 or below and all five Mamba seeds at zero. Mamba stays at or below 0.06 at the first length beyond every ceiling and at 0.01 or below past it.

Nor is the recurrent side one lookup table. On Count-Interp, where the held-out length lies inside the training range, Mamba reaches a closing accuracy of 0.02 and 0.00 under the two hold-out patterns, the behaviour of a lookup table, while the LSTM reaches 0.50 and 0.78 and the Transformer 0.22 and 0.04 (Table 10). The lookup reading holds for Mamba and not for the LSTM, whose ceiling is a failure to extend a rule beyond four training values rather than the absence of one, and the literature on classical RNN counting says the same (Section 6). Widening the training range to 3 to 12 and asking for 13 to 16 gives the same picture with more values. One LSTM seed of three closes length 13 perfectly and length 14 at 0.40, the other two close 13 at 0.13 and 0.00, Mamba closes 13 at 0.02 to 0.16 and nothing beyond, and the two-layer learned-position Transformer decays from 0.63 at 13 to 0.06 at 15 (Table 9).

The judgment task corroborates the distance axis and shows the count axis with a smaller gap. In VERIFY the two-layer Transformer is at chance on distance while recurrent models reach 0.94 to 0.96, whereas on count the recurrent cells score 0.57 to 0.82 against a chance of 0.5, the two-layer Transformers 0.58 to 0.75, and attention stacks of four or more layers 0.94 to 1.00 (Table 23). The direction is that of CLOSE, where two-layer Transformers do not extrapolate count either, and the floor of a binary judgment compresses the gap.

Learning-rate choice on a fresh draw. The lr-matched rate of a family is the rate with its best closing accuracy on the Distance split, which is also a reporting split. To test whether that choice leaks into the reported values, every checkpoint of the scans was scored again on an independent draw of the split, generated from the same definition with a different generator seed. The two draws share no items. For all eight scanned families the rate chosen on the fresh draw is the rate chosen on the original draw, and no cell mean differs by more than 0.01 between the draws. These rates are the standard cells of Tables 1 and 15, 10^{-3} for Mamba and the RNN, 5×10^{-3} for the LSTM, the GRU and the two-layer Transformer, 10^{-2} for S4D and 3×10^{-4} for Mamba-NoSelect.

Emission-range control. In the Count split the correct answer is a run of 7 to 9 tokens, a length that no training item asks a model to emit. In the emission-range control a quarter of the training items is replaced by items of a second role, marked by its own probe token. Such an item is a closed block whose closing run has 7 to 9 tokens, and the target is to emit that run again. The opening length of every CLOSE item stays between 3 and 6, the evaluation splits are unchanged item for item, and the training budget is the same, so three quarters of it go to standard CLOSE items. The control is an additional data module and runner, and no file of the standard pipeline changes.

Every model learns the second role, at 1.00 on held-out items of that role. The Count split does not move. All 18 recurrent runs stay at 0.05 or below, and the twelve-layer Transformer keeps 0.88 against 0.88 for the same seeds of its standard cell (Table 13). The control shifts the distance axis in both directions, the LSTM from 0.83 to 0.60 and Mamba at 10^{-2} from 0.84 to 0.95, and leaves the count axis where it was. The recurrent count failure is therefore not a failure to emit runs of the unseen lengths. Broader training ranges are covered by the ceiling control above, while curricula and auxiliary counting losses were not tested.

Table 13: Emission-range control. A quarter of the training items asks for a closing run of 7 to 9 tokens in a second role, with the CLOSE items and every evaluation split unchanged. Runs (gate-passing in brackets), ID accuracy, closing accuracy on Distance and Count, held-out accuracy of the second role, and the same seeds of the standard cell. Means over gate-passing runs.

			emission-range control			standard cell	
cell	runs	ID	Distance	Count	second role	Distance	Count
Mamba 2L, lr 10^{-3}	3 (3)	1.00	0.92	0.00	1.00	0.93	0.00
Mamba 2L, lr 10^{-2}	3 (3)	1.00	0.95	0.00	1.00	0.84	0.00
LSTM 2L, lr 5×10^{-3}	3 (3)	1.00	0.60	0.02	1.00	0.83	0.00
GRU 2L, lr 5×10^{-3}	3 (3)	1.00	1.00	0.00	1.00	0.99	0.00
RNN 2L, lr 10^{-3}	3 (3)	1.00	0.99	0.00	1.00	0.99	0.00
RNN 2L, lr 10^{-2}	3 (2)	0.98	0.97	0.00	1.00	0.95	0.00
Transformer 12L, anchor lr	3 (3)	1.00	0.41	0.88	1.00	0.49	0.88

Capacity. Parameter counts differ across the roster, and Table 14 lists them beside the two axes. Width does not buy count extrapolation on the attention side. The two-layer Transformer stays at 0.16 or below from width 64 to width 512, while depth at width 128 reaches 0.68 at eight layers with a quarter of the parameters of the widest two-layer model. On the recurrent side no width or depth lifts the count axis, and the LSTM of width 512 reaches 1.00 on Distance with 4.5M parameters and 0.00 on Count. S4D with its width raised to 210, which matches the 0.30M parameters of Mamba, stays at 0.03 to 0.27 on Distance across the five learning rates (Table 6).

Table 14: Parameter count next to the two axes. Width and depth arms of CLOSE at the anchor lr unless a rate is given, runs with gate-passing runs in brackets, closing accuracy on Distance and Count, means over gate-passing runs.

cell	parameters	runs	Distance	Count
Transformer 2L, width 64	0.20M	3 (3)	0.10	0.16
Transformer 2L, width 128	0.59M	3 (3)	0.13	0.04
Transformer 2L, width 256	1.97M	3 (3)	0.18	0.00
Transformer 2L, width 512	7.09M	3 (3)	0.21	0.01
Transformer 4L, width 128	0.99M	3 (3)	0.16	0.28
Transformer 8L, width 128	1.78M	3 (3)	0.28	0.68
Transformer 12L, width 128	2.58M	5 (5)	0.45	0.84
Transformer 16L, width 128	3.37M	3 (3)	0.53	0.67
LSTM 2L, width 64	0.10M	3 (3)	0.51	0.05
LSTM 2L, width 128	0.33M	3 (3)	0.72	0.00
LSTM 2L, width 256	1.18M	3 (3)	0.68	0.06
LSTM 2L, width 512	4.46M	3 (3)	0.35	0.00
LSTM 2L, width 512, lr 5×10^{-3}	4.46M	3 (3)	1.00	0.00
Mamba 2L, width 64	0.10M	3 (3)	0.71	0.00
Mamba 2L, width 128	0.30M	3 (3)	0.81	0.00
Mamba 2L, width 256	1.01M	3 (3)	0.91	0.00
Mamba 2L, width 512	3.65M	3 (3)	0.91	0.00
Mamba 4L, width 512	7.04M	3 (3)	0.73	0.00
S4D 2L, width 128, lr 10^{-2}	0.14M	5 (5)	0.28	0.00
S4D 2L, width 210, lr 10^{-2}	0.30M	3 (3)	0.27	0.00

D DISSOCIATION DETAIL

Table 17 lists every two-layer Transformer cell by positional encoding and learning rate and the deep cells by encoding, and Table 18 the recurrent cells at matched learning rates, by depth and width. Three readings hold across the two tables. No two-layer encoding exceeds 0.32 on Distance at the anchor lr, and ALiBi at lr 10^{-4} is the one two-layer cell above 0.5. NoPE at two layers never passes the gate at any learning rate or budget. On the recurrent side every family has a cell at 0.94 or more on Distance and no cell above 0.03 on Count, and the deep Mamba cells at lr 10^{-2} fail the gate from twelve layers on while the eight-layer cell passes on one seed of five. Per-seed values of every cell are in the release.

The count-axis error classes exist for every run of the fixed-budget campaign, so no row of Table 20 rests on a subset of its runs. The 300,000-step ladder and the training-ceiling control, which enter no main-text table, were not scored with them.

Table 15 gives one standard cell per family at its lr-matched rate with the seed standard deviation and the best seed, the unit behind the family orderings of the text, whereas the family rows of Table 20 pool every cell of the sweep and so also reflect its composition, and Table 16 counts the 137 fixed-budget runs that fail the gate by family and cell.

Table 15: One standard cell per family at its own learning rate (the lr-matched cell of the family, two layers and width 128 unless stated). Mean $\pm$ seed standard deviation over the gate-passing seeds and the best seed, with the number of gate-passing seeds (seeds run) in the second column.

		Distance closing accuracy		Count closing accuracy	
cell	seeds	mean $\pm$ SD	best seed	mean $\pm$ SD	best seed
Mamba 2L, lr 10^{-3}	5 (5)	0.95 ± 0.07	0.99 (seed 2)	0.00 ± 0.00	0.00 (seed 0)
LSTM 2L, lr 5×10^{-3}	5 (5)	0.75 ± 0.11	0.88 (seed 1)	0.00 ± 0.00	0.00 (seed 0)
GRU 2L, lr 5×10^{-3}	5 (5)	0.99 ± 0.02	1.00 (seed 3)	0.00 ± 0.00	0.00 (seed 0)
RNN 2L, lr 10^{-3}	5 (5)	1.00 ± 0.01	1.00 (seed 1)	0.00 ± 0.00	0.00 (seed 0)
Hybrid M A M A 4L, anchor lr	5 (5)	0.96 ± 0.03	1.00 (seed 4)	0.00 ± 0.00	0.00 (seed 0)
Hybrid M A M A M A M A 8L, lr 10^{-3}	5 (5)	0.97 ± 0.02	1.00 (seed 2)	0.04 ± 0.06	0.14 (seed 0)
S4D 2L, lr 10^{-2}	5 (5)	0.28 ± 0.01	0.29 (seed 2)	0.00 ± 0.00	0.00 (seed 0)
Mamba-NoSelect 2L, best lr	5 (5)	0.19 ± 0.10	0.27 (seed 1)	0.00 ± 0.00	0.00 (seed 1)
Transformer 2L, lr 5×10^{-3}	5 (5)	0.20 ± 0.03	0.24 (seed 2)	0.09 ± 0.10	0.21 (seed 3)
Transformer+ALiBi 2L, lr 10^{-4}	5 (5)	0.62 ± 0.29	0.83 (seed 3)	0.11 ± 0.09	0.25 (seed 2)
Transformer 12L, lr 10^{-3}	5 (5)	0.55 ± 0.18	0.74 (seed 4)	0.72 ± 0.10	0.82 (seed 1)
Attention-only NoPE 4L, lr 10^{-4}	5 (5)	0.10 ± 0.10	0.28 (seed 1)	0.92 ± 0.05	0.96 (seed 0)
Attention-only NoPE 8L, lr 10^{-2}	5 (5)	0.35 ± 0.34	0.93 (seed 0)	0.99 ± 0.01	1.00 (seed 2)

Table 16: Gate survivorship of the fixed-budget CLOSE campaign. Runs, runs that pass the ID gate of 0.95, runs that fail it, the pass rate, and the failing cells with their failing-seed count. The comparisons of the paper use the passing runs only.

family	runs	pass	fail	rate	failing cells
Mamba, 1–16L	157	144	13	0.92	mamba L12 d128 lr0.01 (3 of 3), mamba L16 d128 lr0.01 (3 of 3), mamba L4 d512 lr0.01 (3 of 3), mamba L8 d128 lr0.01 (4 of 5)
LSTM, 1–12L	135	126	9	0.93	lstm L12 d128 lr0.005 (3 of 3), lstm L16 d128 lr0.0003 (3 of 3), lstm L16 d128 lr0.005 (3 of 3)
GRU, 1–16L	73	65	8	0.89	gru L12 d128 lr0.005 (3 of 3), gru L16 d128 lr0.005 (3 of 3), gru L8 d128 lr0.005 (2 of 3)
RNN, 1–16L	57	52	5	0.91	rnn L12 d128 lr0.001 (1 of 3), rnn L16 d128 lr0.0003 (1 of 3), rnn L16 d128 lr0.001 (3 of 3)
Hybrid 4L, 0–2 attn layers	63	62	1	0.98	hybrid A=None L4 d128 lr0.01 (1 of 3)
Hybrid 8L, 0–4 attn layers	34	34	0	1.00	–
S4D, 1–16L	165	160	5	0.97	s4d L1 d128 lr0.0001 (5 of 5)
Mamba-NoSelect, 2–16L	150	125	25	0.83	mamba no_select L1 d128 lr0.0001 (5 of 5), mamba no_select L1 d128 lr0.0003 (5 of 5), mamba no_select L1 d128 lr0.001 (5 of 5), mamba no_select L1 d128 lr0.005 (5 of 5), mamba no_select L1 d128 lr0.01 (5 of 5)
Transformer 1–2L, six PEs	288	226	62	0.78	transformer alibi L1 d128 lr0.0001 (3 of 3), transformer alibi L1 d128 lr0.0003 (3 of 3), transformer alibi L2 d128 lr0.0001 slope x0.0833 (3 of 3), transformer alibi L2 d128 lr0.0001 slope x0.25 (2 of 3), transformer learned L1 d128 lr0.0001 (3 of 3), transformer learned L2 d128 lr0.0001 posoff1000 (3 of 3), transformer learned L2 d128 lr0.0003 posoff1000 (3 of 3), transformer learned L2 d128 lr0.0003 win32 (3 of 3), transformer learned L2 d128 lr0.0003 win64 (3 of 3), transformer learned L2 d128 lr1e-05 (3 of 3), transformer none L2 d128 lr0.0001 (3 of 3), transformer none L2 d128 lr0.0003 (3 of 3), transformer none L2 d128 lr0.001 (3 of 3), transformer none L2 d128 lr0.003 (3 of 3), transformer rope L1 d128 lr0.0001 (3 of 3), transformer rope L1 d128 lr0.0003 (3 of 3), transformer sinusoidal L1 d128 lr0.0001 (3 of 3), transformer sinusoidal L1 d128 lr0.0003 (3 of 3), transformer sinusoidal L2 d128 lr0.0003 posoff1000 (3 of 3), transformer t5 L1 d128 lr0.0001 (3 of 3), transformer t5 L1 d128 lr0.0003 (3 of 3)
Transformer 4–16L	148	148	0	1.00	–
Attention-only NoPE, 4–16L	44	44	0	1.00	–
Hybrid 4L, 3 attn layers	12	12	0	1.00	–
Mamba only, hybrid class, 1, 2, 12, 16L	12	12	0	1.00	–
Attention-only NoPE, 1–2L	6	0	6	0.00	hybrid A=0 L1 d128 lr0.0003 (3 of 3), hybrid A=0,1 L2 d128 lr0.0003 (3 of 3)
Mamba variants (no conv, residual position table, symmetric conv)	9	6	3	0.67	mamba sym_conv L2 d128 lr0.0003 (3 of 3)

Table 17: Transformer CLOSE cells by positional encoding, learning rate and depth. Runs and gate-passing runs, ID accuracy over all items, closing accuracy on Distance and Count and the safe-close rate on Distance, means over gate-passing runs. Width 128 and four heads throughout.

cell	runs	ID	Distance	Count	safe
Learned positions, two layers					
lr 10^{-5}	3 (0)	0.92	–	–	–
lr 3×10^{-5}	3 (3)	0.96	0.12	0.29	0.34
lr 10^{-4}	5 (5)	0.99	0.11	0.22	0.17
anchor lr	3 (3)	1.00	0.13	0.04	0.19
lr 10^{-3}	3 (3)	1.00	0.10	0.06	0.19
lr 5×10^{-3}	5 (5)	1.00	0.20	0.09	0.30
lr 10^{-2}	3 (3)	1.00	0.11	0.09	0.28
Sinusoidal, two layers					
anchor lr	5 (5)	0.99	0.11	0.26	0.20
lr 10^{-4}	3 (3)	0.96	0.04	0.32	0.09
lr 10^{-3}	3 (3)	0.99	0.12	0.15	0.22
base 100, anchor lr	3 (3)	0.99	0.06	0.26	0.14
base 1,000	3 (3)	0.98	0.05	0.36	0.09
base 100,000	3 (3)	0.99	0.11	0.28	0.16
RoPE, two layers					
anchor lr	3 (3)	1.00	0.12	0.00	0.25
lr 10^{-4}	3 (3)	1.00	0.11	0.00	0.23
lr 10^{-3}	3 (3)	1.00	0.21	0.00	0.29
ALiBi, two layers					
anchor lr	3 (3)	0.99	0.32	0.05	0.41
lr 10^{-4}	5 (5)	0.99	0.62	0.11	0.73
lr 10^{-3}	3 (3)	0.99	0.19	0.07	0.22
T5 relative bias, two layers					
anchor lr	3 (3)	0.99	0.17	0.04	0.41
lr 10^{-4}	3 (3)	0.96	0.05	0.10	0.27
lr 10^{-3}	3 (3)	1.00	0.02	0.00	0.15
NoPE, two layers					
anchor lr	3 (0)	0.91	–	–	–
lr 10^{-4}	3 (0)	0.90	–	–	–
lr 10^{-3}	3 (0)	0.92	–	–	–
lr 3×10^{-3}	3 (0)	0.92	–	–	–
Sinusoidal, RoPE, ALiBi and T5 at one layer					
sinusoidal 1L, anchor lr	3 (0)	0.94	–	–	–
sinusoidal 1L, lr 10^{-4}	3 (0)	0.91	–	–	–
RoPE 1L, anchor lr	3 (0)	0.92	–	–	–
RoPE 1L, lr 10^{-4}	3 (0)	0.86	–	–	–
ALiBi 1L, anchor lr	3 (0)	0.76	–	–	–
ALiBi 1L, lr 10^{-4}	3 (0)	0.74	–	–	–
T5 1L, anchor lr	3 (0)	0.86	–	–	–
T5 1L, lr 10^{-4}	3 (0)	0.80	–	–	–
Learned positions, depth					
1L, lr 10^{-4}	3 (0)	0.95	–	–	–
4L, anchor lr	3 (3)	1.00	0.16	0.28	0.31
4L, lr 10^{-4}	3 (3)	1.00	0.14	0.32	0.29
8L, anchor lr	3 (3)	1.00	0.28	0.68	0.74
8L, lr 10^{-4}	3 (3)	1.00	0.23	0.71	0.62
8L, lr 10^{-3}	3 (3)	1.00	0.56	0.80	0.85
12L, anchor lr	5 (5)	1.00	0.45	0.84	0.77
12L, lr 10^{-4}	3 (3)	1.00	0.39	0.61	0.87
16L, anchor lr	3 (3)	1.00	0.53	0.67	0.62
16L, lr 10^{-4}	3 (3)	1.00	0.32	0.72	0.88
16L, lr 10^{-3}	3 (3)	1.00	0.36	0.97	0.88
Relative and sinusoidal encodings, depth, lr 10^{-4} *unless stated*					
ALiBi 8L	3 (3)	1.00	0.59	0.47	0.65
ALiBi 8L, anchor lr	3 (3)	1.00	0.36	0.35	0.37
ALiBi 12L	3 (3)	1.00	0.44	0.46	0.46
ALiBi 12L, anchor lr	3 (3)	1.00	0.39	0.35	0.49
ALiBi 16L	3 (3)	1.00	0.36	0.55	0.37
ALiBi 16L, anchor lr	3 (3)	1.00	0.49	0.44	0.62
RoPE 8L	3 (3)	1.00	0.17	0.05	0.29
RoPE 8L, anchor lr	3 (3)	1.00	0.25	0.00	0.39
RoPE 12L	3 (3)	1.00	0.17	0.13	0.28
RoPE 12L, anchor lr	3 (3)	1.00	0.31	0.00	0.47
RoPE 16L	3 (3)	1.00	0.16	0.12	0.33
RoPE 16L, anchor lr	3 (3)	1.00	0.24	0.00	0.39
sinusoidal 8L	3 (3)	1.00	0.18	0.43	0.27
sinusoidal 8L, anchor lr	3 (3)	1.00	0.29	0.46	0.49
sinusoidal 12L	3 (3)	1.00	0.20	0.38	0.28
sinusoidal 12L, anchor lr	3 (3)	1.00	0.28	0.48	0.44
sinusoidal 16L	3 (3)	1.00	0.21	0.34	0.30
sinusoidal 16L, anchor lr	3 (3)	1.00	0.27	0.50	0.49
T5 8L	3 (3)	1.00	0.07	0.62	0.12
T5 8L, anchor lr	3 (3)	1.00	0.12	0.04	0.32
T5 12L	3 (3)	1.00	0.16	0.71	0.59
T5 12L, anchor lr	3 (3)	1.00	0.21	0.09	0.37

