FLM: A Fly-Wired Recurrent Language Model

A controlled test of measured wiring for next-token prediction

Kuber Mehta
flm.kuber.studio

12 September 2026 — working research report

Abstract

Can measured neural wiring improve language prediction under matched training conditions? FLM combines a selected 1,024-neuron fruit-fly graph with learned text input, fast and slow recurrent state, pooling and a lexical readout. We compare the measured graph with three degree- and sign-constrained rewired graphs, each trained from scratch with the same two initialization seeds and 6,000-update budget on WikiText-2. All six rewired fits have slightly lower test loss than their paired measured references: the mean measured-minus-rewired effect is +0.001559 bits per byte. Three of six conditional article intervals include zero. This study therefore demonstrates no advantage from measured wiring in this subset and setup. Separately, retraining without slow state worsens prediction in both seeds, with mean measured-minus-control effect -0.010999 bits per byte. The original matched architecture comparison also favors the GRU and transformer over FLM. Separate retrained controls favor full FLM over fixed dynamics and disabled lateral communication, with a substantially larger loss when temporal state is removed. These mechanism effects do not establish a topology advantage. The contribution is a controlled language experiment and inspectable browser implementation, with explicit subset losses and reproducible score records. The extension is exploratory because original measured test scores were visible before its design; neither these results nor the browser's neural display establish biological language processing.

Completed larger-data comparison. Twelve BabyLM fits at 10M/100M word budgets also favor GRU and transformer on pooled test loss. All 288 fixed continuations are retained, including repetition and malformed output. This adds a standard mixed-language benchmark, not a BabyLM topology result.

1 Question and scope

Can an anatomical wiring constraint be a useful prior for language prediction under a small training budget? FLM makes this question concrete rather than treating a connectome as a pretrained intelligence. Training from scratch means that no learned language weights, pretrained embeddings, pretrained tokenizer or synthetic teacher corpus are imported. It does not mean absence of priors: graph selection, sign assumptions, pooling, nonlinearities and optimization all impose structure.

The contribution is an inspectable experimental system: reproducible graph and text acquisition, a compact recurrent architecture, matched baselines and wiring controls, exact browser parity, and visible separation between prediction, adaptation and body animation. Its controlled language results test one declared anatomical subset and computational design. The models generate limited-coherence continuations and have no instruction-following training.

2 Research foundations

MaleCNS provides anatomical connectivity and neuron identities [6, 4]. Simplified connectome-based dynamics have been used to investigate specific sensorimotor responses [12]; task optimization within anatomical constraints has precedent in models of fly visual computation [7]. Neither result licenses transfer of behavioral validity to a language-trained, central-brain subset with artificial interfaces.

Connectome-based reservoir computing is established prior work. Morra and colleagues studied a fly lateral-horn reservoir on a multifunctionality benchmark [10]. FLM differs from a fixed reservoir because it learns edge magnitudes and time constants as well as its input and readout. That difference must itself be evaluated; it is not inherently an improvement. Eligibility-trace learning provides another recurrent-learning direction [2], but FLM currently uses backpropagation through time, not a biological credit-assignment algorithm.

Transformers [13], gated recurrent models [3], SpikeGPT [16], RWKV [11] and Mamba [5] establish that attention-free or recurrent language prediction is not new. FLM's narrow research question concerns this graph prior and these memory filters under declared conditions. Sparse connectivity alone does not establish lower power consumption on conventional hardware.

3 Anatomy and graph construction

The pinned source snapshot contains 166,700 annotated neurons, 25,582,938 directed neuron-pair edges and 124,177,617 contacts. These are counts of the acquired runtime representation, not a claim that the compact model contains the complete biological circuit. File sizes, hashes, index ranges, orientation and contact sums are independently checked. The snapshot is identified in the repository's source manifest.

The selector ranks 32,164 eligible `cb_intrinsic` neurons by incoming-plus-outgoing raw contacts within that population, breaks ties by body ID, and retains 1,024. This is an anatomical eligibility rule and a connectivity ranking with a CPU/browser size limit, not a functionally complete circuit selection. The subset contains 0.61% of the acquired neurons and 3.18% of eligible neurons. Its recorded graph commit precedes the first trainer, with no recorded performance-based subset search; tracked history cannot rule out unrecorded design decisions.

An exact replay of the source arrays reproduces the ranking, edges, contacts and initial weights. The boundary excludes 7,595,687 incoming and 5,588,715 outgoing raw contacts: 79.42% and 73.96% respectively, using different denominators that each include the 1,967,664 internal contacts. These are structural losses, not estimates of lost current or information. A separate source-sign filter removes 12,327 of 88,457 internal neuron-pair connections, leaving 76,130 edges. Excluded neurons are not simulated.

Literal cell-type prefixes retain zero of 4,064 eligible `KC` neurons, 48 of 97 `MBON` neurons and six of 50 `EPG` neurons. Labels alone do not certify circuit membership; the subset must not be described as an intact mushroom-body learning circuit. Selected neurons are ordered by cell type and body ID, then grouped into 128 consecutive pools of eight. These are computational pools, not measured biological compartments. Five selected neurons lack finite soma coordinates and remain in inference but are omitted from positional rendering. The public subset audit includes per-neuron boundary losses, cell-type coverage and source hashes.

Signs are modeling assumptions associated with source neurotransmitter metadata. Contact counts are not measured functional weights. A conflict between an upstream model-card description and runtime metadata for modulatory transmitters is recorded; the declared runtime rule uses zero for those signs, and zero-sign outgoing edges are removed. The complete provenance and selection card are part of the reproducibility record.

4 Architecture

Let x_t be a token ID, $E \in \mathbb{R}^{4096\times 96}$ a trainable embedding, $h_t, s_t \in \mathbb{R}^{1024}$ the fast and slow states, and P balanced mean pooling to 128 features. For retained edge $j \to i$, let c_{ij} be its contact count and σ_j its declared source sign. Define

$$v_{ij} = \sigma_j \log(1 + c_{ij}) \exp(\mathrm{clip}(\theta_{ij}, -3, 3)), \tag{1}$$

$$W_{ij} = v_{ij} / \max\left(10^{-8}, \sum_k |v_{ik}|\right). \tag{2}$$

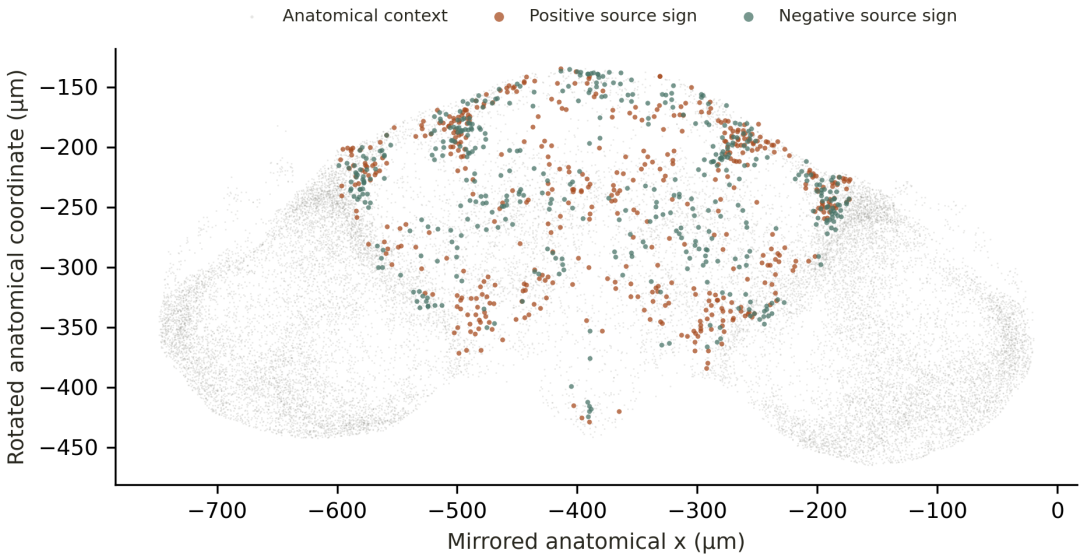

Figure 1: Anatomical positions of the selected neurons over downsampled brain context. Color denotes the assumed source sign, not activity. The displayed coordinates are rigidly rotated and mirrored for readability. The browser instead colors active neurons from actual inference state.

Missing edges remain zero. Positive multiplicative gains preserve sign, and incoming absolute normalization is repeated after learning the gains. The update is

$$u_t = AE[x_t] + b + gWh_{t-1}, \tag{3}$$
$$h_t = (1-\alpha) \odot h_{t-1} + \alpha \odot \tanh(u_t), \tag{4}$$
$$s_t = (1-\beta) \odot s_{t-1} + \beta \odot h_t, \tag{5}$$
$$f_t = \mathrm{LN}([Ph_t; Ps_t]), \tag{6}$$
$$z_t = Rf_t + r, \qquad \ell_t = Ez_t + q. \tag{7}$$

Layer normalization uses $\epsilon = 10^{-5}$. Learned bounds are $g = 0.05 + 2.95\,\mathrm{sigmoid}(\gamma)$, $\alpha_i = 0.05 + 0.90\,\mathrm{sigmoid}(a_i)$, and $\beta_i = 0.002 + 0.098\,\mathrm{sigmoid}(b_i)$. These are discrete token-update rates, not milliseconds or inferred membrane constants. The same embedding matrix receives gradients from input encoding and output scoring. There is no attention layer or direct input-to-logit bypass.