cell	runs	ID	Distance	Count	safe
T5 16L	3 (3)	1.00	0.32	0.65	0.63
T5 16L, anchor lr	3 (3)	1.00	0.23	0.40	0.58
RoPE 4L	3 (3)	1.00	0.53	0.05	0.58
RoPE 4L, anchor lr	3 (3)	1.00	0.28	0.00	0.47
ALiBi 4L	3 (3)	1.00	0.43	0.41	0.45
ALiBi 4L, anchor lr	3 (3)	1.00	0.39	0.39	0.69
T5 4L	3 (3)	1.00	0.15	0.51	0.34
T5 4L, anchor lr	3 (3)	1.00	0.08	0.01	0.28
sinusoidal 4L, anchor lr	3 (3)	1.00	0.24	0.29	0.41
RoPE 8L, lr 10^{-3}	3 (3)	1.00	0.39	0.00	0.43
RoPE 12L, lr 10^{-3}	3 (3)	1.00	0.31	0.00	0.40
RoPE 16L, lr 10^{-3}	3 (3)	1.00	0.35	0.00	0.44

Table 18: Recurrent CLOSE cells at matched learning rates, by depth and width. Columns as in Table 17. The anchor-lr depth and width cells of these families are in Table 11.

cell	runs	ID	Distance	Count	safe
Mamba					
1L, lr 10^{-3}	5 (5)	0.96	0.53	0.00	0.60
1L, lr 10^{-2}	3 (3)	0.96	0.87	0.00	0.96
2L, lr 10^{-2}	5 (5)	1.00	0.85	0.00	0.88
4L, lr 10^{-3}	3 (3)	1.00	0.47	0.00	0.49
4L, lr 10^{-2}	3 (3)	1.00	1.00	0.00	1.00
4L, width 256, anchor lr	3 (3)	1.00	0.78	0.00	0.81
4L, width 512, anchor lr	3 (3)	1.00	0.73	0.00	0.79
4L, width 512, lr 10^{-2}	3 (0)	0.41	–	–	–
8L, lr 10^{-2}	5 (1)	0.58	1.00	0.00	1.00
12L, lr 10^{-3}	3 (3)	1.00	0.76	0.00	0.84
12L, lr 10^{-2}	3 (0)	0.41	–	–	–
16L, lr 10^{-3}	3 (3)	1.00	0.64	0.00	0.66
16L, lr 10^{-2}	3 (0)	0.42	–	–	–
LSTM					
1L, lr 5×10^{-3}	3 (3)	1.00	0.90	0.00	0.90
2L, lr 5×10^{-3}	5 (5)	1.00	0.75	0.00	0.77
2L, width 512, lr 5×10^{-3}	3 (3)	1.00	1.00	0.00	1.00
2L, width 256, lr 10^{-3}	3 (3)	1.00	0.45	0.00	0.54
2L, width 256, lr 5×10^{-3}	3 (3)	1.00	0.99	0.00	0.99
2L, width 512, lr 10^{-3}	3 (3)	1.00	0.62	0.00	0.65
4L, lr 10^{-3}	3 (3)	1.00	0.41	0.00	0.42
4L, lr 5×10^{-3}	5 (5)	1.00	0.46	0.00	0.49
4L, lr 10^{-2}	3 (3)	1.00	0.77	0.00	0.81
8L, lr 10^{-3}	3 (3)	1.00	0.28	0.03	0.34
12L, anchor lr	3 (3)	1.00	0.31	0.00	0.34
12L, lr 5×10^{-3}	3 (0)	0.07	–	–	–
16L, anchor lr	3 (0)	0.07	–	–	–
16L, lr 5×10^{-3}	3 (0)	0.07	–	–	–
GRU					
1L, anchor lr	3 (3)	1.00	0.94	0.00	0.95
2L, lr 5×10^{-3}	5 (5)	1.00	0.99	0.00	1.00
4L, anchor lr	3 (3)	1.00	0.49	0.00	0.51
4L, lr 10^{-3}	3 (3)	1.00	0.78	0.01	0.83
4L, lr 5×10^{-3}	3 (3)	1.00	0.96	0.00	0.99
8L, anchor lr	3 (3)	1.00	0.44	0.00	0.45
8L, lr 10^{-3}	3 (3)	1.00	0.30	0.00	0.30
8L, lr 5×10^{-3}	3 (1)	0.38	0.99	0.00	0.99
12L, anchor lr	3 (3)	1.00	0.36	0.00	0.36
12L, lr 5×10^{-3}	3 (0)	0.07	–	–	–
16L, anchor lr	3 (3)	1.00	0.69	0.00	0.84
16L, lr 5×10^{-3}	3 (0)	0.07	–	–	–
2L, width 64	3 (3)	1.00	0.70	0.02	0.87
2L, width 256	3 (3)	1.00	0.92	0.00	0.94
2L, width 512	3 (3)	1.00	0.77	0.00	0.79
2L, width 256, lr 10^{-3}	3 (3)	1.00	0.93	0.01	0.94
2L, width 256, lr 5×10^{-3}	3 (3)	1.00	1.00	0.00	1.00
2L, width 512, lr 10^{-3}	3 (3)	1.00	0.98	0.00	1.00
2L, width 512, lr 5×10^{-3}	3 (3)	1.00	1.00	0.00	1.00
RNN					
1L, anchor lr	3 (3)	1.00	0.94	0.00	0.94
2L, lr 10^{-2}	5 (5)	1.00	0.95	0.00	0.98
4L, anchor lr	3 (3)	1.00	0.94	0.00	0.99
4L, lr 10^{-3}	3 (3)	1.00	0.53	0.00	0.63
8L, anchor lr	3 (3)	1.00	0.53	0.00	0.56

cell	runs	ID	Distance	Count	safe
8L, lr 10^{-3}	3 (3)	1.00	0.87	0.00	0.97
12L, anchor lr	3 (3)	1.00	0.95	0.00	0.98
12L, lr 10^{-3}	3 (2)	0.98	0.91	0.00	0.94
16L, anchor lr	3 (2)	0.67	0.97	0.00	0.99
16L, lr 10^{-3}	3 (0)	0.33	–	–	–
2L, width 64	3 (3)	1.00	0.90	0.00	1.00
2L, width 256	3 (3)	1.00	0.99	0.00	1.00
2L, width 512	3 (3)	1.00	1.00	0.00	1.00

Cells by axis. Table 19 sorts every gate-passing cell by the axes on which its mean closing accuracy reaches 0.5. A reversal of the dissociation would be a cell with a recurrent block that reaches 0.5 on Count alone, or an attention-only cell that reaches it on Distance alone. The first kind does not occur at the cell level in 221 cells. At the run level it occurs once in 798 runs, the top placement of the single-Mamba stacks at 0.59 on Count with 0.02 on Distance (Table 37). The second kind occurs in nine of 132 cells, eight ALiBi cells and the four-layer RoPE cell at lr 10^{-4}, none above 0.62, and ALiBi is the encoding with a built-in distance penalty (Appendix G).

Every other departure from the crossed pattern fails both axes. These are the 59 cells of S4D and Mamba-NoSelect, none above 0.29 on Distance, 35 cells with a recurrent block, and 97 attention-only cells. Twenty-eight of the 35 sit between 0.30 and 0.50 on Distance, most of them at the anchor lr or under a changed task condition, and the seven below are the `AMAM` stack at 0.27, three LSTM condition cells at 0.04 to 0.27, the eight-layer LSTM at lr 10^{-3} at 0.28, and the two single-Mamba stacks with the Mamba layer in the upper half, at 0.02 and 0.16. No attention-only cell of one or two layers reaches 0.5 on Count, 72 cells in all, and the RoPE cells stay at 0.13 or below on Count at every depth.

Three cells carry both axes, the sixteen-layer learned-position Transformer at the anchor lr, at 0.53 and 0.67, the twelve-layer one at lr 10^{-3}, at 0.55 and 0.72, the standard cell of its row in Table 1, and the eight-layer one at lr 10^{-3}, at 0.56 and 0.80.

Table 19: Every gate-passing CLOSE cell of the fixed-budget campaign by the axes on which its mean closing accuracy reaches 0.5, with the runs counted by the same rule in brackets. The attention-only cells that reach 0.5 on Distance alone are ALiBi (8 cells) and RoPE (1 cell), at 0.52 to 0.62.

cells (runs)	all	Distance only	Count only	both	neither
Mamba, LSTM, GRU, RNN and hybrids with a Mamba block	162 (513)	127 (389)	0 (1)	0 (0)	35 (123)
S4D and Mamba-NoSelect	59 (285)	0 (0)	0 (0)	0 (0)	59 (285)
attention only	132 (418)	9 (27)	23 (83)	3 (17)	97 (291)

Table 20: The family rows of Figure 1 over the 1,216 gate-passing CLOSE runs. Distance has gaps scaled $\times 12$, and Count has unseen $\ell = 7$ to 9. Best and mean are closing accuracy over runs, safe the mean safe-close rate, and the last five columns the share of Count items in each error class. Rows pool every cell of the sweep.

		Distance (content length)			Count (safe length)			error classes on the Count split (share of items)				
family	runs	best	mean	safe	best	mean	safe	none	short	inner match	no-stop	long
Mamba, 1–16L	144	1.00	0.75	0.78	0.02	0.00	0.00	0.36	0.40	0.21	0.02	0.00
LSTM, 1–12L	126	1.00	0.55	0.62	0.34	0.02	0.02	0.14	0.76	0.09	0.00	0.00
GRU, 1–16L	65	1.00	0.79	0.82	0.17	0.00	0.00	0.28	0.57	0.14	0.00	0.00
RNN, 1–16L	52	1.00	0.89	0.93	0.00	0.00	0.00	0.42	0.33	0.25	0.00	0.00
Hybrid 4L, 0–2 attn layers	62	1.00	0.75	0.83	0.32	0.01	0.01	0.16	0.71	0.11	0.00	0.00
Hybrid 4L, 3 attn layers	12	0.94	0.44	0.63	0.59	0.26	0.26	0.05	0.56	0.11	0.01	0.00
AAAM	3	0.03	0.02	0.14	0.59	0.45	0.45	0.05	0.40	0.09	0.00	0.00
MAAA	3	0.94	0.88	0.95	0.30	0.22	0.22	0.06	0.61	0.11	0.00	0.00
AAMA	3	0.33	0.16	0.53	0.31	0.11	0.11	0.06	0.68	0.14	0.00	0.00
AMAA	3	0.90	0.71	0.88	0.30	0.25	0.25	0.03	0.55	0.11	0.06	0.00
Hybrid 8L, 0–4 attn layers	34	1.00	0.88	0.91	0.41	0.02	0.02	0.18	0.64	0.15	0.00	0.00
S4D, 1–16L	160	0.29	0.18	0.20	0.21	0.01	0.02	0.16	0.68	0.12	0.01	0.01
Mamba-NoSelect, 2–16L	125	0.41	0.25	0.28	0.01	0.00	0.00	0.33	0.56	0.10	0.01	0.00
Transformer 1–2L, six PEs	226	0.83	0.19	0.31	0.55	0.09	0.09	0.17	0.59	0.13	0.01	0.01
Transformer 4–16L	148	0.86	0.30	0.50	0.99	0.39	0.45	0.05	0.40	0.06	0.02	0.07
learned positions	43	0.86	0.35	0.70	0.99	0.63	0.64	0.04	0.33	0.03	0.00	0.01
sinusoidal	24	0.34	0.22	0.35	0.62	0.41	0.41	0.06	0.43	0.09	0.00	0.01
RoPE	33	0.78	0.29	0.41	0.27	0.03	0.03	0.13	0.72	0.11	0.00	0.00
ALiBi	24	0.69	0.43	0.51	0.64	0.43	0.43	0.04	0.47	0.06	0.00	0.00
T5	24	0.62	0.17	0.40	0.89	0.38	0.72	0.01	0.12	0.05	0.09	0.34
Attention-only NoPE, 4–16L	44	0.93	0.18	0.49	1.00	0.96	0.96	0.00	0.02	0.02	0.00	0.00
Mamba only, hybrid class, 1, 2, 12, 16L	12	0.86	0.54	0.60	0.00	0.00	0.00	0.29	0.55	0.16	0.00	0.00
Mamba variants, 2L	6	0.97	0.76	0.78	0.10	0.02	0.02	0.36	0.43	0.19	0.00	0.00

Notes on Table 20. The table counts the runs with ID accuracy ≥ 0.95 and excludes the budget-ladder runs. Every family is at its best learning rate somewhere in its row, and best is the best single run of the row. None is a zero-length close, an output that emits no fence token and stops at once, and the five error shares and the mean closing accuracy of a row sum to one. The rows cover every run of the pool.

The indented rows split the four-layer stacks with a single Mamba layer by the position of that layer and the deep Transformer row by positional encoding, where RoPE stays at 0.04 on Count against 0.38 to 0.63 for the other four, and the last two rows hold the runs that belong to no family of the text, the Mamba-only stacks of the hybrid class and the two Mamba-variant cells, no convolution and residual position table. The distance axis is the one on which content length acts in LatentMD, and the count axis the one on which the safe-length choice acts. None, short, inner match and no-stop leave the fence unclosed, the family-level counterpart of E2 (Unclosed Fence), and a long close is the counterpart of an over-safe E0 (Correct) (Section 5). Bootstrap intervals over runs are in Appendix K and per-cell values in the released repository.

E SEGMENT DECOMPOSITION

Every segment-split evaluation of the campaign was rescored on 11 and 12 September with the answer-0 items excluded, so that the tables here report closing accuracy like the rest of the paper. The all-items scores are kept in the release. Table 22 gives the six splits of the rescored standard cells of Table 1, the two-layer cells at the anchor lr, the two-layer encodings, the deep learned-position Transformers and the ladder stacks, and Table 23 the VERIFY cells.

Why the Distance split is anchored at twelvefold. The factor of twelve was fixed with the campaign configuration before any model was trained. Figure 6 and Table 29 show what the choice does, in closing accuracy at scales $\times 4$ to $\times 24$ of each segment, and Figure 7 redraws Figure 1 with the Distance axis anchored at $\times 4$ for the 66 cells that have that scale. In Figure 6 the four-layer Mamba at lr 10^{-2} stays at 1.00 on Gaps through $\times 16$ and at 0.95 at $\times 24$, the deep learned-position Transformers fall from above 0.9 at $\times 4$ to below one half at $\times 12$, and the four-layer NoPE stack

falls from 0.54 to 0.06. The factor was chosen beforehand, so the rest of this paragraph describes the choice by its outcome and does not derive it.

Two facts show what the anchor does. First, the attention side's distance limit is graded and moves with depth, so a scale has to be named. At $\times 4$, 29 of the 32 attention cells with a ladder are at or above one half on Gaps, and the deep learned-position cells carry the length at 0.78 to 0.96 while they extrapolate count at 0.61 to 0.84, so seven cells occupy the upper right of Figure 7, the six deep learned-position cells and the eight-layer NoPE stack at 0.74 and 0.99, and the attention side shows no dissociation. At $\times 12$ three attention cells remain at or above one half, the two-layer ALiBi cell at 0.61, the eight-layer ALiBi cell at 0.61 and the sixteen-layer learned-position cell at the anchor lr at 0.54, and the region above 0.6 on both axes is empty.

Second, the recurrent side's count limit does not move with the scale, and its distance score at the lr-matched rate stays near ceiling at $\times 12$, at 0.95 for the two-layer Mamba, 0.99 for the GRU, 1.00 for the RNN and 0.97 for the alternating hybrid, with the LSTM at 0.76.

Twelvefold is therefore the smallest scale of the ladder at which fewer than a tenth of the attention cells still carry the length while the lr-matched recurrent cells still do, that is, the smallest scale at which both limits are visible on one split. At $\times 16$ and $\times 24$ the anchor lr recurrent cells fall as well, the two-layer Mamba at the anchor lr to 0.75 and 0.70 and the pure four-layer stack of the ladder class to 0.33 and 0.23, while the four-layer Mamba at lr 10^{-2} keeps 0.95 and the alternating hybrid 0.93 at $\times 24$. Table 21 lists the standard cells of Table 1 at the three scales. At $\times 24$ five of the six nonlinear or selective rows stay above every attention-only row, at 0.79 to 0.99, and the LSTM falls to 0.50, below the two-layer ALiBi cell at 0.57. The $\times 12$ column of that table is the ladder's own draw of 2,000 items and can differ from Table 1 in the second decimal.