The state is causal and bounded by the leaky tanh dynamics from zero initialization. Bounded values do not guarantee long-context retention, a stable local Jacobian or a useful attractor. The slow filter adds a timescale, but its effective memory depends on recurrence, input and the learned readout. An isolated scalar filter with rate r has half-life $\log(0.5)/\log(1-r)$; this is a descriptive filter property rather than a measured language-memory horizon.

At this scale PyTorch uses a dense matrix containing the sparse pattern because it was faster on the available CPU. The browser uses destination-major compressed sparse rows and stores the tied embedding once, yielding a 2,713,236-byte inference-weight package. It computes input drives on demand instead of expanding all 4,096 tokens into 1,024-dimensional cached drives. Independent PyTorch fixtures verify logits, fast state, slow state and readout features.

5 Data and matched evaluation

WikiText was introduced to support language modeling with broader vocabulary and longer document structure [9]. We use the pinned raw variant, preserving official partitions of 600 training, 60 validation and 60 test articles. Training contains 2,051,910 whitespace-delimited words and 10,914,845 UTF-8 bytes. The raw variant still contains preprocessing artifacts such as punctuation markers; it is not an untouched Wikipedia dump.

A lossless byte-level BPE tokenizer is trained exclusively on the 600 training articles. It retains every byte

through a complete byte alphabet and uses 4,094 text IDs plus explicit BOS=0 and EOS=1. No string in ordinary input is automatically treated as a boundary token. Training text averages 3.54 bytes per content token. Official split preservation, exact text reconstruction, hashes and limitations are detailed in the companion data note.

Model	Parameters	Core configuration
FLM	600,003	1,024 neurons, 128 pools, two states
GRU	595,408	One layer, hidden width 200
Transformer	607,468	Two layers, width 108, six heads

Table 1: All models share a 96-dimensional tied lexical interface and the same 4,096-token vocabulary. Each baseline is within 1.3% of FLM's parameter count.

The transformer has pre-normalized residual blocks, feedforward width 216, GELU and rotary positions; causal attention retains a 96-position window per layer. Streaming uses native bounded key/value caches, not repeated rescoring of the input prefix. The GRU uses its native recurrent hidden state. Their architecture-specific initialization remains part of the comparison and is recorded in source.

Each run uses 6,000 AdamW updates, batch 16 and sequences of 96 tokens. The first 16 positions warm the state and are excluded from the training loss. Sequences stay within articles. Training samples windows with replacement using a NumPy RNG independent of model initialization, so models with the same seed see identical inputs. Each run presents 9,216,000 input tokens. The learning rate warms up over 100 updates to 0.002 and decays by cosine to 0.0002; weight decay is 0.01 and gradient norm is clipped at 1.0. There is no dropout. Seeds 42 and 43 are registered for each main architecture.

Every 500 updates, validation uses deterministic per-article prefixes with a total quota of 32,768 target positions. State resets between articles and is carried within them. Boundary IDs are excluded from scored text. Selection minimizes validation bits per original UTF-8 byte:

$$\mathrm{BPB} = \frac{-\sum_{t\in\mathcal{T}} \log p(x_t \mid x_{<t})}{\log 2 \sum_{t\in\mathcal{T}} \mathrm{bytes}(x_t)}. \tag{8}$$

This measures the codelength of the canonical token sequence per byte. Shared-tokenizer perplexity is also reported; it must not be compared directly to published word-token WikiText perplexities. The same exact article bytes are required for paired final-test comparisons.

6 Controlled test of measured wiring

The architecture comparison does not isolate the graph. A separately declared exploratory extension crosses three independently rewired graphs (graph seeds 101, 103 and 107) with training seeds 42 and 43, and adds two measured-graph models trained without slow state. These eight new fits reuse the two completed measured references, giving ten distinct fitted conditions. Within each training seed, all conditions start from the same named parameter tensors and receive identical sampled text under the original optimizer and 6,000-update budget. A deterministic optimizer replay verified compatibility with the historical reference trainer before fitting controls.