The dissociation of Section 3 is a statement at the twelvefold scale. At $\times 4$ it is one-sided, recurrence carries distance only and deep attention carries both, which is the reading of Section 3 that the attention side's limit is a boundary that depth moves and the recurrent side's a ceiling that only the removal of recurrent layers lifts (Table 37). At $\times 24$ the learned-position Transformers have no point, because some of their stretched sequences exceed the 1,024-position table, and the two-layer ALiBi cell is the only attention cell left above one half, at 0.57.

Table 21: The distance axis at three gap scales for the standard cells of Table 1. Closing accuracy on Gaps at $\times 4$, $\times 12$ and $\times 24$ from the scale ladder, 2,000 items per scale, mean over gate-passing seeds, with the rank of the row in brackets. A dash marks a learned-position cell whose stretched sequences exceed its 1,024-position table.

model	layers	lr	seeds	$\times 4$	$\times 12$	$\times 24$
Mamba	2	10^{-3}	5	1.00 (1)	0.95 (5)	0.79 (5)
LSTM	2	5×10^{-3}	5	0.97 (7)	0.76 (6)	0.50 (7)
GRU	2	5×10^{-3}	5	1.00 (1)	0.99 (2)	0.96 (2)
RNN	2	10^{-3}	5	1.00 (1)	1.00 (1)	0.99 (1)
Hybrid MAMA	4	3×10^{-4}	5	1.00 (1)	0.97 (3)	0.93 (3)
Hybrid MAMAMAMA	8	10^{-3}	5	1.00 (1)	0.97 (3)	0.88 (4)
S4D	2	10^{-2}	5	0.51 (12)	0.29 (10)	0.27 (8)
Mamba-NoSelect	2	3×10^{-4}	5	0.45 (13)	0.19 (11)	0.14 (10)
Transformer	2	5×10^{-3}	5	0.73 (9)	0.19 (11)	–
Transformer+ALiBi	2	10^{-4}	5	0.73 (9)	0.61 (7)	0.57 (6)
Transformer	12	10^{-3}	5	0.98 (6)	0.54 (8)	–
Attention-only NoPE	4	10^{-4}	5	0.58 (11)	0.10 (13)	0.02 (11)
Attention-only NoPE	8	10^{-2}	5	0.83 (8)	0.36 (9)	0.18 (9)

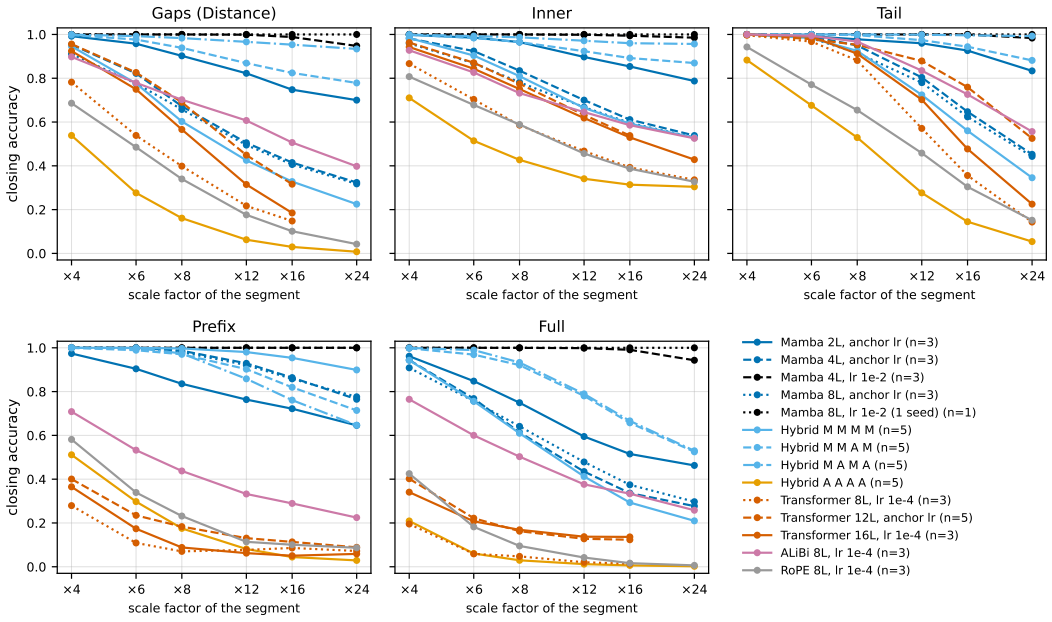


Figure 6: Closing accuracy on Gaps, Inner, Tail, Prefix and Full for the deep cells and the attention-layer ladder, closing accuracy from $\times 4$ to $\times 24$, mean over gate-passing seeds. The four-layer Mamba at lr 10^{-2} is the black dashed line. The deep learned-position Transformers have no $\times 24$ point because their stretched sequences exceed the position table.

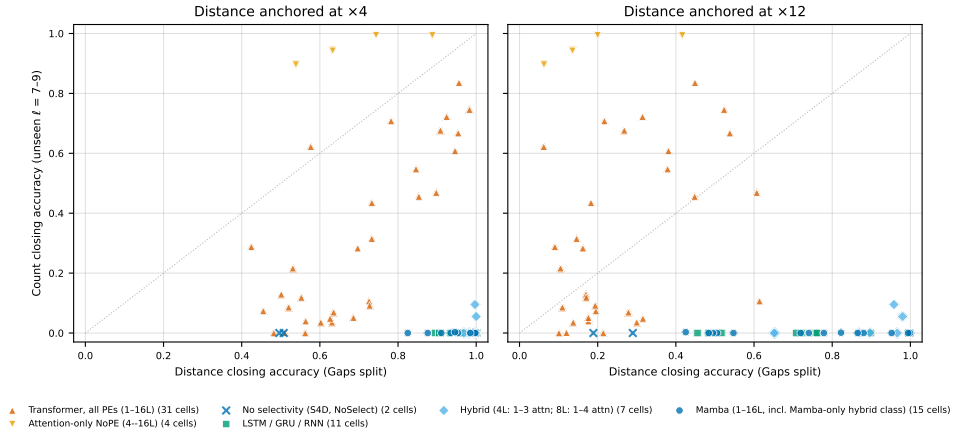


Figure 7: Figure 1 with the Distance axis anchored at $\times 4$ (left) and at $\times 12$ (right), cell means over gate-passing seeds for the 66 cells with a $\times 4$ closing-accuracy file. At $\times 4$ the deep learned-position Transformers sit in the upper right, so the attention side shows no dissociation; at $\times 12$ that region is empty. The recurrent cells stay on the horizontal axis at either scale.

Table 22: Six segment splits of CLOSE, closing accuracy per split, mean over gate-passing runs (runs column), for the rescored standard cells of Table 1, the anchor two-layer cells and encodings, the deep learned-position cells and the ladder stacks. Prefix and Pad read against ID, the rest against Distance, the last two columns the already-closed accuracy of Tail and Inner.

cell	runs	prefix	inner	tail	pad	gaps	full	ans-0 tail	ans-0 inner
Mamba-NoSelect, 2L, anchor lr	5	0.39	0.44	0.24	0.20	0.19	0.14	0.11	0.65
Mamba, 2L, anchor lr	3	0.76	0.90	0.96	0.26	0.82	0.60	0.72	0.94
Mamba, 2L, lr 10^{-3}	5	1.00	0.98	1.00	0.41	0.95	0.95	0.67	0.99
Mamba, 2L, lr 10^{-2}	5	1.00	0.95	0.95	0.75	0.86	0.86	0.94	0.95
Mamba, 4L, anchor lr	3	0.93	0.70	0.80	0.62	0.50	0.44	0.19	0.90
Mamba, 8L, anchor lr	3	0.92	0.66	0.78	0.48	0.49	0.47	0.00	0.79
Mamba, 12L, lr 10^{-3}	3	0.98	0.88	0.95	0.77	0.76	0.72	0.00	0.83
Mamba, 16L, lr 10^{-3}	3	0.74	0.76	0.94	0.55	0.64	0.53	0.00	0.86
LSTM, 2L, anchor lr	3	0.61	0.82	0.93	0.16	0.72	0.42	0.14	0.99
LSTM, 2L, lr 5×10^{-3}	5	1.00	0.88	0.93	1.00	0.76	0.75	0.90	0.98
GRU, 2L, lr 5×10^{-3}	5	1.00	0.99	1.00	1.00	0.99	0.99	0.99	1.00
GRU, 2L, anchor lr	5	0.98	0.91	0.91	0.81	0.82	0.79	0.57	0.94
RNN, 2L, lr 10^{-3}	5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
RNN, 2L, lr 10^{-2}	5	0.68	0.97	0.96	0.67	0.94	0.65	1.00	0.99
RNN, 2L, anchor lr	5	0.81	0.95	1.00	0.73	0.95	0.78	0.61	1.00
S4D, 2L, lr 10^{-3}	5	0.36	0.39	0.16	0.22	0.14	0.13	0.05	0.66
S4D, 2L, lr 10^{-2}	5	0.26	0.46	0.29	0.35	0.28	0.25	0.03	0.83
Transformer+ALiBi, 2L, lr 10^{-4}	5	0.30	0.68	0.58	0.19	0.62	0.34	0.03	0.97
Transformer+RoPE, 2L, lr 10^{-4}	3	0.62	0.44	0.27	0.00	0.11	0.02	0.01	0.26
Transformer+sinusoidal, 2L, anchor lr	5	0.30	0.34	0.25	0.00	0.12	0.04	0.14	0.45
Transformer+T5, 2L, anchor lr	3	0.22	0.42	0.39	0.00	0.17	0.04	0.20	0.10
Transformer, 2L, lr 10^{-4}	5	0.07	0.35	0.24	0.00	0.11	0.01	0.24	0.85
Transformer, 2L, anchor lr	3	0.18	0.38	0.34	0.00	0.14	0.03	0.44	0.37
Transformer, 2L, lr 5×10^{-3}	5	0.37	0.45	0.47	0.00	0.20	0.04	0.64	0.20
Transformer, 4L, anchor lr	3	0.19	0.41	0.40	0.00	0.16	0.02	0.72	0.52
Transformer, 8L, anchor lr	3	0.09	0.67	0.63	0.00	0.27	0.14	1.00	0.87
Transformer, 8L, lr 10^{-4}	3	0.08	0.47	0.57	0.00	0.22	0.02	0.78	0.59
Transformer, 12L, lr 10^{-3}	5	0.12	0.73	0.90	0.00	0.54	0.13	1.00	0.93
Transformer, 12L, anchor lr	5	0.14	0.63	0.88	0.04	0.45	0.13	0.99	0.70
Transformer, 16L, lr 10^{-4}	3	0.06	0.62	0.70	0.04	0.32	0.14	0.95	0.73
Hybrid A A A A, 4L, lr 10^{-4}	5	0.13	0.36	0.32	0.00	0.10	0.00	0.19	0.64
Hybrid M M M M, 4L, anchor lr	5	0.98	0.66	0.72	0.74	0.43	0.41	0.52	0.79
Hybrid A A A A, 4L, anchor lr	5	0.08	0.34	0.28	0.00	0.06	0.01	0.74	0.42
Hybrid M A M A, 4L, anchor lr	5	0.86	0.97	1.00	0.54	0.97	0.79	1.00	0.98
Hybrid M M A M, 4L, anchor lr	5	0.91	0.92	0.98	0.59	0.87	0.78	0.55	0.89
Hybrid A A A A A A A A, 8L, anchor lr	5	0.32	0.45	0.64	0.00	0.20	0.10	0.48	0.67
Hybrid A A A A A A A A, 8L, lr 10^{-2}	5	0.48	0.58	0.58	0.00	0.36	0.22	0.87	0.57
Hybrid M A M A M A M A, 8L, anchor lr	5	1.00	0.97	0.99	0.39	0.96	0.95	1.00	1.00
Hybrid M A M A M A M A, 8L, lr 10^{-3}	5	1.00	0.98	1.00	0.86	0.97	0.97	0.78	1.00

Table 23: VERIFY on the standard splits and the six segment splits, accuracy (chance 0.5), mean over gate-passing runs. Rows at the anchor lr unless stated, width 128.

cell	runs	ID	Distance	Count	prefix	inner	tail	pad	gaps	full
LSTM 2L, anchor lr	3 (3)	1.00	0.80	0.78	0.61	0.85	0.96	0.60	0.80	0.66
LSTM 2L, lr 10^{-2}	3 (3)	1.00	0.91	0.70	1.00	0.93	1.00	0.99	0.91	0.91
LSTM 1L	3 (3)	1.00	0.92	0.82	0.69	0.95	0.98	0.62	0.92	0.73
LSTM 4L	3 (3)	1.00	0.85	0.77	0.73	0.90	0.95	0.74	0.84	0.71
LSTM 8L	3 (3)	1.00	0.85	0.75	0.59	0.91	0.92	0.50	0.85	0.84
Mamba 2L, anchor lr	3 (3)	1.00	0.94	0.66	0.86	0.97	1.00	0.92	0.94	0.93
Mamba 2L, lr 10^{-3}	3 (3)	1.00	0.96	0.57	1.00	0.97	1.00	0.94	0.96	0.96
Mamba 1L	3 (0)	0.94	–	–	–	–	–	–	–	–
Mamba 4L	3 (3)	1.00	0.83	0.67	0.98	0.88	0.98	0.85	0.83	0.83
Mamba 8L	3 (3)	1.00	0.85	0.71	0.97	0.88	0.98	0.75	0.85	0.85
Transformer 2L, anchor lr	3 (3)	1.00	0.50	0.75	0.51	0.53	0.56	0.50	0.49	0.50
Transformer 1L	3 (3)	1.00	0.53	0.73	0.50	0.51	0.55	0.50	0.53	0.50
Transformer 4L	3 (3)	1.00	0.37	0.94	0.64	0.46	0.61	0.50	0.37	0.42
Transformer 8L	3 (3)	1.00	0.67	0.98	0.48	0.76	0.90	0.49	0.66	0.52
Transformer+ALiBi 2L	3 (3)	0.99	0.78	0.58	0.70	0.81	0.87	0.78	0.78	0.74
Transformer+RoPE 2L	3 (3)	1.00	0.57	0.62	0.73	0.64	0.81	0.56	0.56	0.57
Hybrid M M M M	3 (3)	1.00	0.91	0.65	0.99	0.92	0.99	0.95	0.91	0.90
Hybrid M M A M	3 (3)	1.00	0.92	0.74	0.89	0.94	0.99	0.78	0.92	0.86
Hybrid M A M A	3 (3)	1.00	0.97	0.62	0.95	0.98	1.00	0.77	0.97	0.91
Hybrid A A A A	3 (3)	1.00	0.50	1.00	0.62	0.54	0.77	0.18	0.49	0.34
Hybrid M M M M, lr 10^{-3}	3 (3)	1.00	0.89	0.71	0.88	0.90	0.98	0.83	0.88	0.79
Hybrid M A M A, lr 10^{-3}	3 (3)	1.00	0.85	0.70	0.94	0.92	0.92	0.84	0.85	0.80

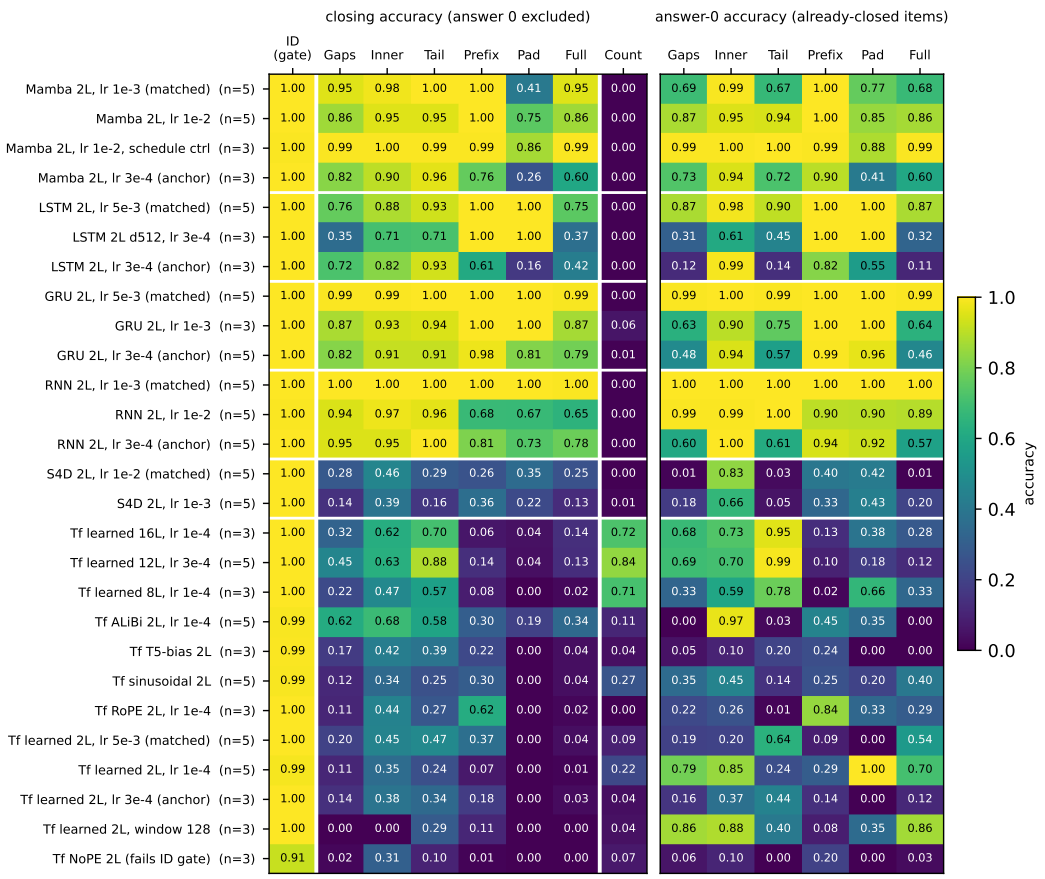


Figure 8: Segment signatures of the CLOSE cells discussed in the text, mean over seeds. Left, closing accuracy on the six segment splits and on Count, with ID accuracy over all items as the gate. Right, accuracy on the already-closed items. Prefix and Pad are read against ID, the other splits against Distance, and rows tagged matched are the lr-matched cells.

Segment signatures (detail for Section 4.1). In Figure 8 Gaps is the Distance split, and the rows are cells at the anchor lr and at their lr-matched rate and Transformers of each positional encoding and depth at their best learning rate. Mamba and the RNN at lr 10^{-2}, S4D at 10^{-3} and the two-layer Transformer at 10^{-4} keep a row because analyses of Section 4 name those cells.

After a stretch of filler that follows the closing run, an under-tuned recurrent model closes again. Mamba at 12 and 16 layers shows the same loss at its largest, with closing accuracy on Tail of 0.95 and 0.94 and no already-closed item answered correctly. The training distribution that removes the prefix collapse in Section 4.2 restores the anchor Mamba's Pad. The deep Transformer prefix collapse is more than half complete at a fourfold scale, where the twelve- and sixteen-layer learned-position cells sit at 0.31 to 0.47 in closing accuracy (Table 30), and deep relative-position Transformers collapse on Prefix and Pad as well, in a graded rather than threshold-like way.

In VERIFY depth revives only the distance component of the learned-position Transformer, with eight-layer Inner at 0.76 and Tail at 0.90 while Prefix and Pad stay at 0.49, and among two-layer Transformers only the distance-biased encodings verify at all, ALiBi at 0.78 and RoPE at 0.57.

F ERROR TAXONOMY

Table 24 gives the no-stop share and the inner-match share by segment split, Table 25 the count-axis classes, and Table 26 the OPEN detail under content scales up to 48. Two signatures are visible

only here. The no-stop error of the two-layer learned-position cell at the anchor lr is a Pad and distance signature, 1.00 on Pad and 0.70 on Gaps against 0.01 on Prefix, and the inner-match share of the deep learned-position cells is a Prefix signature, 0.34 of wrong outputs at twelve layers against 0.15 or less on every other split (Table 24). The sixteen-layer ALiBi and RoPE cells close short in 0.43 and 0.42 of items, the largest short-close shares among the deep cells, and the attention-only four-layer stack never stops in 0.56 of items.

On Count the safe-close rate stays within 0.05 of closing accuracy in all but 38 of the 1,216 runs, and 26 of those are T5 Transformers. Their cells of four to sixteen layers count and then close long, at a safe-close rate of 0.82 to 0.94 against a closing accuracy of 0.51 to 0.71 at lr 10^{-4} and of 0.36 to 0.80 against 0.01 to 0.40 at the anchor lr.

The none column is the zero-length close of Section 3, scored on every gate-passing run with the same generator, 2,000 items per split, and subtracted from the short column of the taxonomy, so the run-level shares are consistent to within 0.03. On Distance it is an attention-side class, 0.60 of items for the two-layer learned-position cell at lr 10^{-4} and 0.00 to 0.01 for the recurrent anchors. On Count it is a recurrent-side class, 0.14 to 0.42 of items by family, and the eight-layer NoPE stack never produces it.

On ID every cell is at 0.98 or above at every margin. On Distance the recurrent cells move by a few points in either direction between margin 1 and margin 2, Mamba at the anchor lr from 0.71 to 0.81 and at lr 10^{-2} from 0.84 to 0.82, the attention cells sit near zero at every margin except ALiBi at 0.63 to 0.88, and the items without an inner run score highest in every cell. The cost of an inner run is its presence and not its margin, and margin 0 has no open instance in the suite, since a run that reaches ℓ closes the fence.

Table 24: Two error signatures by segment split, mean over gate-passing runs. The no-stop share is the fraction of open items whose output never emits STOP, and the inner-match share is the fraction of wrong outputs that close at the length of an inner run of the item. Full is omitted; it tracks Gaps.

		no-stop share					inner-match share among wrong outputs				
cell	runs	prefix	inner	tail	pad	gaps	prefix	inner	tail	pad	gaps
Mamba 2L, anchor lr	3	0.00	0.00	0.00	0.00	0.00	0.07	0.36	0.21	0.26	0.45
LSTM 2L, anchor lr	3	0.00	0.00	0.00	0.00	0.00	0.26	0.25	0.38	0.24	0.32
Transformer 2L, anchor lr	3	0.01	0.43	0.39	1.00	0.70	0.21	0.01	0.05	0.00	0.00
Transformer 2L, lr 10^{-4}	5	0.02	0.04	0.11	0.00	0.12	0.22	0.04	0.11	0.00	0.03
Transformer+ALiBi 2L, lr 10^{-4}	5	0.00	0.00	0.02	0.00	0.01	0.23	0.11	0.21	0.08	0.20
Transformer+RoPE 2L, lr 10^{-4}	3	0.01	0.25	0.14	0.67	0.48	0.13	0.06	0.18	0.00	0.01
Transformer 12L	5	0.01	0.01	0.00	0.66	0.04	0.34	0.15	0.02	0.01	0.06
Transformer 16L, lr 10^{-4}	3	0.00	0.01	0.00	0.28	0.03	0.34	0.15	0.00	0.05	0.03
Hybrid M M M M	5	0.00	0.00	0.04	0.00	0.09	0.12	0.24	0.22	0.24	0.25
Hybrid M A M A	5	0.00	0.00	0.00	0.00	0.00	0.26	0.17	0.13	0.22	0.25

Table 25: Error classes of CLOSE on the Count split, share of open items, mean over gate-passing runs, for the rescored standard cells of Table 1, the two-layer families at the anchor lr, the two-layer encodings, the deep Transformers and the ladder stacks. None is the zero-length close as in Table 27. Family means over all gate-passing runs are the last five columns of Table 20.

cell	runs	exact	none	short	inner match	no-stop	long
Mamba 2L, anchor lr	3	0.00	0.47	0.33	0.20	0.00	0.00
Mamba 2L, lr 10^{-3}	5	0.00	0.44	0.21	0.35	0.01	0.00
Mamba 2L, lr 10^{-2}	5	0.00	0.32	0.33	0.28	0.07	0.00
LSTM 2L, anchor lr	3	0.00	0.15	0.77	0.09	0.00	0.00
LSTM 2L, lr 5×10^{-3}	5	0.00	0.21	0.70	0.09	0.00	0.00
GRU 2L, anchor lr	5	0.00	0.46	0.28	0.25	0.00	0.00
GRU 2L, lr 5×10^{-3}	5	0.00	0.19	0.75	0.07	0.00	0.00
RNN 2L, anchor lr	5	0.00	0.38	0.42	0.19	0.00	0.00
RNN 2L, lr 10^{-3}	5	0.00	0.44	0.30	0.26	0.00	0.00
S4D 2L, lr 10^{-3}	5	0.01	0.13	0.65	0.14	0.03	0.03
S4D 2L, lr 10^{-2}	5	0.00	0.14	0.75	0.11	0.00	0.00
Transformer 2L, anchor lr	3	0.03	0.11	0.74	0.12	0.00	0.00
Transformer 2L, lr 10^{-4}	5	0.21	0.10	0.57	0.10	0.00	0.01
Transformer 2L, lr 5×10^{-3}	5	0.09	0.13	0.60	0.14	0.01	0.02
Transformer+sinusoidal 2L	5	0.26	0.15	0.47	0.12	0.00	0.00
Transformer+ALiBi 2L, lr 10^{-4}	5	0.11	0.15	0.56	0.18	0.00	0.00
Transformer+RoPE 2L, lr 10^{-4}	3	0.00	0.29	0.60	0.11	0.00	0.00
Transformer+T5 2L	3	0.04	0.18	0.60	0.11	0.01	0.06
Transformer 4L	3	0.28	0.10	0.54	0.07	0.00	0.02
Transformer 8L, lr 10^{-4}	3	0.71	0.02	0.24	0.02	0.00	0.00
Transformer 12L	5	0.83	0.01	0.14	0.01	0.00	0.00
Transformer 16L, lr 10^{-4}	3	0.72	0.03	0.23	0.01	0.00	0.00
Transformer+ALiBi 8L	3	0.47	0.03	0.43	0.06	0.00	0.00
Transformer+ALiBi 16L	3	0.55	0.03	0.37	0.04	0.00	0.00
Transformer+RoPE 8L	3	0.05	0.06	0.77	0.11	0.00	0.00
Transformer+RoPE 16L	3	0.12	0.07	0.71	0.10	0.00	0.01
Hybrid M M M M	5	0.00	0.16	0.75	0.09	0.00	0.00
Hybrid M A M A	5	0.00	0.14	0.76	0.10	0.00	0.00
Hybrid A A A A	5	0.90	0.00	0.07	0.02	0.00	0.00

Table 26: OPEN detail. ID and Distance accuracy of the opening choice, then at content-gap scales 12, 24 and 48 the safe-length failure rate (an opening length at or below the longest inner run) and the share of items opened exactly at the inner maximum (eq.). Means over seeds at the anchor lr unless stated; the two-layer Transformer uses ALiBi.

				×12		×24		×48	
cell	runs	ID	Distance	failure	eq.	failure	eq.	failure	eq.
LSTM 2L	3 (3)	1.00	1.00	0.00	0.00	–	–	–	–
Mamba 2L	3 (3)	1.00	1.00	0.00	0.00	–	–	–	–
Transformer+ALiBi 2L	3 (3)	1.00	0.60	0.28	0.21	–	–	–	–
Hybrid M M M M	3 (3)	1.00	0.96	0.03	0.03	0.15	0.14	0.32	0.26
Hybrid M M A M	3 (3)	1.00	1.00	0.00	0.00	0.00	0.00	0.03	0.03
Hybrid M A M A	3 (3)	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Hybrid A A A A	3 (3)	1.00	0.77	0.22	0.12	0.32	0.16	0.41	0.19
Hybrid M M M M, lr 10^{-3}	3 (3)	1.00	0.93	0.07	0.06	0.21	0.15	0.37	0.20
Hybrid M A M A, lr 10^{-3}	3 (3)	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 27: Error classes of CLOSE on the Distance split, closing accuracy and then the share of items in each class, mean over seeds. The recurrent rows are anchor lr cells.

CLOSE cell	exact	inner match	none	short	no-stop	long
Mamba, anchor lr	0.81	0.07	0.01	0.07	0.00	0.04
LSTM, anchor lr	0.72	0.09	0.00	0.15	0.00	0.04
Transformer 2L, anchor lr	0.13	0.00	0.06	0.05	0.70	0.05
Transformer 2L, lr 10^{-4}	0.11	0.03	0.60	0.09	0.12	0.06
Transformer+ALiBi 2L, lr 10^{-4}	0.62	0.08	0.00	0.17	0.01	0.11
Transformer+RoPE 2L, lr 10^{-4}	0.11	0.01	0.18	0.08	0.49	0.12
Transformer+T5 2L	0.17	0.02	0.02	0.08	0.47	0.24
Transformer 8L, lr 10^{-4}	0.23	0.02	0.07	0.02	0.27	0.39
Transformer 12L	0.45	0.03	0.10	0.06	0.04	0.32
Transformer 16L, lr 10^{-4}	0.33	0.02	0.04	0.03	0.03	0.56
Transformer+ALiBi 8L	0.59	0.08	0.01	0.25	0.00	0.06

Error classes (detail for Section 4.1). Table 27 gives the fingerprint of each family at the anchor lr for the recurrent rows and at each encoding's best rate for the attention rows. The none column and the LatentMD counterpart of each class are as in the notes on Table 20 (Appendix D). Recurrent models close at the length of another run in 0.07 of items for Mamba and 0.09 for the LSTM. The zero-length close is the one class the recurrent models do not show on this split, at most 0.01 of items. On the Prefix split deep Transformers close at an inner run's length instead of failing to stop (Table 24).

F.1 THE SYNTHETIC ERROR CLASSES READ THROUGH THE LATENTMD FORMAT METRICS

LatentMD scores a generation with three format metrics (Lee & Kang, 2026). F1 (Instruction Compliance) asks whether the response follows the wrapping directive of the prompt. F2 (Fence Balance) walks the fence lines with a stack, where a bare fence of the same family and of at least the opening length closes the current block, and passes when the stack is empty at the end of the response. F3 (Safe Length) asks whether the outer fence is longer than every same-family run inside it, either a priori, when the prompt states the inner run, or post hoc, from the realized content. The four labels follow from F2 (Fence Balance) and F3 (Safe Length) alone. E0 (Correct) is F2 $\wedge$ F3, E1 (Premature Closure) is $\neg$F2 with an interior bare fence at least as long as the outer one, E2 (Unclosed Fence) is $\neg$F2 without such a fence, and E3 (Latent Collision) is F2 $\wedge \neg$F3.

A synthetic item and its output become a Markdown document by one reading. A run of n fence tokens is a bare fence line of n backticks, filler is content, and STOP is the end of the file. The suite has no info strings, so no run inside an open fence opens a nested block, and the stack of F2 (Fence Balance) reduces to the rule of Section 2. A run shorter than the open length is content and a run at least as long closes the fence.

Three consequences follow. F1 (Instruction Compliance) has no counterpart, because no item carries a wrapping directive. F3 (Safe Length) post hoc passes by construction in CLOSE and in the safe items of VERIFY, because every inner run is shorter than ℓ. And E3 (Latent Collision) has no instance in any task, because a bare run that reaches ℓ always closes the fence, so a collision is never masked by balance. Figure 9 shows the verdict of each metric on one worked item per task, with dark squares for fence tokens, light squares for filler, P for the probe, $\dashv$ for the stop token and $\cdots$ for a run that never stops, and Figure 13 in Appendix J draws the three minimal LatentMD examples next to their synthetic counterparts.

Two classes need a word. The inner-length match of CLOSE is E2 (Unclosed Fence) under the metrics, since the emitted run is shorter than ℓ and the fence stays open. The paper places it in the collision family of Table 39 because of its cause, a model that closes at the length of an inner run has confused the Wrapper with a Content run, which is the deficit behind E1 (Premature Closure) and E3 (Latent Collision). That placement is by deficit and not by label. The second class is the answer-0 error of Section 4, a model that emits a run after a fence has already closed. The run opens a new fence that never closes, so F2 (Fence Balance) fails, and because the earlier closing run is an interior bare match, the mechanical LatentMD rule labels the document E1 (Premature Closure).

The deficit is a lost record of the fence state, and Table 39 carries it as its own row, placed in neither family.

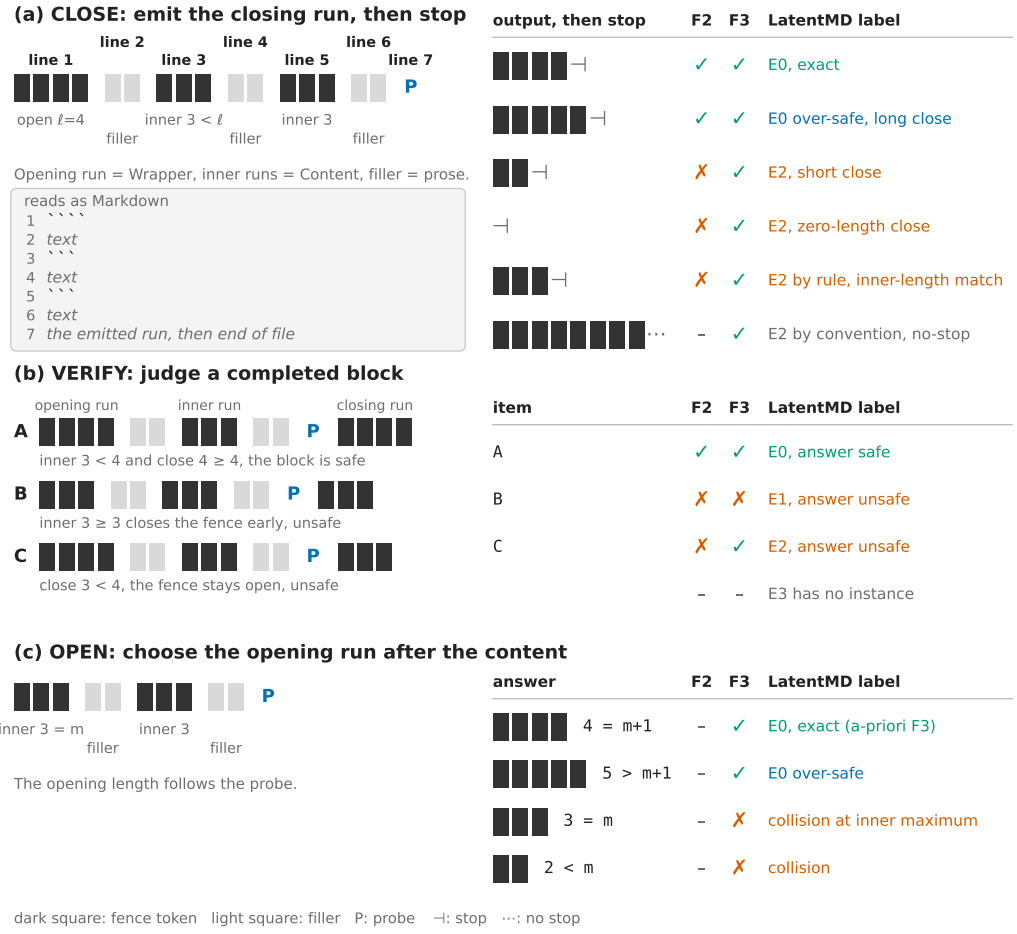


Figure 9: One worked item per task read as a Markdown document and scored with F2 (Fence Balance) and F3 (Safe Length). (a) A CLOSE item, $\ell = 4$, with its six output classes. (b) The three VERIFY item types, answered by the F2 $\wedge$ F3 verdict. (c) An OPEN item with $m = 3$ and four answers.

G POSITIONAL ENCODING AND MECHANISM DETAIL

Detail behind Section 4.2. The worst-head column of Table 28 is the base accuracy after zeroing the single most harmful head. In the position-shift experiment, restoring the opening run, the prompt or the readout position alone makes the shifted model worse than no restoration, because a partial restoration mixes two coordinate frames. A linear readout probe at the probe position, fitted on one seed per model in August, is reported here and supports no claim of the paper. It found that a collapsed two-layer Transformer still carries the count in its representation, since a probe fitted on Distance items recovers the answer at 0.81 while a probe fitted on ID items transfers to Distance below the majority baseline, and that the same transfer succeeds for the recurrent models in proportion to their task score, from 0.52 to 0.91. The reading is that the readout direction of the attention model moves with the distribution while the representation keeps the information. The probe is single-seed, single-position and linear, and Mamba's residual stream needs a standardized ridge before the probe fits at all.