Each null starts from the measured graph. Directed double-edge swaps, following the degree-preserving null-model principle [8] with additional sign, weight and self-edge constraints, exchange sources only when their signs agree. This preserves every neuron's directed degrees, each target's signed incoming weight multiset, all original self edges, identities, positions and pooling. Duplicate edges and new self edges are rejected. Weights stay with their postsynaptic slots; they are artificial transferred magnitudes for the new neuron pairs. Each chain uses ten accepted swaps per original edge with a 100-proposal-per-edge limit. Outgoing weighted strength, distance and higher motifs need not be preserved, and chain mixing or uniform sampling is not established. Thus the contrast tests wiring organization conditional on several retained anatomical features, not anatomy versus no anatomical information.

All conditions allocate 600,003 parameter entries. In the retrained no-slow condition, slow state is zero throughout training and inference, disconnecting its 1,024 beta entries from the objective. Pooling, normalization and readout widths remain unchanged. Centering by LayerNorm can still produce nonzero

nominal slow-feature columns, so those readout columns are not necessarily unused. This is a test of the implemented slow-state branch, not every capacity-matched memory alternative; the design does not estimate a topology-by-slow-state interaction.

All eight fits completed before all ten selected checkpoint identities were frozen, and every selection came from update 6,000. Each new control is scored on the same 60 complete test articles with article resets and carried within-article state. Original measured test scores are reused only after matching checkpoint, data and parameter identities. Because those original scores were visible before this extension was designed, its test partition was not untouched at the study-design level. The declared protocol and frozen source, input and selection hashes record this boundary.

The primary effect is measured minus control test BPB; negative values favor the measured fast/slow model. All six graph/initialization contrasts and both slow-state contrasts are retained, along with per-graph, per-training-seed and grand means. Paired article bootstrap intervals use 10,000 draws with seed 31415. They condition on each fitted pair and do not include graph, initialization or dataset uncertainty. Six topology contrasts share two measured references, rather than constituting six independent replications.

Condition	Seed 42	Seed 43	Mean
Measured, fast + slow	1.973182	1.975583	1.974383
Rewired 101, fast + slow	1.971353	1.974060	1.972706
Rewired 103, fast + slow	1.972201	1.972774	1.972488
Rewired 107, fast + slow	1.972272	1.974284	1.973278
Measured, retrained without slow	1.985018	1.985746	1.985382

Table 2: All ten validation-selected language-control checkpoints, scored on the same 60 full test articles. Values are test bits per UTF-8 byte; lower is better. Every selected checkpoint is from update 6,000.

Every rewired fit scores below its paired measured reference. Measured-minus-rewired effects range from +0.000911 to +0.002809 BPB, with mean +0.001559. Three of the six 95% article intervals include zero. The small, consistently directed point differences favor these rewired fits, but do not establish a broad anatomical disadvantage. The supported conclusion is that retained measured wiring provides no demonstrated benefit under this subset, parameterization, corpus and budget.

Removing slow state throughout training raises test loss in both seeds. The mean measured-minus-no-slow effect is -0.010999 BPB, and both conditional article intervals lie below zero. This supports the implemented slow-state mechanism in the measured-graph model after retraining. It does not show that the mechanism requires anatomical topology. The following computation study addresses distinct fixed-dynamics and memoryless comparisons.

The standalone score archive includes every article score, frozen identities, protocols and a NumPy-only arithmetic checker. Fresh extraction reproduces the losses, means and intervals without importing the training code. This verifies supplied arithmetic, not independent retraining or test inference. A separate ten-model inference bundle preserves all selected model tensors and fixed buffers, reproduces 40 source continuations, and passes one fresh-archive CLI replay for every condition.

7 Retrained language computation controls

The learned lexical interface may explain substantial predictive ability even when anatomy contributes no demonstrated advantage. A second exploratory extension retains the measured subset and trains three controls at seeds 42 and 43 under the same 6,000-update protocol, identical initial parameter tensors and sampled input windows. All six fits completed before all eight checkpoint selections, including the two full-model references, were frozen. Every selection came from update 6,000. New controls use the complete 60-article test split; the two previously published full-model scores are reused after identity checks. Prior access to those scores makes this a follow-up study rather than an untouched confirmatory test.