Table 28: Where the far-distance failure sits. Clean is the teacher-forced closing-token accuracy at training gaps, corrupt the same items with gaps $\times 12$. Recovery is the fraction of the lost accuracy regained when one site is patched from the clean run, the two sites of Section 4. Two-layer cells unless stated, the 8-layer cells at lr 10^{-4}, mean over seeds.

cell	clean $\rightarrow$ corrupt	attn. output	layer-0 resid.	worst head	verdict
RoPE 2L, lr 3×10^{-4}	$1.00 \rightarrow 0.11$	$+0.92$	$+0.40$	0.28	retrieval
RoPE 2L, lr 10^{-4}	$1.00 \rightarrow 0.11$	$+0.75$	$+0.25$	0.28	retrieval
ALiBi 2L, lr 3×10^{-4}	$0.99 \rightarrow 0.28$	$+0.69$	$+0.24$	0.04	retrieval (one layer-0 head)
ALiBi 2L, lr 10^{-4}	$0.98 \rightarrow 0.60$	-0.24	-0.21	0.00	partly collapsed, no single site
learned 2L, lr 3×10^{-4}	$1.00 \rightarrow 0.15$	$+0.01$	$+0.64$	0.33	readout position
learned 2L, lr 10^{-4}	$0.98 \rightarrow 0.09$	$+0.13$	$+0.49$	0.26	readout position
sinusoidal 2L	$0.99 \rightarrow 0.12$	-0.02	$+0.07$	0.31	no single site
T5 2L	$0.99 \rightarrow 0.18$	-0.16	0.00	0.15	no single site (two heads)
learned 8L	$1.00 \rightarrow 0.20$	$+0.03$	-0.10	0.64	no single site
ALiBi 8L	$1.00 \rightarrow 0.59$	-0.09	-0.28	0.12	no single site (one head)
RoPE 8L	$1.00 \rightarrow 0.19$	$+0.02$	-0.12	0.80	no single site

Tables 29, 30 and 31 give the dense gap and prefix sweeps at nine scales and the inner-run sweep in closing accuracy, the first two for the standard cells of Table 1 and the cells named in the text and the third for every cell of the 15 September rescoring, and Table 32 the free-running rescoring of the readout patches of Table 28.

Table 29: Dense gap sweep, closing accuracy on the Gaps split at nine scale factors, 2,000 items per scale, mean over gate-passing seeds, for the standard cells of Table 1 and the cells named in the text. Learned-position cells have no value at $\times 24$ when the stretched sequences exceed the 1,024-position table. Gaps at $\times 12$ is the Distance split.

cell	runs	$\times 1$	$\times 2$	$\times 3$	$\times 4$	$\times 6$	$\times 8$	$\times 12$	$\times 16$	$\times 24$
2 layers										
LSTM, 2L, lr 5×10^{-3}	5	–	–	–	0.97	–	–	0.76	–	0.50
GRU, 2L, lr 5×10^{-3}	5	–	–	–	1.00	–	–	0.99	–	0.96
RNN, 2L, lr 10^{-3}	5	–	–	–	1.00	–	–	1.00	–	0.99
Mamba, 2L, lr 10^{-3}	5	1.00	1.00	1.00	1.00	1.00	0.99	0.95	0.89	0.79
Mamba-NoSelect, 2L, anchor lr	5	–	–	–	0.45	–	–	0.19	–	0.14
Transformer+ALiBi, 2L, lr 10^{-4}	5	0.98	0.92	0.81	0.73	0.67	0.64	0.61	0.60	0.57
Transformer+RoPE, 2L, anchor lr	3	1.00	0.87	0.68	0.51	0.34	0.22	0.12	0.07	0.02
Transformer+RoPE, 2L, lr 10^{-4}	3	1.00	0.89	0.66	0.48	0.30	0.21	0.10	0.06	0.03
Transformer, 2L, anchor lr	3	1.00	0.91	0.75	0.60	0.40	0.27	0.14	0.08	–
Transformer, 2L, lr 5×10^{-3}	5	1.00	0.94	0.85	0.73	0.49	0.37	0.19	0.13	–
S4D, 2L, lr 10^{-2}	5	–	–	–	0.51	–	–	0.29	–	0.27
4 layers										
Mamba, 4L, lr 10^{-2}	3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.95
Hybrid A A A A, 4L, anchor lr	5	1.00	0.96	0.76	0.54	0.28	0.16	0.06	0.03	0.01
Hybrid A A A A, 4L, lr 10^{-4}	5	–	–	–	0.58	–	–	0.10	–	0.02
Hybrid A M M M, 4L, anchor lr	3	1.00	1.00	1.00	0.98	0.90	0.81	0.65	0.53	0.38
Hybrid M A M A, 4L, anchor lr	5	1.00	1.00	1.00	1.00	0.99	0.98	0.97	0.95	0.93
Hybrid M M A M, 4L, anchor lr	5	1.00	1.00	1.00	1.00	0.98	0.94	0.87	0.82	0.78
Hybrid M M M A, 4L, anchor lr	3	1.00	1.00	1.00	0.97	0.86	0.77	0.65	0.60	0.55
Hybrid M M M M, 4L, anchor lr	5	1.00	1.00	0.99	0.95	0.78	0.60	0.43	0.33	0.23
Hybrid M M M M, 4L, lr 10^{-3}	3	1.00	1.00	1.00	0.98	0.94	0.87	0.72	0.61	0.48
Transformer, 4L, anchor lr	3	1.00	0.97	0.88	0.70	0.45	0.32	0.16	0.10	–
8 layers										
Hybrid A A A A A A A A, 8L, lr 10^{-2}	5	–	–	–	0.83	–	–	0.36	–	0.18
Hybrid M A M A M A M A, 8L, lr 10^{-3}	5	–	–	–	1.00	–	–	0.97	–	0.88
Transformer+ALiBi, 8L, lr 10^{-4}	3	1.00	1.00	0.97	0.90	0.78	0.70	0.61	0.51	0.40
Transformer+RoPE, 8L, lr 10^{-4}	3	1.00	0.96	0.83	0.69	0.49	0.34	0.18	0.10	0.04
Transformer, 8L, anchor lr	3	1.00	1.00	0.97	0.91	0.72	0.54	0.27	0.15	–
Transformer, 8L, lr 10^{-4}	3	1.00	0.99	0.92	0.78	0.54	0.40	0.22	0.15	–
12 layers										
Transformer+ALiBi, 12L, lr 10^{-4}	3	1.00	1.00	0.95	0.85	0.70	0.57	0.45	0.36	0.27
Transformer+RoPE, 12L, lr 10^{-4}	3	1.00	0.93	0.70	0.50	0.32	0.23	0.17	0.13	0.10
Transformer, 12L, anchor lr	5	1.00	1.00	0.99	0.96	0.83	0.69	0.45	0.32	–
Transformer, 12L, lr 10^{-3}	5	–	–	–	0.98	–	–	0.54	–	–
Transformer, 12L, lr 10^{-4}	3	1.00	1.00	0.99	0.95	0.80	0.64	0.38	0.24	–
16 layers										
Transformer+ALiBi, 16L, lr 10^{-4}	3	1.00	0.99	0.94	0.85	0.66	0.53	0.38	0.28	0.20
Transformer+RoPE, 16L, lr 10^{-4}	3	1.00	0.93	0.71	0.55	0.35	0.24	0.17	0.13	0.09
Transformer, 16L, anchor lr	3	1.00	1.00	0.99	0.95	0.83	0.72	0.54	0.41	–
Transformer, 16L, lr 10^{-4}	3	1.00	1.00	0.98	0.92	0.75	0.57	0.31	0.18	–

Table 30: Dense prefix sweep, closing accuracy on the Prefix split at nine scale factors of the filler before the opening run, 2,000 items per scale, mean over gate-passing seeds, same cells as Table 29. Prefix at $\times 12$ is the split of Section 4.

cell	runs	$\times 1$	$\times 2$	$\times 3$	$\times 4$	$\times 6$	$\times 8$	$\times 12$	$\times 16$	$\times 24$
2 layers										
LSTM, 2L, lr 5×10^{-3}	5	–	–	–	–	–	–	–	–	–
GRU, 2L, lr 5×10^{-3}	5	–	–	–	–	–	–	–	–	–
RNN, 2L, lr 10^{-3}	5	–	–	–	–	–	–	–	–	–
Mamba, 2L, lr 10^{-3}	5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.93
Mamba-NoSelect, 2L, anchor lr	5	–	–	–	–	–	–	–	–	–
Transformer+ALiBi, 2L, lr 10^{-4}	5	0.98	0.84	0.66	0.55	0.41	0.34	0.30	0.29	0.26
Transformer+RoPE, 2L, anchor lr	3	1.00	0.95	0.89	0.86	0.82	0.83	0.85	0.79	0.53
Transformer+RoPE, 2L, lr 10^{-4}	3	1.00	0.94	0.86	0.80	0.74	0.70	0.63	0.49	0.29
Transformer, 2L, anchor lr	3	1.00	0.64	0.39	0.29	0.21	0.19	0.18	0.17	0.10
Transformer, 2L, lr 5×10^{-3}	5	1.00	0.56	0.39	0.32	0.26	0.27	0.37	0.34	0.23
S4D, 2L, lr 10^{-2}	5	–	–	–	–	–	–	–	–	–
4 layers										
Mamba, 4L, lr 10^{-2}	3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Hybrid A A A A, 4L, anchor lr	5	1.00	0.91	0.67	0.51	0.30	0.18	0.08	0.04	0.03
Hybrid A A A A, 4L, lr 10^{-4}	5	–	–	–	–	–	–	–	–	–
Hybrid A M M M, 4L, anchor lr	3	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.88	0.75
Hybrid M A M A, 4L, anchor lr	5	1.00	1.00	1.00	1.00	1.00	0.97	0.86	0.76	0.65
Hybrid M M A M, 4L, anchor lr	5	1.00	1.00	1.00	1.00	0.99	0.97	0.90	0.82	0.71
Hybrid M M M A, 4L, anchor lr	3	1.00	1.00	1.00	1.00	1.00	0.99	0.96	0.93	0.88
Hybrid M M M M, 4L, anchor lr	5	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.95	0.90
Hybrid M M M M, 4L, lr 10^{-3}	3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96
Transformer, 4L, anchor lr	3	1.00	0.57	0.33	0.21	0.14	0.17	0.19	0.17	0.11
8 layers										
Hybrid A A A A A A A A, 8L, lr 10^{-2}	5	–	–	–	–	–	–	–	–	–
Hybrid M A M A M A M A, 8L, lr 10^{-3}	5	–	–	–	–	–	–	–	–	–
Transformer+ALiBi, 8L, lr 10^{-4}	3	1.00	1.00	0.88	0.71	0.53	0.44	0.33	0.29	0.22
Transformer+RoPE, 8L, lr 10^{-4}	3	1.00	0.98	0.77	0.58	0.34	0.23	0.11	0.10	0.09
Transformer, 8L, anchor lr	3	1.00	0.69	0.44	0.31	0.15	0.11	0.09	0.08	0.10
Transformer, 8L, lr 10^{-4}	3	1.00	0.71	0.44	0.28	0.11	0.07	0.08	0.09	0.07
12 layers										
Transformer+ALiBi, 12L, lr 10^{-4}	3	1.00	1.00	0.97	0.82	0.61	0.45	0.26	0.18	0.10
Transformer+RoPE, 12L, lr 10^{-4}	3	1.00	0.98	0.80	0.61	0.42	0.34	0.25	0.25	0.21
Transformer, 12L, anchor lr	5	1.00	0.78	0.56	0.40	0.23	0.18	0.13	0.11	0.09
Transformer, 12L, lr 10^{-3}	5	–	–	–	–	–	–	–	–	–
Transformer, 12L, lr 10^{-4}	3	1.00	0.74	0.45	0.31	0.14	0.08	0.07	0.06	0.06
16 layers										
Transformer+ALiBi, 16L, lr 10^{-4}	3	1.00	1.00	0.90	0.69	0.38	0.21	0.06	0.03	0.04
Transformer+RoPE, 16L, lr 10^{-4}	3	1.00	0.99	0.82	0.67	0.48	0.40	0.30	0.28	0.24
Transformer, 16L, anchor lr	3	1.00	0.86	0.65	0.47	0.27	0.16	0.08	0.06	0.05
Transformer, 16L, lr 10^{-4}	3	1.00	0.80	0.51	0.37	0.17	0.09	0.06	0.05	0.06

Table 31: Inner-run sweep in closing accuracy, 2,000 items per condition, mean over gate-passing seeds, for the cells of Table 29 that were scored. The Inner split at the training scale with the number of inner runs fixed at k instead of drawn from 0 to 2, so k inner runs and $k+1$ gaps of 2 to 8 tokens lie between the opening run and the probe.

cell	runs	$k=1$	$k=2$	$k=3$	$k=4$	$k=6$	$k=8$
1 layer							
LSTM, 1L, lr 5×10^{-3}	3	1.00	1.00	1.00	1.00	0.99	0.95
Mamba, 1L, lr 10^{-2}	3	0.94	0.95	0.54	0.38	0.23	0.22
Mamba, 1L, lr 10^{-3}	5	0.93	0.94	0.69	0.50	0.32	0.27
2 layers							
LSTM, 2L, anchor lr	3	1.00	1.00	1.00	1.00	0.99	0.91
LSTM, 2L, lr 10^{-2}	3	1.00	1.00	1.00	1.00	1.00	0.98
LSTM, 2L, width 512, lr 5×10^{-3}	3	1.00	1.00	1.00	1.00	1.00	1.00
Mamba, 2L, anchor lr	3	1.00	1.00	1.00	0.99	0.89	0.73
Mamba, 2L, lr 10^{-2}	5	1.00	1.00	1.00	1.00	1.00	0.99
Mamba, 2L, lr 10^{-2}, constant	3	1.00	0.99	1.00	0.98	0.98	0.96
Mamba, 2L, lr 10^{-3}	5	1.00	1.00	1.00	1.00	0.99	0.92
Transformer+ALiBi, 2L, anchor lr	3	0.99	0.98	0.84	0.61	0.48	0.52
Transformer+ALiBi, 2L, lr 10^{-3}	3	0.99	0.99	0.86	0.65	0.50	0.50
Transformer+ALiBi, 2L, lr 10^{-3}, constant	3	0.97	0.96	0.86	0.69	0.51	0.46
Transformer+ALiBi, 2L, lr 10^{-3}, step drop	3	0.98	0.98	0.84	0.63	0.44	0.46
Transformer+ALiBi, 2L, lr 10^{-4}	5	0.99	0.99	0.88	0.68	0.58	0.61
Transformer+RoPE, 2L, anchor lr	3	1.00	1.00	0.97	0.79	0.49	0.42
Transformer+RoPE, 2L, lr 10^{-3}	3	1.00	1.00	0.92	0.58	0.27	0.26
Transformer+RoPE, 2L, lr 10^{-4}	3	1.00	1.00	0.94	0.78	0.54	0.39
Transformer+T5, 2L, anchor lr	3	0.99	0.98	0.84	0.55	0.34	0.37
Transformer+sinusoidal, 2L, anchor lr	5	0.99	0.99	0.63	0.25	0.28	0.37
Transformer+sinusoidal, 2L, base 100, anchor lr	3	0.99	0.97	0.61	0.22	0.23	0.32
Transformer+sinusoidal, 2L, base 1000, anchor lr	3	0.97	0.96	0.54	0.19	0.28	0.38
Transformer+sinusoidal, 2L, base 100000, anchor lr	3	1.00	0.99	0.68	0.28	0.30	0.36
Transformer, 2L, anchor lr	3	1.00	1.00	0.66	0.27	0.22	0.27
Transformer, 2L, lr 10^{-2}	3	1.00	0.99	0.87	0.57	0.34	0.27
Transformer, 2L, lr 10^{-4}	5	0.98	0.97	0.69	0.32	0.20	0.27
Transformer, 2L, lr 5×10^{-3}	5	1.00	1.00	0.96	0.82	0.57	0.39
4 layers							
LSTM, 4L, anchor lr	3	1.00	1.00	1.00	1.00	0.96	0.79
LSTM, 4L, lr 10^{-2}	3	1.00	1.00	1.00	1.00	1.00	0.97
LSTM, 4L, lr 5×10^{-3}	5	1.00	1.00	1.00	0.99	0.86	0.73
Mamba, 4L, anchor lr	3	1.00	1.00	1.00	1.00	0.94	0.85
Mamba, 4L, lr 10^{-2}	3	1.00	1.00	1.00	1.00	1.00	1.00
Mamba, 4L, lr 10^{-3}	3	1.00	1.00	1.00	1.00	0.97	0.86
Mamba, 4L, width 256, anchor lr	3	1.00	1.00	1.00	1.00	1.00	0.97
Mamba, 4L, width 512, anchor lr	3	1.00	1.00	1.00	1.00	1.00	0.98
Hybrid A A A A, 4L, anchor lr	5	1.00	1.00	0.85	0.54	0.28	0.39
Hybrid A M M M, 4L, anchor lr	3	1.00	1.00	1.00	1.00	0.94	0.78
Hybrid M A M A, 4L, anchor lr	5	1.00	1.00	1.00	1.00	1.00	1.00
Hybrid M A M A, 4L, lr 10^{-3}	3	1.00	1.00	1.00	1.00	0.99	0.96
Hybrid M M A M, 4L, anchor lr	5	1.00	1.00	1.00	1.00	1.00	1.00
Hybrid M M M A, 4L, anchor lr	3	1.00	1.00	1.00	1.00	0.99	0.97
Hybrid M M M M, 4L, anchor lr	5	1.00	1.00	1.00	1.00	0.97	0.78
Hybrid M M M M, 4L, lr 10^{-3}	3	1.00	1.00	1.00	1.00	0.96	0.90
Transformer, 4L, anchor lr	3	1.00	1.00	0.98	0.92	0.43	0.15
Transformer, 4L, lr 10^{-4}	3	1.00	1.00	0.99	0.89	0.34	0.14
8 layers							
LSTM, 8L, anchor lr	3	1.00	1.00	1.00	1.00	0.94	0.78
Mamba, 8L, anchor lr	3	1.00	1.00	1.00	1.00	0.99	0.93
Mamba, 8L, lr 10^{-2}	1	1.00	1.00	1.00	1.00	1.00	1.00
Transformer+ALiBi, 8L, lr 10^{-4}	3	1.00	1.00	1.00	0.99	0.83	0.59
Transformer+RoPE, 8L, lr 10^{-4}	3	1.00	1.00	1.00	0.93	0.54	0.38
Transformer+T5, 8L, lr 10^{-4}	3	1.00	1.00	0.87	0.54	0.16	0.29
Transformer+sinusoidal, 8L, lr 10^{-4}	3	1.00	1.00	0.99	0.95	0.71	0.43
Transformer, 8L, anchor lr	3	1.00	1.00	0.98	0.87	0.59	0.35
Transformer, 8L, lr 10^{-4}	3	1.00	1.00	0.99	0.95	0.60	0.30
12 layers							
Transformer+ALiBi, 12L, lr 10^{-4}	3	1.00	1.00	1.00	0.99	0.80	0.57
Transformer+RoPE, 12L, lr 10^{-4}	3	1.00	1.00	0.99	0.88	0.43	0.25
Transformer, 12L, anchor lr	5	1.00	1.00	0.98	0.92	0.73	0.48
Transformer, 12L, lr 10^{-4}	3	1.00	1.00	1.00	0.97	0.79	0.56
16 layers							
Transformer+ALiBi, 16L, lr 10^{-4}	3	1.00	1.00	1.00	0.99	0.65	0.40
Transformer+RoPE, 16L, lr 10^{-4}	3	1.00	1.00	0.98	0.82	0.40	0.33
Transformer, 16L, anchor lr	3	1.00	1.00	1.00	1.00	0.92	0.72
Transformer, 16L, lr 10^{-4}	3	1.00	1.00	0.99	0.95	0.70	0.45

Table 32: Free-running counterpart of Table 28, the same patches with the whole closing tail generated by greedy decoding. Clean and corrupt are closing accuracy at training gaps and at gaps $\times 12$ on 300 drawn open items with an inner run (count in brackets), recovery is as in Table 28, and the last column is the share of corrupt items that never stop. Mean over seeds.

cell	seeds (items)	clean $\rightarrow$ corrupt	attn. output	layer-0 resid.	no-stop
RoPE 2L, lr 3×10^{-4}	3 (170)	$1.00 \rightarrow 0.11$	$+0.92$	$+0.40$	0.73
RoPE 2L, lr 10^{-4}	3 (170)	$1.00 \rightarrow 0.11$	$+0.75$	$+0.25$	0.52
ALiBi 2L, lr 3×10^{-4}	3 (170)	$0.99 \rightarrow 0.28$	$+0.69$	$+0.25$	0.00
ALiBi 2L, lr 10^{-4}	5 (171)	$0.98 \rightarrow 0.60$	-0.24	-0.21	0.01
learned 2L, lr 3×10^{-4}	3 (170)	$1.00 \rightarrow 0.15$	$+0.01$	$+0.64$	0.69
learned 2L, lr 10^{-4}	5 (171)	$0.98 \rightarrow 0.09$	$+0.13$	$+0.49$	0.13
sinusoidal 2L	5 (171)	$0.99 \rightarrow 0.12$	-0.02	$+0.07$	0.27
T5 2L	3 (170)	$0.99 \rightarrow 0.18$	-0.16	$+0.00$	0.47
learned 8L	3 (170)	$1.00 \rightarrow 0.21$	$+0.03$	-0.15	0.27
ALiBi 8L	3 (170)	$1.00 \rightarrow 0.58$	-0.09	-0.22	0.00
RoPE 8L	3 (170)	$1.00 \rightarrow 0.21$	$+0.01$	-0.14	0.16

The panels of Figure 10 run over the key positions on a log scale, and the title of each panel gives the layer-1 mass on the opening run for the drawn item and its mean over 64 items, with the mean mass on the inner runs and on the prefix.



Figure 10: Attention from the probe position, final checkpoints of seed 0, one open CLOSE item with two inner runs at training gaps (left) and with its gaps stretched twelvefold (right). Each panel is the head-averaged attention row of the probe query, layer 0 dashed and layer 1 solid. The red band is the opening run, the orange lines the inner fence tokens and the grey band the prefix.

Table 33: Prefix-randomized training, all nine arms, closing accuracy on Prefix, Pad and Gaps before (standard training, prefixes of 1 to 4 tokens) and after training with prefixes of 1 to 48 tokens, mean over gate-passing seeds (seeds before / after in the second column). The splits are generated with the standard prefix range in both cases.

		Prefix		Pad		Gaps	
arm	seeds	before	after	before	after	before	after
Transformer 2L, anchor lr	3 / 3	0.18	0.96	0.00	0.00	0.14	0.33
Transformer 2L, lr 10^{-4}	5 / 3	0.07	0.94	0.00	0.00	0.11	0.23
Transformer+sinusoidal 2L	5 / 3	0.30	1.00	0.00	0.00	0.12	0.29
Transformer+ALiBi 2L, lr 10^{-4}	5 / 3	0.30	0.99	0.19	0.49	0.62	0.25
Transformer+RoPE 2L, lr 10^{-4}	3 / 3	0.62	1.00	0.00	0.00	0.11	0.31
Transformer+T5 2L	3 / 3	0.22	0.98	0.00	0.00	0.17	0.08
LSTM 2L, anchor lr	3 / 3	0.61	1.00	0.16	1.00	0.72	0.47
Mamba 2L, anchor lr	3 / 3	0.76	1.00	0.26	1.00	0.82	0.76
Hybrid M A M A 4L, anchor lr	5 / 3	0.86	1.00	0.54	1.00	0.97	0.69

Table 34: Layer-ablation map. Each layer is skipped in turn (identity for an attention block, no residual update for a Mamba block) and the six segment splits are rescored on 1,000 items in closing accuracy, mean over seeds. Layers are indexed from the input, and the last two columns give the intact Gaps and Pad closing accuracy.

cell	runs	critical layers	redundant	Pad carriers	gaps	pad
Mamba 2L, anchor lr	3	0, 1	0/2	none	0.83	0.28
Mamba 4L, anchor lr	3	0, 1	0/4	0, 1	0.50	0.64
Mamba 8L, anchor lr	3	0, 1, 2	4/8	0	0.49	0.50
Mamba 4L, lr 10^{-2}	3	0, 1, 3	0/4	0, 1, 3	1.00	1.00
Mamba 8L, lr 10^{-2}	1	0, 1, 7	5/8	0, 1, 7	1.00	1.00
Transformer 2L, lr 10^{-4}	5	0, 1	0/2	none	0.10	0.00
Transformer 4L, anchor lr	3	0, 1, 2, 3	0/4	none	0.16	0.00
Transformer 8L, anchor lr	3	0, 1, 2, 3, 4	1/8	none	0.26	0.00
Transformer 8L, lr 10^{-4}	3	0, 1, 2, 3, 4	1/8	none	0.21	0.00
Transformer 12L, anchor lr	5	0, 1, 2, 3, 4, 5	2/12	none	0.45	0.04
Transformer 16L, anchor lr	3	0, 1, 3, 4, 5, 6, 7	8/16	none	0.54	0.00
Transformer 16L, lr 10^{-4}	3	0, 1, 3, 6	5/16	none	0.31	0.04
Transformer+ALiBi 2L, lr 10^{-4}	5	0, 1	0/2	0	0.62	0.20
Transformer+ALiBi 8L	3	0, 1, 2, 3	4/8	none	0.61	0.08
Transformer+ALiBi 12L	3	0, 1, 2, 3	6/12	none	0.45	0.05
Transformer+ALiBi 16L	3	0, 1, 3, 4	8/16	none	0.38	0.08
Transformer+RoPE 2L, lr 10^{-4}	3	0, 1	0/2	none	0.10	0.00
Transformer+RoPE 8L	3	0, 1, 3, 4	2/8	none	0.17	0.00
Transformer+RoPE 12L	3	0	3/12	none	0.16	0.06
Transformer+RoPE 16L	3	0, 1	11/16	none	0.17	0.00
Hybrid M M M M	5	0, 1	2/4	0, 1	0.42	0.75
Hybrid M M A M	5	0, 1, 2	0/4	0, 1, 2	0.87	0.59
Hybrid M A M A	5	0, 1, 2, 3	0/4	0, 1	0.97	0.54
Hybrid A A A A	5	0, 1, 2, 3	0/4	none	0.06	0.00

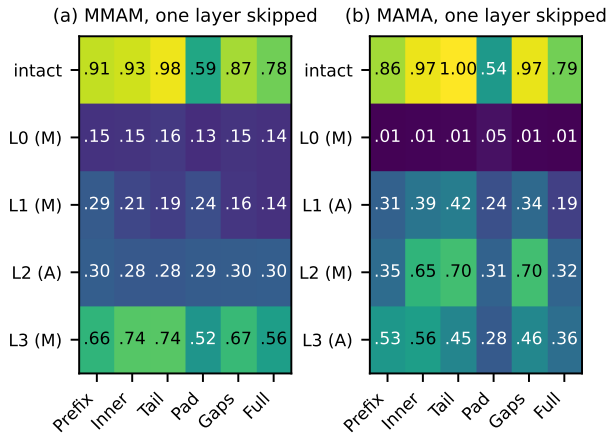


Figure 11: Layer ablation of the four-layer MMAM (a) and MAMA (b) hybrids, closing accuracy per split with one layer skipped, five seeds.

Mechanism (detail for Sections 4.2 to 4.4). Under prefix insertion the ALiBi distance penalty favours the opening run while its content term leaks to the prefix. Prefix-randomized training restores Prefix in closing accuracy from 0.18 to 0.96 for learned positions, from 0.30 to 1.00 for sinusoidal, from 0.30 to 0.99 for ALiBi, from 0.62 to 1.00 for RoPE, from 0.22 to 0.98 for T5, and from 0.61, 0.76 and 0.86 to 1.00 for the LSTM, Mamba and the alternating hybrid (Table 33). Its cost on Gaps closing accuracy is a drop from 0.82 to 0.76 for Mamba, from 0.72 to 0.47 for the LSTM, from 0.97 to 0.69 for the alternating hybrid and from 0.62 to 0.25 for ALiBi, while the encodings that were already near zero move by a few points. A training arm with random position offsets failed the gate and is not used.

The ALiBi cell at lr 10^{-4} used elsewhere in the paper does not collapse at twelvefold gaps and recovers 0.03 under readout patching. Learned, RoPE and sinusoidal models have no decisive head. In most conditions the first closing token is right and the failure is in how many tokens to emit and when to stop. The free-running rescoring of the patches uses 300 drawn items and counts an item as exact only if every fence token and the stop are right, and under stretched gaps the corrupt RoPE and anchor-rate learned models never stop in 73% and 69% of items while the two ALiBi cells stop in every item and in all but 2%, so the patching verdicts hold for the whole emitted tail (Table 32).

ALiBi at lr 10^{-4} does not cross one half on Gaps through $\times 24$ at two layers, where it holds 0.57, and crosses between 16 and 24 at eight layers and between 8 and 12 at twelve and sixteen (Table 29). A position shift of 12 collapses the twelve- and sixteen-layer models to 0.04 and 0.08 at lr 10^{-4}. In the single-attention hybrid the two Mamba layers below the attention layer are the critical ones for distance, since skipping either drops Gaps from 0.87 to 0.15 or 0.16, against 0.30 without the attention layer.

Five of sixteen learned-position layers at lr 10^{-4} and eight at the anchor lr, eight of sixteen ALiBi layers and eleven of sixteen RoPE layers can be skipped one at a time with no split losing more than a fifth of its closing accuracy (Table 34). In that table a layer is critical when skipping it halves at least one split, redundant when every split keeps at least 80% of its intact accuracy, and a Pad carrier when skipping it halves Pad, with splits below 0.1 intact ignored.

On the recurrent side a convolution of width 2 gives 0.85 and of width 8 gives 0.98 on Distance for the anchor Mamba (Table 11), the twelve- and sixteen-layer Mamba grid has no middle learning rate, and we did not map the eight-layer circuit, ablate the query, key and value composition paths, or measure relay mass between heads.

H ATTENTION-LAYER LADDER

Table 36 lists the four- and eight-layer cells of the ladder in CLOSE, and Figure 12 the three tasks and the safe-length failure rate at content-gap scales 12 to 48. At lr 10^{-3} the pure stack gains on CLOSE, 0.42 to 0.72, while the alternating stack loses, so the best cell of each stack sits at a different rate. At lr 5×10^{-3} and 10^{-2} the pure stack gives 0.59 and 1.00 on Distance, the second on the two seeds of three that pass the gate, and the alternating stack 0.70 and 0.74, so on CLOSE the best rung of the pure stack matches the best cell of the alternating stack, and Count stays at zero at every rate.

Table 35 gives the seed-level crossed interaction of Section 4 for the ladder and for the two-layer cells, with 10,000 bootstrap resamples in which each group is resampled with replacement, and a positive Δ whose interval excludes zero is the crossed interaction at the seed level. In the dense gap sweep of Table 29 the pure four-layer stack crosses one half between $\times 8$ and $\times 12$ at the anchor lr and between $\times 16$ and $\times 24$ at lr 10^{-3}, the stack with its one attention layer at the bottom crosses between $\times 16$ and $\times 24$, the three other mixed stacks stay above one half through $\times 24$, at 0.55 to 0.93, and the attention-only stack crosses between $\times 4$ and $\times 6$.

Table 35: Seed-level crossed interaction. For every gate-passing seed s of a recurrent-component cell (R) and of an attention-only cell (A), D_s is its Distance closing accuracy minus its Count closing accuracy. Δ is the mean of D_s over the R seeds minus the mean over the A seeds, with a 95% bootstrap interval over seeds, and the last column is the share of (R, A) seed pairs with $D_R > D_A$.

pair	seeds R / A	D_s, R	D_s, A	Δ [95% CI]	pairs $D_R > D_A$
four-layer ladder, M M M M against A A A A	5 / 5	+0.34, +0.45, +0.52, +0.45, +0.32	-0.70, -0.75, -0.87, -0.94, -0.92	+1.25 [+1.15, +1.36]	1.00
four-layer ladder, M A M A against A A A A	5 / 5	+0.93, +0.96, +0.94, +1.00, +1.00	-0.70, -0.75, -0.87, -0.94, -0.92	+1.80 [+1.71, +1.88]	1.00
two-layer anchors, Mamba against Transformer, anchor lr	3 / 3	+0.65, +0.96, +0.82	+0.08, +0.13, +0.08	+0.71 [+0.55, +0.87]	1.00
lr-matched cells, Mamba 10^{-3} against Transformer 5×10^{-3}	5 / 5	+0.83, +0.97, +0.99, +0.98, +0.98	+0.19, +0.03, +0.22, -0.04, +0.13	+0.84 [+0.75, +0.94]	1.00
lr-matched cells, LSTM 5×10^{-3} against Transformer 5×10^{-3}	5 / 5	+0.83, +0.88, +0.78, +0.64, +0.65	+0.19, +0.03, +0.22, -0.04, +0.13	+0.65 [+0.52, +0.77]	1.00
eight-layer stacks, M×8 against NoPE A×8	3 / 5	+0.61, +0.43, +0.83	-0.75, -0.65, -0.95, -0.98, -0.64	+1.42 [+1.21, +1.65]	1.00

Table 36: The attention-layer ladder, CLOSE closing accuracy on Distance and Count, mean over gate-passing runs (runs column, gate-passing in brackets). Four-layer stacks with 0 to 4 NoPE attention layers at the anchor lr, the pure and the alternating stack at three higher rates, and eight-layer stacks at the anchor lr. M is a Mamba layer and A an attention layer, from the bottom.

cell	runs	Distance	Count
Four layers, anchor lr, five seeds for CLOSE			
M M M M	5 (5)	0.42	0.00
M M A M	5 (5)	0.87	0.00
M A M A	5 (5)	0.96	0.00
A A A A	5 (5)	0.06	0.90
Four layers, lr 10^{-3}			
M M M M	3 (3)	0.72	0.00
M A M A	3 (3)	0.89	0.00
Four layers, lr 5×10^{-3}, CLOSE *only*			
M M M M	3 (3)	0.59	0.00
M A M A	3 (3)	0.70	0.00
Four layers, lr 10^{-2}, CLOSE *only*			
M M M M	3 (2)	1.00	0.00
M A M A	3 (3)	0.74	0.00
Eight layers, anchor lr			
M×8	3 (3)	0.63	0.00
1 attention of 8	3 (3)	0.84	0.00
2 attention of 8 (M M M A M M M A)	3 (3)	0.86	0.00
3 attention of 8 (M A M M A M M A)	3 (3)	0.93	0.01
M A M A M A M A	5 (5)	0.95	0.10
A×8	5 (5)	0.20	0.99

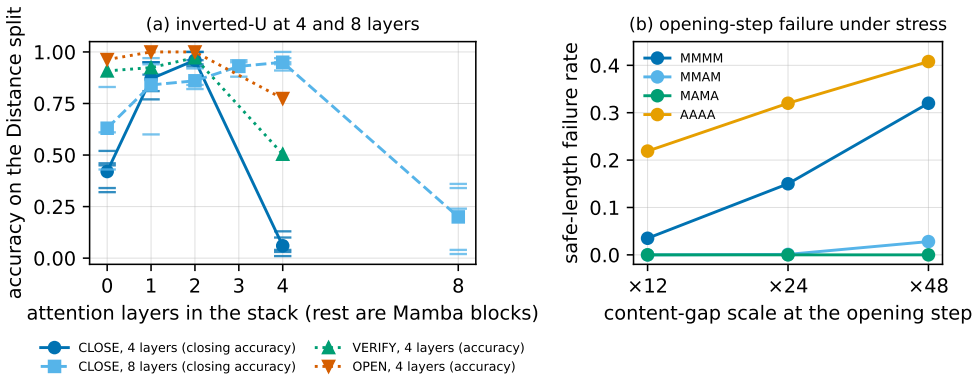


Figure 12: The attention-layer ladder. (a) Four- and eight-layer hybrids with 0 to 4 and 0 to 8 NoPE attention layers, anchor lr, three or five seeds per cell. CLOSE is closing accuracy on Distance, VERIFY and OPEN are accuracy. (b) safe-length failure of the four-layer cells at OPEN gap scales 12 to 48.

Ladder controls (detail for Section 4.3). Three controls hold the reading of the ladder. Retraining the pure and the alternating stack at lr 10^{-3} keeps the order on CLOSE, with the pure stack at 0.72. Moving the single attention layer to the bottom or the top lowers closing accuracy on Distance from 0.87 to 0.66 and 0.65, more than the 0.09 that a second layer adds, and bottom placement loses the Pad tolerance, because a NoPE attention layer then faces the padding without a recurrent layer beneath it.

Two limits apply beyond the learning rate. The hybrid gain is tied to the training distribution and budget, since prefix-randomized training lowers the alternating cell's closing accuracy on Gaps from 0.97 to 0.69, and the Pad tolerance of hybrids varies across seeds, so we do not claim it. Ten times the training budget lowers the four cells from 0.42, 0.87, 0.96 and 0.06 to 0.33, 0.61, 0.84 and 0.03 in the same order, and it lifts the alternating cell's count extrapolation to 0.23 and 0.27 in two of its three seeds, all of it at the first length beyond the training maximum.

The failure signature follows the dominant component. Pure Mamba closes short, at times at the length of another run, pure attention fails by not stopping, and the hybrids leave a few long closes. Under a 48-fold content scale pure Mamba opens exactly one short and pure attention opens far short.