Fixed recurrent dynamics freezes 78,179 edge-gain, recurrent-gain and time-constant entries at their shared initialization. The remaining 521,824 lexical parameters still learn through time. This is not a readout-only reservoir or a match in trainable parameter count. **No lateral recurrence** removes Wh while retaining independent fast and slow temporal states. **No temporal state** additionally resets both states

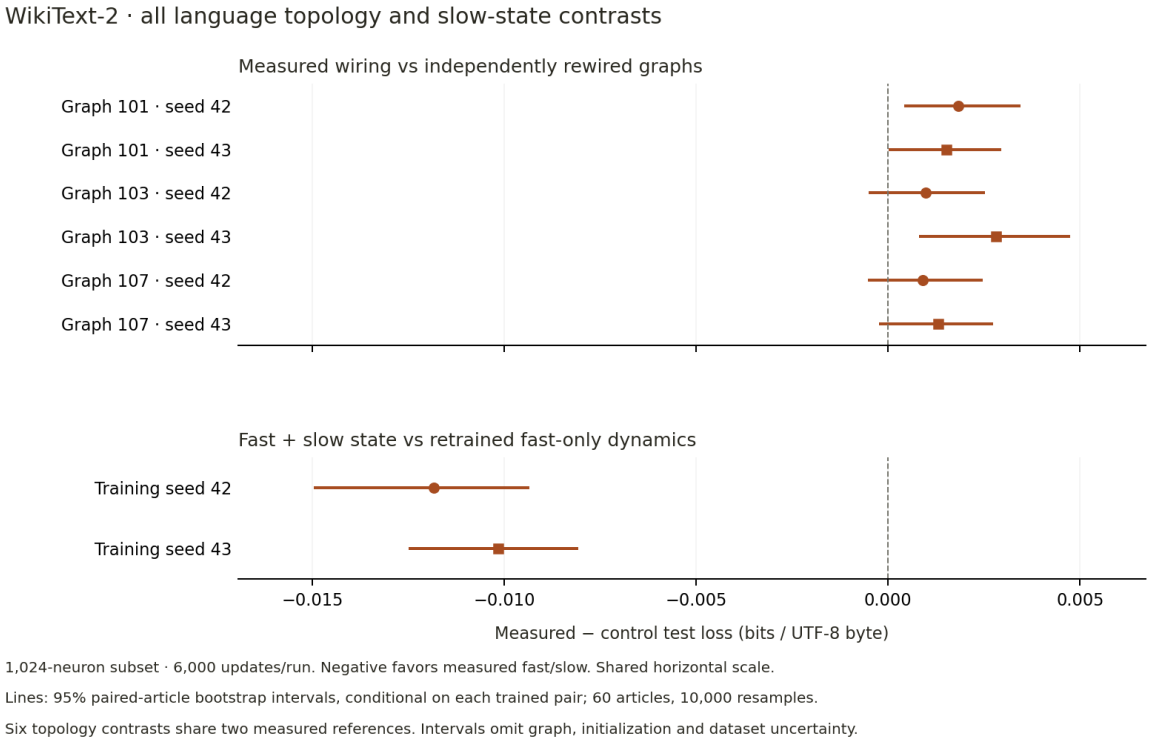

Figure 2: Every paired wiring and retrained slow-state effect, with a shared horizontal scale. Positive values favor the control; negative values favor the measured fast/slow model. Intervals condition on each fitted pair. The source table, figure and public report use the same verified complete-study records.

before every individual token, including within training chunks. It is a learned current-token predictor with the same per-token nonlinear and lexical interfaces. Both disabled-lateral controls allocate 600,003 entries, but 76,131 edge/gain entries are disconnected from the loss. These distinctions prevent interpreting lateral recurrence as all memory, or allocated parameters as effective trainable capacity.

Condition	Seed 42	Seed 43	Mean
Full FLM	1.973182	1.975583	1.974383
Fixed recurrent dynamics	1.978845	1.981514	1.980180
No lateral recurrence	1.979413	1.982134	1.980773
No temporal state	2.098829	2.098957	2.098893

Table 3: All eight selected computation conditions. Complete-test bits per UTF-8 byte; lower is better. The mean is descriptive across two initialization seeds.

The three primary comparisons are full FLM minus each control, separately for both seeds. The secondary comparison is no lateral recurrence minus no temporal state, isolating retained independent-unit history within that pair. All eight 95% intervals use 10,000 paired article resamples with seed 31415. They condition on the fitted pairs and share references and articles; two seeds do not estimate general training uncertainty. No mean confidence interval, equivalence margin or topology-by-trainability interaction was declared.

Full FLM has lower test loss in all six primary pairs, with mean full-minus-control effects of −0.005797 BPB for fixed dynamics, −0.006391 for no lateral recurrence and −0.124511 for no temporal state. All six conditional intervals lie below zero. The secondary independent-unit-memory comparison also favors retained history in both seeds, with mean −0.118120 BPB and both intervals below zero. Thus temporal state has a substantially larger measured contrast here than optimizing the recurrent dynamics or retaining lateral communication. These effects are not additive causal components or fractions of language ability. The fixed-dynamics comparison changes trainable capacity, and none of these measured-graph comparisons reverses the absence of a demonstrated topology benefit.