Stacks with a single recurrent layer. The ladder stops at two attention layers of four, so a further control trains the four placements of a single Mamba layer among three NoPE attention layers at the anchor lr, three seeds each (Table 37). This is the control that asks whether a single recurrent layer removes count extrapolation, and in the table M is a Mamba layer and A a NoPE attention layer, written from the bottom of the stack to the top. All twelve runs pass the gate.

Count extrapolation does not vanish with one recurrent layer. It falls from 0.90 in the attention-only stack to 0.45 with the Mamba layer on top, to 0.25 and 0.22 with it in the second and the first position and to 0.11 in the third. With two Mamba layers it is 0.00 in four of the six arrangements and 0.13 and 0.17 in the other two (AMAM and MAAM, with one seed of three at 0.32 and 0.27), and with three or four it is 0.00 in every arrangement. One run, the top placement at 0.59, is the only run of the fixed-budget campaign with a recurrent block that passes one half on Count.

Distance depends on where the recurrent layer sits. The stack carries the length with the Mamba layer in the lower half, at 0.88 and 0.71, and loses it with the layer in the upper half, at 0.16 and 0.02. The two lower placements are the only four-layer cells that hold both axes above 0.2. Each recurrent layer therefore lowers count extrapolation, the first one most, a second one removes it in most arrangements, and a third removes it in all of them at this budget.

Table 37: Four-layer stacks of the hybrid class at the anchor lr by the number of Mamba layers, closing accuracy on the Distance and Count splits, mean ± standard deviation over the gate-passing seeds, and the Count value of every seed.

stack	Mamba layers	seeds	Distance	Count	Count by seed
MMMM	4	5	0.42 ± 0.09	0.00 ± 0.01	0.00, 0.00, 0.00, 0.01, 0.00
AMMM	3	3	0.66 ± 0.25	0.00 ± 0.00	0.00, 0.00, 0.00
MAMM	3	3	0.87 ± 0.14	0.00 ± 0.00	0.00, 0.00, 0.00
MMAM	3	5	0.87 ± 0.08	0.00 ± 0.00	0.00, 0.00, 0.00, 0.00, 0.00
MMMA	3	3	0.65 ± 0.20	0.00 ± 0.00	0.00, 0.00, 0.00
AAMM	2	3	0.79 ± 0.22	0.00 ± 0.00	0.00, 0.00, 0.00
AMAM	2	3	0.27 ± 0.13	0.13 ± 0.17	0.32, 0.04, 0.02
AMMA	2	3	0.71 ± 0.32	0.00 ± 0.00	0.00, 0.00, 0.00
MAAM	2	3	0.97 ± 0.02	0.17 ± 0.15	0.27, 0.00, 0.25
MAMA	2	5	0.96 ± 0.03	0.00 ± 0.00	0.00, 0.00, 0.00, 0.00, 0.00
MMAA	2	3	0.89 ± 0.05	0.00 ± 0.00	0.00, 0.00, 0.00
AAAM	1	3	0.02 ± 0.01	0.45 ± 0.14	0.46, 0.59, 0.32
AAMA	1	3	0.16 ± 0.14	0.11 ± 0.17	0.31, 0.00, 0.03
AMAA	1	3	0.71 ± 0.18	0.25 ± 0.05	0.30, 0.20, 0.26
MAAA	1	3	0.88 ± 0.10	0.22 ± 0.10	0.26, 0.11, 0.30
AAAA	0	5	0.06 ± 0.05	0.90 ± 0.11	0.70, 0.89, 0.98, 0.96, 0.95

I OPEN-K

The OPEN-K pilot asks whether a model can commit to a safe opening length from knowledge held in its weights rather than from content it has seen. The identifiers, the auxiliary loss, the held-out split and the two arms are defined in Appendix A. Table 38 lists the best joint setting of each anchor family and its isolated control, chosen by held-out commit among the 46 cells of the pilot. With the auxiliary channel trained jointly, all three anchor families commit on held-out identifiers at their own learning rate (LSTM 0.94 at lr 5×10^{-3}, Mamba and the Transformer at or near 1.00), while the isolated arms stay near the chance level of 0.17, so the transfer runs through the shared channel and not through memorization.

Table 38: The OPEN-K pilot, the best joint setting of each anchor family by held-out commit and its isolated control. Held-out commit is the accuracy of the opening choice $m + 1$ on identifiers never seen as OPEN-K pairs in training, held-out probe the accuracy of the direct length query. Seed means, seed standard deviation in units of the last digit, anchor lr unless stated.

model	setting	regime	seeds	held-out commit	held-out probe
LSTM	$\lambda = 0.5$, lr 0.005	joint	3	0.94(5)	0.93(6)
	same	isolated	3	0.17(10)	0.02(3)
Mamba	$\lambda = 0.5$	joint	3	1.00(0)	0.99(1)
	same	isolated	3	0.21(7)	0.00(0)
Transformer	$\lambda = 0.5$	joint	3	1.00	1.00
	same	isolated	3	0.04(4)	0.00(0)

J BRIDGE DETAIL

Table 39: Family-level correspondence between the LatentMD error types and axes and the MBT quantities.

LatentMD	MBT counterpart
E0 (Correct)	exact minimum close. A long close is an over-safe E0.
collision family {E1, E3}, an inner run reaches the outer length, visible (E1) or masked by balance (E3)	VERIFY premature closure, OPEN safe-length failure, CLOSE inner-length match excess
unclosed family {E2}, the outer fence is never validly closed	CLOSE short close and no-stop, VERIFY unclosed
E1 (Premature Closure) by the mechanical rule alone, a run written after the fence has closed reopens it and never closes	CLOSE answer-0 error, the lost record of a closed fence (Tail at the anchor lr), placed in neither family
common factor of both families, no margin discipline (F3)	carrying the margin, detecting a collision, choosing a margin
D-Outer-Length	ℓ, the Count split
B-axis (inner nesting)	inner-run count k
D-Family and D-Cross-Family	fixed fence versus bound fence
content length	the Distance split and the segment splits
nine Transformer models	the architecture axis
A-axis (outer constraint), hints H1 to H3, the Literal role	no counterpart
content correctness C1 to C3	removed by design

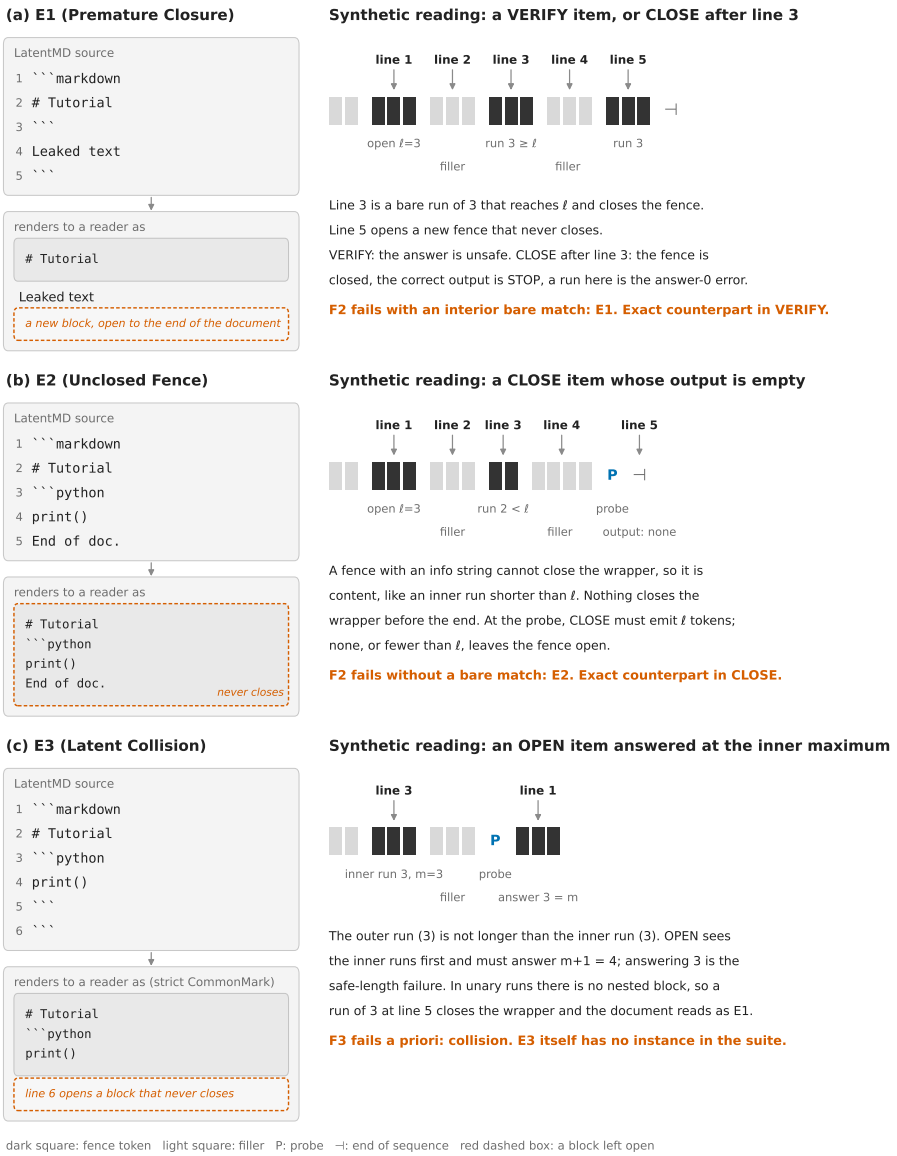


Figure 13: The three minimal failure examples of LatentMD (Lee & Kang, 2026), redrawn from its Figure 2, with the document a reader sees below each source, next to the synthetic sequence that instantiates the same situation. Markdown lines are numbered above the segment of the token stream they correspond to, and a red dashed box marks a block left open.

In Figure 13 the E1 (Premature Closure) example (a) is a VERIFY item whose inner run reaches ℓ, and after that line a CLOSE item whose correct output is STOP. The E2 (Unclosed Fence) example (b) is a CLOSE item whose output is empty or shorter than ℓ. The E3 (Latent Collision) example (c) is an OPEN item answered at the inner maximum. Its latent form needs a nested block that an info string opens, which unary runs cannot express, so in the suite the same document reads as E1 (Premature Closure).

Table 40: The deployed models of Section 5, their attention layers and the positional encoding of those layers, as documented by their authors. Labels follow LatentMD.

model	class	attention layers	positional encoding	source
Falcon-Mamba-7B	pure SSM	none	none	Zuo et al., 2024
Jamba-Mini-1.7	hybrid	1 in 8 layers	none	Lieber et al., 2024
Falcon-H1-7B, 34B	parallel hybrid	attention and SSM heads in every block, channels 2:1:5 for SSM, attention and MLP	RoPE, base 10^{11}	Zuo et al., 2025
Granite-4.0-H-Tiny, Small	hybrid	1 in 10 blocks	none	IBM Granite Team, 2025
Nemotron-Nano-9B-v2	hybrid	6 of 62 layers	none	NVIDIA, 2025
Nemotron-H-47B	hybrid	about 8% of layers	none	Blakeman et al., 2025
Llama-3.1-8B, 70B	Transformer	all	RoPE, base 5×10^5	Grattafiori et al., 2024
Gemma-2-9B	Transformer	all	RoPE	Gemma Team, 2024
Gemma-3-27B	Transformer	all, local and global	RoPE, base 10^4 local and 10^6 global	Gemma Team, 2025
Qwen3-7B, 32B	Transformer	all	RoPE	Yang et al., 2025
GPT-4o, Gemini-2.5-Flash, Claude-Sonnet-4	Transformer	undisclosed	undisclosed	Lee & Kang, 2026

The eight models are the instruction-tuned open-weight models with a recurrent component that remained after the exclusions of Section 5, which were fixed before any generation. Decoding used temperature 0, top-p 0.9, 7,168 output tokens and a fixed seed on the 4,179 prompts of the LatentMD grid and the 248 natural prompts, and the released LatentMD scoring tool was run unchanged. Table 41 gives the E2 (Unclosed Fence) rate by content length under both truncation rules for all seventeen models, Table 42 the two main end states of the unclosed fences with their counts, Wilson intervals and the content-restricted denominator, and Table 43 the hint axis with wrap compliance. Content length is measured in lines from the outer opening fence to its close or the end of the output, trunc. is the number of truncated records and n the number of generations scored under the frozen rule.

The size tiers of Table 41 follow the LatentMD groups of the Transformers by total parameter count, small at 7 to 9B, mid at 27 to 34B and large at 47B and above, so the three mixture-of-experts hybrids sit by their totals, Granite-4.0-H-Tiny at 7B with 1B active, Granite-4.0-H-Small at 32B with 9B active and Jamba-Mini-1.7 at 52B with 12B active. The denominator of Table 41 and of Figure 14 is the set of valid generations that opened at least one fence, which is 79.7% of Falcon-Mamba's generations and 95% to 100% for every other model. The content-restricted denominator of Table 42 ensures that a never-closed fence cannot come from missing content alone, and in that table the nine LatentMD Transformers are above the rule and the eight held-out models below it.

The hint columns of Table 43 are boundary-failure rates under the wrapped constraint, with no hint (the mean of the four wrapped cells, the LatentMD baseline) and with hints H1 (rule), H2 (one-shot example) and H3 (length instruction). The four wrapped cells of Table 43 are A2 with the inner-block conditions B1 to B4, compliance there is the F1 verdict, every entry pools the prompts of the four conditions, and this pooling reproduces the LatentMD hint figure within 0.001 for its nine models.

The nine Transformers come from the released LatentMD generations, produced under the same decoding settings and re-aggregated with the same functions as the eight, since LatentMD reports no quantity by content length. Every LatentMD input of this comparison is public (Lee & Kang, 2026). The LatentMD dataset on the Hugging Face Hub, released under CC BY-SA 4.0, holds the prompt grid with its natural and hint prompts and the responses of all nine models, and the code repository linked from that paper holds the generation scripts and the scoring tool.

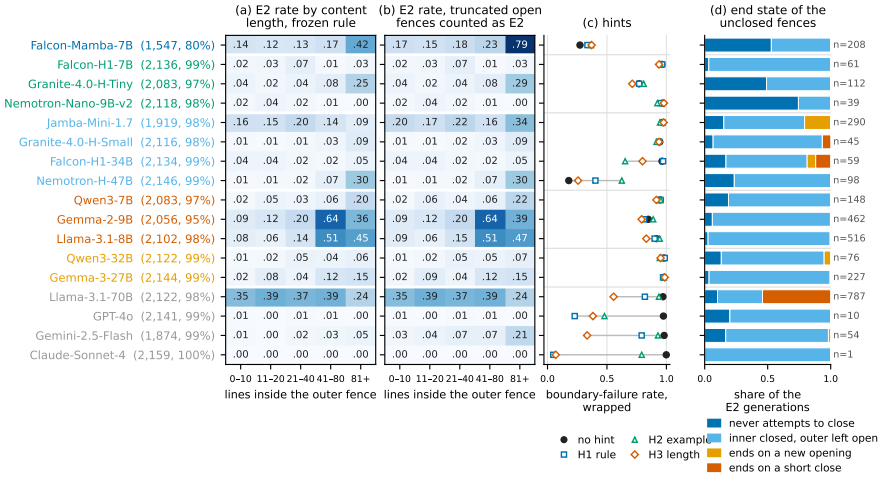


Figure 14: The seventeen models by class (Section 5.2). (a) E2 (Unclosed Fence) rate by content length under the frozen rule and (b) with truncated open fences counted. (c) Boundary-failure rate without a hint and under H1 to H3. (d) End state of the unclosed fences, n their number. Brackets give the denominator of (a) and (b), the valid generations that open a fence, and its share of all valid generations.

Table 41: E2 (Unclosed Fence) rate by content length for the eight held-out models and the nine LatentMD Transformers, the curves of Figure 14a and b. Under the frozen rule truncated generations are excluded, under the second rule a truncated generation whose outer fence is still open counts as E2 (Unclosed Fence).

		frozen rule, lines					truncations counted as E2, lines						
model	class	0–10	11–20	21–40	41–80	>80	0–10	11–20	21–40	41–80	>80	trunc.	n
Falcon-Mamba-7B	pure SSM, small	0.14	0.12	0.13	0.17	0.42	0.17	0.15	0.18	0.23	0.79	181	1547
Jamba-Mini-1.7	hybrid, large	0.16	0.15	0.20	0.14	0.09	0.20	0.17	0.22	0.16	0.34	197	1919
Falcon-H1-7B	hybrid, small	0.02	0.03	0.07	0.01	0.03	0.02	0.03	0.07	0.01	0.03	0	2136
Falcon-H1-34B	hybrid, mid	0.04	0.04	0.02	0.02	0.05	0.04	0.04	0.02	0.02	0.05	0	2134
Granite-4.0-H-Tiny	hybrid, small	0.04	0.02	0.04	0.08	0.25	0.04	0.02	0.04	0.08	0.29	5	2083
Granite-4.0-H-Small	hybrid, mid	0.01	0.01	0.01	0.03	0.09	0.01	0.01	0.02	0.03	0.09	1	2116
Nemotron-Nano-9B-v2	hybrid, small	0.02	0.04	0.02	0.01	0.00	0.02	0.04	0.02	0.01	0.00	0	2118
Nemotron-H-47B	hybrid, large	0.01	0.01	0.02	0.07	0.30	0.01	0.01	0.02	0.07	0.30	0	2146
Qwen3-7B	Transformer, small	0.02	0.05	0.03	0.06	0.20	0.02	0.06	0.04	0.06	0.22	19	2083
Gemma-2-9B	Transformer, small	0.09	0.12	0.20	0.64	0.36	0.09	0.12	0.20	0.64	0.39	3	2056
Llama-3.1-8B	Transformer, small	0.08	0.06	0.14	0.51	0.45	0.09	0.06	0.15	0.51	0.47	13	2102
Qwen3-32B	Transformer, mid	0.01	0.02	0.05	0.04	0.06	0.01	0.02	0.05	0.05	0.07	8	2122
Gemma-3-27B	Transformer, mid	0.02	0.08	0.04	0.12	0.15	0.02	0.09	0.04	0.12	0.15	2	2144
Llama-3.1-70B	Transformer, large	0.35	0.39	0.37	0.39	0.24	0.35	0.39	0.37	0.39	0.24	1	2122
GPT-4o	Transformer, large	0.01	0.00	0.01	0.01	0.00	0.01	0.00	0.01	0.01	0.00	0	2141
Gemini-2.5-Flash	Transformer, large	0.01	0.00	0.02	0.03	0.05	0.03	0.04	0.07	0.07	0.21	230	1874
Claude-Sonnet-4	Transformer, large	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	2159

Table 42: The two main end states of the unclosed fences. Never is the share of E2 (Unclosed Fence) generations with no fence line after the outer opening, with a 95% Wilson interval, and inner the share that close the inner blocks and then stop. The right half restricts the denominator to E2 (Unclosed Fence) generations whose content passed the LatentMD content check.

	every E2 (Unclosed Fence) generation			content-correct E2 (Unclosed Fence) generations		
model	n	never [95% CI]	inner	n	never [95% CI]	inner
Qwen3-7B	148	0.19 [0.13, 0.26]	0.81	45	0.11 [0.05, 0.23]	0.89
Gemma-2-9B	462	0.06 [0.04, 0.09]	0.94	15	0.00 [-0.00, 0.20]	1.00
Llama-3.1-8B	516	0.03 [0.02, 0.04]	0.97	157	0.03 [0.01, 0.06]	0.97
Qwen3-32B	76	0.13 [0.07, 0.23]	0.82	30	0.10 [0.04, 0.26]	0.80
Gemma-3-27B	227	0.04 [0.02, 0.07]	0.96	43	0.02 [0.00, 0.12]	0.95
Llama-3.1-70B	787	0.10 [0.08, 0.12]	0.36	401	0.03 [0.01, 0.04]	0.41
GPT-4o	10	0.20 [0.06, 0.51]	0.80	1	0.00 [0.00, 0.79]	1.00
Gemini-2.5-Flash	54	0.17 [0.09, 0.29]	0.81	49	0.16 [0.09, 0.29]	0.84
Claude-Sonnet-4	1	0.00 [0.00, 0.79]	1.00	1	0.00 [0.00, 0.79]	1.00
Falcon-Mamba-7B	208	0.53 [0.46, 0.60]	0.47	41	0.27 [0.16, 0.42]	0.73
Jamba-Mini-1.7	290	0.15 [0.12, 0.20]	0.64	165	0.19 [0.14, 0.25]	0.61
Falcon-H1-7B	61	0.03 [0.01, 0.11]	0.97	22	0.04 [0.01, 0.22]	0.95
Falcon-H1-34B	59	0.17 [0.10, 0.28]	0.64	3	0.00 [0.00, 0.56]	1.00
Granite-4.0-H-Tiny	112	0.49 [0.40, 0.58]	0.51	45	0.53 [0.39, 0.67]	0.47
Granite-4.0-H-Small	45	0.07 [0.02, 0.18]	0.87	17	0.00 [0.00, 0.18]	1.00
Nemotron-Nano-9B-v2	39	0.74 [0.59, 0.85]	0.26	32	0.81 [0.65, 0.91]	0.19
Nemotron-H-47B	98	0.23 [0.16, 0.33]	0.77	25	0.12 [0.04, 0.30]	0.88

Table 43: The hint axis for the eight held-out models. None is the no-hint baseline over the four wrapped cells, and H1, H2 and H3 are the rule hint, the one-shot example and the explicit length instruction. Compliance is the share of prompts on which the model followed the wrap instruction, and boundary failure is given over all prompts and over the compliant prompts alone.

	wrap compliance (F1)				boundary failure, all prompts				boundary failure, compliant prompts			
model	none	H1	H2	H3	none	H1	H2	H3	none	H1	H2	H3
Falcon-Mamba-7B	0.26	0.28	0.24	0.42	0.27	0.34	0.37	0.37	0.72	0.79	0.93	0.69
Jamba-Mini-1.7	0.92	0.91	0.96	0.91	0.96	0.97	0.95	0.98	0.96	0.96	0.95	0.97
Falcon-H1-7B	1.00	0.99	1.00	0.99	0.96	0.96	0.95	0.94	0.96	0.97	0.95	0.94
Falcon-H1-34B	0.99	0.99	0.98	0.98	0.96	0.97	0.65	0.80	0.96	0.97	0.65	0.79
Granite-4.0-H-Tiny	0.72	0.76	0.75	0.71	0.77	0.77	0.81	0.71	0.97	0.95	0.93	0.91
Granite-4.0-H-Small	0.95	0.83	0.96	0.97	0.94	0.94	0.92	0.94	0.95	0.94	0.92	0.95
Nemotron-Nano-9B-v2	0.99	1.00	1.00	1.00	0.97	0.95	0.92	0.98	0.97	0.95	0.92	0.98
Nemotron-H-47B	0.11	0.33	0.55	0.14	0.18	0.40	0.62	0.26	0.85	0.92	0.95	0.80

Positional encodings of the deployed models (detail for Section 4.2). Of the deployed models read in Section 5, the six open-weight Transformers use RoPE (Grattafiori et al., 2024; Gemma Team, 2024; 2025; Yang et al., 2025), as does the attention branch of Falcon-H1 (Zuo et al., 2025), while the attention layers of Jamba, Granite-4.0-H and the two Nemotron hybrids use no positional encoding (Lieber et al., 2024; IBM Granite Team, 2025; Blakeman et al., 2025; NVIDIA, 2025), and the three API models do not disclose theirs. Table 40 lists the sources. RoPE collapses at two layers and remains the lowest deep encoding at every depth, measured at the default base and without the base scaling that deployed models apply (Chen et al., 2023; Peng et al., 2024), so no sentence about the RoPE-based models of the cohort follows from it.

Tokenizers of the deployed models. The suite writes a fence of length n as n identical tokens. A deployed model reads the same fence through a subword tokenizer, and Table 44 shows how the tokenizers of fifteen of the seventeen deployed models split a run of 3 to 9 backticks. The tokenizers of Claude Sonnet 4 and Gemini-2.5-Flash are not public. Every vocabulary holds backtick-only entries for runs of one, two and three, and those of Gemma-2-9B and Falcon-H1-34B also hold one for four. Under o200k_base, the tokenizer of GPT-4o, the three entries are the tokens 63, 26178 and 168394, and a run of seven backticks is the vector (26178, 26178, 168394). Thirteen of the fifteen models write a run as pairs and end an odd run with the triple, and the other two write fours first. The default fence of three backticks is therefore one token, and the lengths 4 to 9 take two to five tokens.

Three differences from the unary runs follow. The number of tokens is about half the length, the alphabet has three or four symbols, and the parity of the length sits in the last token. The position

of the fence matters as well. For GPT-4o, Llama-3.1, Qwen3 and Granite-4.0 the last backtick of a fence that ends its line merges with the newline, so a closing fence of length four is a triple and a single where the opener before an info string is two pairs, and the two Nemotron models merge a final triple with the newline. Under these tokenizers a model cannot close by repeating the tokens of the opener. Whether the count axis survives this encoding is untested, and a control that trains CLOSE on merged tokens is left to future work.

Table 44: A run of n backticks under the tokenizers of the deployed models, written as the number of backticks in each token. Opener is a fence before an info string and closer a fence that ends its line, and an underlined token also holds the newline. The second column lists the lengths of the vocabulary entries made of backticks only. The suite writes the same run as n identical tokens.

models	entries	fence	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$	$n=9$
GPT-4o (o200k_base), Llama-3.1-8B, 70B, Qwen3-7B, 32B, Granite-4.0-H-Tiny, Small	1, 2, 3	opener	3	2+2	2+3	2+2+2	2+2+3	2+2+2+2	2+2+2+3
		closer	2+1	3+1	2+2+1	2+3+1	2+2+2+1	2+2+3+1	2+2+2+2+1
Gemma-2-9B, Falcon-H1-34B	1, 2, 3, 4	either	3	4	2+3	4+2	4+3	4+4	4+2+3
Gemma-3-27B, Falcon-Mamba-7B, Falcon-H1-7B, Jamba-Mini-1.7	1, 2, 3	either	3	2+2	2+3	2+2+2	2+2+3	2+2+2+2	2+2+2+3
Nemotron-Nano-9B-v2, Nemotron-H-47B	1, 2, 3	opener	3	2+2	2+3	2+2+2	2+2+3	2+2+2+2	2+2+2+3
		closer	3	2+2	2+3	2+2+2	2+2+3	2+2+2+2	2+2+2+3

Confounds of the cohort. Five confounds attach to every sentence about the cohort. The pure SSM class holds one model, Falcon-Mamba-7B, because the rules fixed before any generation removed the other candidates. The Mamba and Mamba-2 base checkpoints have no instruction tuning, so the wrap-compliance gate F1 has no meaning for them, Mamba-Codestral-7B has no chat template, mamba-chat 2.8B is a small community fine-tune of 2023 excluded for quality, the three Zamba2 instruct models have a 4,096-token context below the 7,168-token output budget, and RWKV-7 and RecurrentGemma are recurrent but outside the Mamba-based class on which the cohort was defined. Its content quality is low, with a heading-structure pass rate of 42.7%. The families were trained on different data. Compliance conditioning changes the denominators. The Transformer reference spans three tiers, so size is matched only within the 7 to 9 billion group.

Cohort (detail for Section 5). Falcon-Mamba-7B fails at a boundary in 21.0% of its 1,940 valid generations. On the 627 prompts that carry a wrap instruction it follows the instruction at a rate of 0.26, and among those it fails at 0.72. Each scored model produced 4,427 generations. The three API models sit in the large tier by LatentMD's own grouping, which places the models with undisclosed sizes with the largest open model.

The RoPE cells of the suite do not extrapolate count at any depth we trained, from 0.00 at two layers at three learning rates to 0.13 at twelve layers at lr 10^{-4} (Figure 3, Table 17). Beyond two layers the encodings other than learned positions were trained at lr 10^{-4} and at the anchor lr with three seeds, and RoPE also at 10^{-3} at eight, twelve and sixteen layers, so the statement covers this grid and no other rate. Section 5.1 reads it as consistent with the safe-length result of the deployed RoPE Transformers, and we do not carry it over as a cause, since those models differ in scale, data and training length.

Above the 7 to 9 billion size, Llama-3.1-70B is flat at 0.35 to 0.39 over the first 80 lines, the three API models stay at or below 0.05 and the two mid models below 0.15, while Nemotron-H-47B rises from 0.007 to 0.30 and Jamba falls from 0.16 to 0.09. Truncation occurs in 181, 197 and 230 records of the two highest-share models and Gemini-2.5-Flash against 0 to 19 elsewhere. Counting a truncated generation whose outer fence is still open as E2 turns the fall of Jamba into a rise of $+0.14$ and Gemini's $+0.04$ into $+0.19$. Exclusion removes the generations that kept writing without closing, and inclusion inflates the long bins mechanically because the token limit correlates with content length.

The E3 share of failures of the nine Transformers runs from 9% to 98%. Under the length instruction Gemini-2.5-Flash falls to 0.33, Llama-3.1-70B to 0.56 and Claude-Sonnet-4 to 0.07. Falcon-Mamba and Nemotron-H-47B follow the wrap instruction on 0.26 and 0.11 of the no-hint prompts, the hints

move that compliance to 0.24 to 0.42 and 0.14 to 0.55, and among the compliant prompts the failure rate rises under the rule hint and the example, from 0.72 to 0.79 and 0.93 for Falcon-Mamba and from 0.85 to 0.92 and 0.95 for Nemotron-H-47B, and stays near its baseline under the length instruction, at 0.69 and 0.80, so under the first two hints the worsening is not an artifact of newly compliant outputs, while the other hybrids stay at 0.65 to 0.98 and the one improvement is Falcon-H1-34B under the example, from 0.96 to 0.65. Over all prompts the three hints move Falcon-Mamba from 0.27 to 0.34, 0.37 and 0.37, and the other hybrids improve by at most 0.06.

In the short bins Llama-3.1-70B is the exception at 34.7%. The share of unclosed fences that never write another fence line is 0.74 for Nemotron-Nano-9B-v2, 0.53 for Falcon-Mamba-7B, 0.49 for Granite-4.0-H-Tiny, 0.23 for Nemotron-H-47B, 0.17 for Falcon-H1-34B, 0.15 for Jamba-Mini-1.7, 0.07 for Granite-4.0-H-Small and 0.03 for Falcon-H1-7B (Table 42), so it follows neither the SSM share nor the size tier. The other half of Falcon-Mamba's unclosed fences close the inner blocks and then end without closing the outer one, which is also the dominant end state of every Transformer but Llama-3.1-70B. Neither the pure SSM nor Jamba shows a short close, and the one model whose unclosed fences mostly end in a short close is Llama-3.1-70B, at 54% of 787, a Transformer. A generation that writes no inner code block writes no inner fence, which is why the content-restricted shares are reported next to the full ones.

Falcon-Mamba's wrap compliance of 0.26 also differs from the 0.99 and 0.72 of the two hybrids with the highest never-closed shares, and its share of 0.53 (208 fences) has a 95% Wilson interval of 0.46 to 0.60. Among the nine Transformers the never-closed end state takes 0.03 to 0.19 of the unclosed fences of the three 7 to 9B models. These facts come from the short bins and the end state and do not depend on the truncation rule.

What does not transfer in full (detail for Section 5.3). The Tail split probes the memory of a closed fence, which under-tuned synthetic models lose by closing again, and not the decision to close at all, which well-tuned 7B models omit, so the two failures point in opposite directions. The cohort adds no count measurement, so the count half of the dissociation is tested against LatentMD's Transformers only, and the cohort cannot weigh the post-training account against the architectural one. Of the synthetic hybrids only the one-of-eight cell is near a deployed attention ratio, and none is a parallel hybrid like Falcon-H1. OPEN opens after the content is seen, which is not the situation of a deployed model, and the families differ in training data. One E2 rate hides a never-attempted close and a lost close, which is why the end state should be reported. The same grammar applies to other surfaces, such as adherence to a default hash count in Rust raw strings, and we leave those as falsifiable future work.

K CELL INVENTORY

The full inventory of the campaign, 604 cells and 1,987 runs with seed means, standard deviations, gate counts, parameter counts and wall-clock times, is the file `cells.csv` of the released repository. The file `cited_cells.csv` in the same repository lists the 243 cells whose numbers appear in a figure, a table or the text of this paper, with the value of each seed and the place where the cell is cited, and the scripts of the repository produce every table and figure of the paper from these files and the run records. Table 45 gives the uncertainty of each row of Table 20 at the unit that varies, the run. Items within a run are not independent draws of a family's accuracy, so the interval is a percentile bootstrap over the runs of the row (10,000 resamples), and the last column of each half is the share of runs at or above 0.5.

Table 45: Table 20 with 95% bootstrap intervals over runs for the mean closing accuracy of the family rows of Figure 1 other than the single-Mamba stacks, on the Distance and Count splits, and the share of runs at or above 0.5.

		Distance closing accuracy			Count closing accuracy		
family	runs	mean	95% CI over runs	runs ≥ 0.5	mean	95% CI over runs	runs ≥ 0.5
Mamba, 1–16L	144	0.75	[0.71, 0.78]	0.85	0.00	[0.00, 0.00]	0.00
LSTM, 1–12L	126	0.55	[0.50, 0.59]	0.48	0.02	[0.01, 0.03]	0.00
GRU, 1–16L	65	0.79	[0.73, 0.85]	0.85	0.00	[0.00, 0.01]	0.00
RNN, 1–16L	52	0.89	[0.84, 0.94]	0.94	0.00	[0.00, 0.00]	0.00
Hybrid 4L, 0–2 attn layers	62	0.75	[0.69, 0.80]	0.81	0.01	[0.00, 0.03]	0.00
Hybrid 8L, 0–4 attn layers	34	0.88	[0.84, 0.92]	0.97	0.02	[0.00, 0.05]	0.00
S4D, 1–16L	160	0.18	[0.17, 0.20]	0.00	0.01	[0.01, 0.01]	0.00
Mamba-NoSelect, 2–16L	125	0.25	[0.23, 0.26]	0.00	0.00	[0.00, 0.00]	0.00
Transformer 1–2L, six PEs	226	0.19	[0.17, 0.21]	0.08	0.09	[0.07, 0.10]	0.01
Transformer 4–16L	148	0.30	[0.27, 0.33]	0.15	0.39	[0.34, 0.43]	0.36
Attention-only NoPE, 4–16L	44	0.18	[0.13, 0.23]	0.07	0.96	[0.94, 0.98]	1.00

L TASK DIFFERENCES FROM PRIOR SYNTHETIC BENCHMARKS

Table 46 places the closest prior synthetic settings beside the suite and claims no priority for any element.

Table 46: The closest prior synthetic settings, what each evaluates and what MBT changes, as read from the cited papers. Scores are not comparable across settings.

prior setting	what it evaluates	what MBT changes
Flip-Flop (Liu et al., 2023a), star-free and counter languages (Sarrof et al., 2024)	one bit written, ignored and read, and next-token prediction on 27 regular and counter languages at in-distribution and longer lengths	one fence length carried over gaps stretched twelvefold and reproduced at unseen values in the same transduction, with distractor runs and a fence that can close and reopen
counting (Weiss et al., 2018; Suzgun et al., 2019; Chang & Bisk, 2025), recognizers (Butoi et al., 2025), the EOS decision (Newman et al., 2020)	acceptance or next-token prediction on counting languages, and whether learning to stop lowers extrapolation	free generation scored as exact emission followed by STOP, VERIFY as the recognition counterpart, an emission control and a Count-Interp interpolation control
associative recall (Arora et al., 2024; Okpekpe & Orvieto, 2025) and repeated copying (Jelassi et al., 2024)	many key-value pairs or long strings, in the recall revisit trained and tested at one length under a learning-rate scan	one bounded value, out-of-distribution splits that no run trains on, and a five-rate scan per family with the lr-matched cell chosen on Distance
MAD (Poli et al., 2024), Retrievit (Pantazopoulos et al., 2026), subquadratic comparisons (Hartl et al., 2026)	hybrid designs on recall, selective copying, compression, retrieval, accumulation and state tracking	four- and eight-layer stacks of Mamba and NoPE attention blocks on exact boundary generation, with no score carried across tasks
Mamba-2 (Dao & Gu, 2024) and Mamba-3 (Lahoti et al., 2026)	later state updates and discretizations, tested on parity and modular arithmetic	the first published Mamba algorithm in the suite, and Mamba-2 blocks only inside the deployed cohort
attention glitches (Liu et al., 2023a), NoPE (Kazemnejad et al., 2023), randomized positions (Ruoss et al., 2023)	attention dilution and positional margins, removal of explicit position, and randomized indices in training	a position-index shift with the tokens fixed, real filler insertion, layer ablation and readout patching at two and eight layers