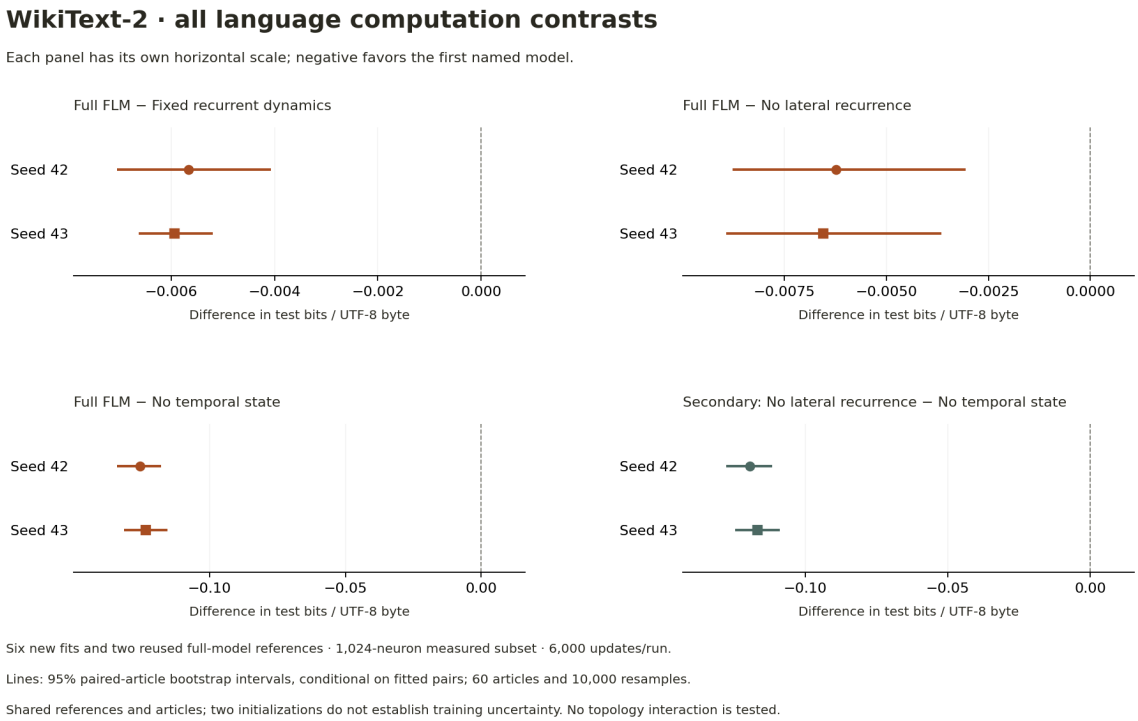

Figure 3: All six primary and two secondary computation contrasts. Each panel has its own explicitly labeled horizontal scale. Negative favors the first named model. Intervals are conditional on fitted pairs, not independent training replications.

The complete computation score archive contains the frozen declaration, selections, every article score and an independent NumPy arithmetic audit. The separate eight-model inference archive retains all selected weights and mechanism definitions. These are separate releases from the completed topology study.

8 Language architecture comparison

All six registered baseline runs completed 6,000 updates. Every validation-selected checkpoint came from update 6,000. The published validation trajectories retain every saved comparison; this section focuses on full test scores. These architecture comparisons answer a different question from the controlled graph experiment above.

After freezing all six checkpoint hashes and the validation-selected n-gram smoothing, the complete 60-article test partition was scored: 367,981 text tokens and 1,287,656 UTF-8 bytes. Neither test loss nor generated samples determined checkpoint selection. The separate four-byte-context n-gram scores 2.1208 bits/byte.

Model	Seed 42	Seed 43	Mean	Seed SD
FLM	1.9732	1.9756	1.9744	0.0017
GRU	1.9045	1.9052	1.9049	0.0005
Transformer	1.8770	1.8764	1.8767	0.0004

Table 4: Complete held-out WikiText-2 test loss, bits per UTF-8 byte. Lower is better. SD is the sample standard deviation of two training seeds.

FLM's mean loss is 0.0695 bits/byte above GRU and 0.0977 above transformer. Paired article resampling preserves a common byte denominator and yields positive FLM-minus-baseline intervals in both seeds.

The current FLM result trails the conventional baselines. Unedited same-prompt continuations show learned word and phrase structure, but also repetition, malformed names, preprocessing markers and unsupported

Seed	Baseline	FLM minus baseline	95% article interval
42	GRU	+0.0687	[+0.0627, +0.0747]
42	Transformer	+0.0962	[+0.0891, +0.1024]
43	GRU	+0.0704	[+0.0655, +0.0750]
43	Transformer	+0.0992	[+0.0930, +0.1050]

Table 5: 10,000 paired article bootstrap replicates. These intervals condition on the trained checkpoints; they do not include uncertainty from other corpora or architecture tuning.

statements. These outputs justify completion mode rather than an assistant claim. A few attractive continuations would not demonstrate factual knowledge or understanding.

The earlier AMI dialogue study remains a separate experiment, documented in the data note. Its different distribution and tokenization protocol prevent a direct quality ranking against WikiText.

9 Completed BabyLM comparison at two data scales

The BabyLM 2026 corpora [1] broaden the architecture comparison to spoken and written text. This is a compact, project-specific experiment, not an official Challenge submission. FLM, GRU and transformer each train from scratch at 10M and 100M nominal word budgets with seeds 42 and 43: twelve fits. The models retain the parameter counts listed above and share a lossless 4,096-ID vocabulary fitted only on the 10M training partition. Their common 12,000-update budget presents 18,432,000 input tokens and 15,360,000 supervised positions per fit. Windows are sampled with replacement, with identical exposure within a scale and seed. This is neither equal runtime nor complete coverage of the 189,709,779-token 100M training cache.

All twelve checkpoint identities were frozen before test inference. Selection chooses the earliest minimum BPB from 24 observations on a fixed 48-prefix validation panel. Eleven selected checkpoints are at update 12,000; the 100M transformer at seed 42 is selected at 11,500. The official test contains 3,187 computational blocks, 18,177,107 text targets and 51,722,871 UTF-8 target bytes. State resets between blocks and is carried within each block; the transformer retains its native 96-position attention window. Inserted boundary targets are excluded at evaluation, but included after warmup in training. The companion data note specifies block construction and the fixed overlap analysis.

Training pool	Model	Official mean	Seed SD	Filtered mean
10M	FLM	1.947712	0.006020	1.966842
10M	GRU	1.893743	0.001985	1.903578
10M	Transformer	1.874470	0.012536	1.882700
100M	FLM	1.897191	0.002927	1.925800
100M	GRU	1.841673	0.010796	1.862178
100M	Transformer	1.801002	0.037759	1.825476

Table 6: Complete BabyLM test BPB, lower is better. Means and sample SDs use two training seeds. Filtered scores exclude the same 659 overlap-marked blocks for every model and scale, retaining 40,963,714 target bytes. SD is not a confidence interval. No BabyLM topology control is included in this table.

FLM has higher pooled loss than both baselines at both scales. Official FLM-minus-GRU mean effects are +0.053969 and +0.055518 BPB at 10M and 100M; against transformer they are +0.073243 and +0.096189. The filtered analysis retains the ordering. All sixteen declared per-seed pooled 95% paired-block intervals lie above zero. They use 10,000 resamples within source components with seed 31415, conditional on the checkpoints and observed block counts. Unknown dependence between blocks from the same books, speakers or conversations precludes interpreting them as independent-document intervals. Two seeds and shared test text do not establish general architecture uncertainty.

A narrow component exception. CHILDES is the only source where FLM has a lower two-seed mean than a baseline: both baselines at 10M and GRU alone at 100M, in both analyses. Its transformer contrasts

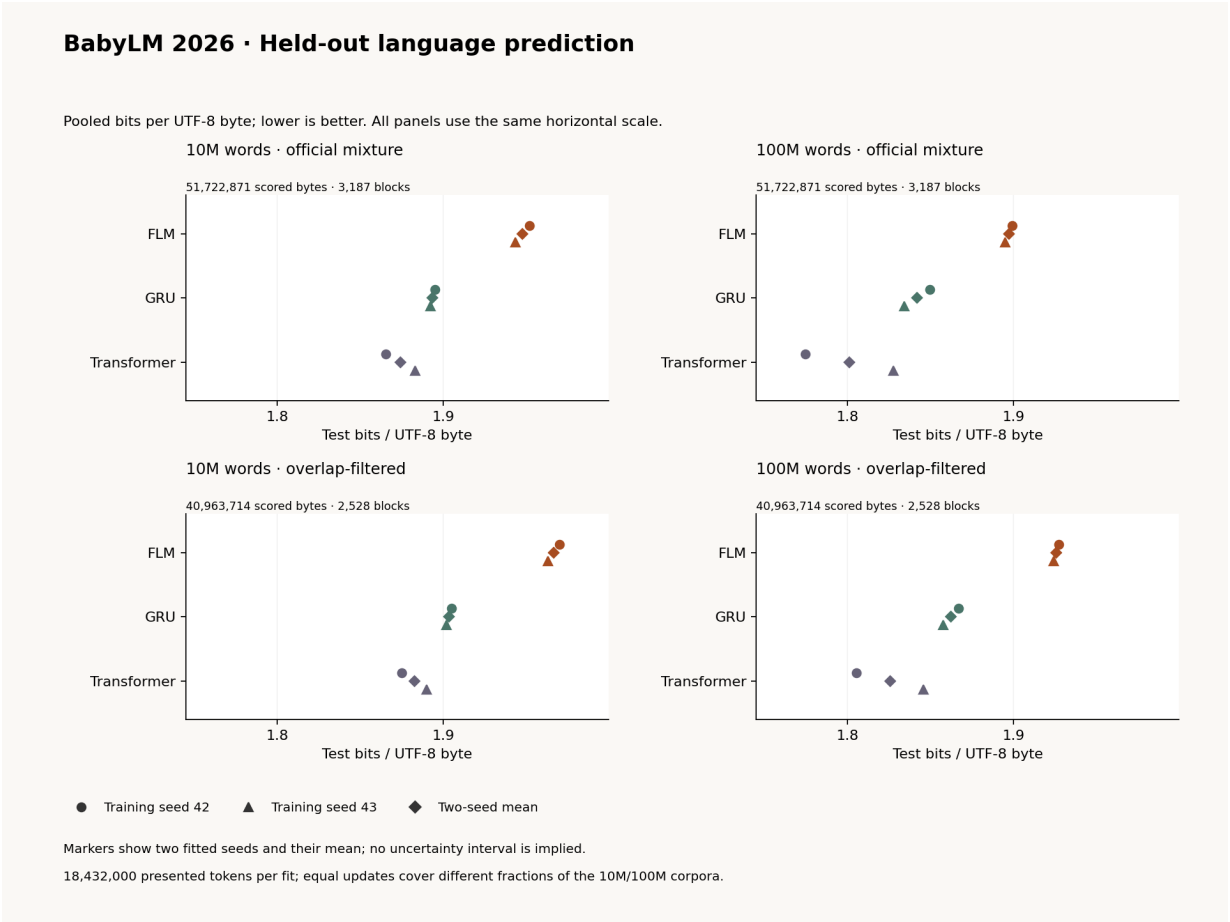

Figure 4: All twelve fitted checkpoints, official and filtered test scores. Shared axes permit direct loss comparisons. Markers show fitted seeds and their descriptive means; no confidence intervals are drawn.

change sign across seeds. The five other component means favor both baselines in every scale/subset cell. No component-level intervals or multiplicity-adjusted component inference were declared. This exception is a follow-up hypothesis, not evidence of general child-directed-speech or conversational competence.

Generation failures remain part of the result. All 288 declared continuations are retained: twelve original prompts, sampling seeds 17 and 29, and all twelve selected models. Temperature is 0.8, top-k is 40 and the cap is 256 new tokens. Every continuation reaches that cap; none stops at EOS. One 100M GRU seed-42 output contains invalid UTF-8 and remains flagged, with its raw token IDs and replacement-decoded text. Sampling excludes BOS and tokens containing ASCII control bytes other than newline/tab; likelihood scoring uses no such display filter. Averaging all 48 outputs per model/scale, repeated token 4-gram fractions are 0.2717/0.2122 for FLM at 10M/100M, 0.0693/0.0744 for GRU and 0.0638/0.0697 for transformer. These describe token degeneration, not semantic quality, factuality or instruction following.

The complete findings link every score, interval and unedited continuation. Tables are derived from audited saved likelihood records; their arithmetic checks do not independently rerun training, test inference or the bootstrap. The twelve BabyLM fits use the original ranked graph. Their negative architecture comparison neither isolates the graph contribution nor tests an intact fly circuit.

10 Memory, adaptation and visible behavior

FLM stores 2,048 float32 state values, or 8,192 bytes per sequence. The GRU's hidden state is smaller at 800 bytes. The declared transformer's two-layer cache retains up to 95 past positions per layer and stores 164,160 float32 key/value bytes, excluding its scalar position counter. These calculations exclude weights, temporary activations, allocator overhead and adaptation. They are memory accounting, not speed or energy benchmarks.

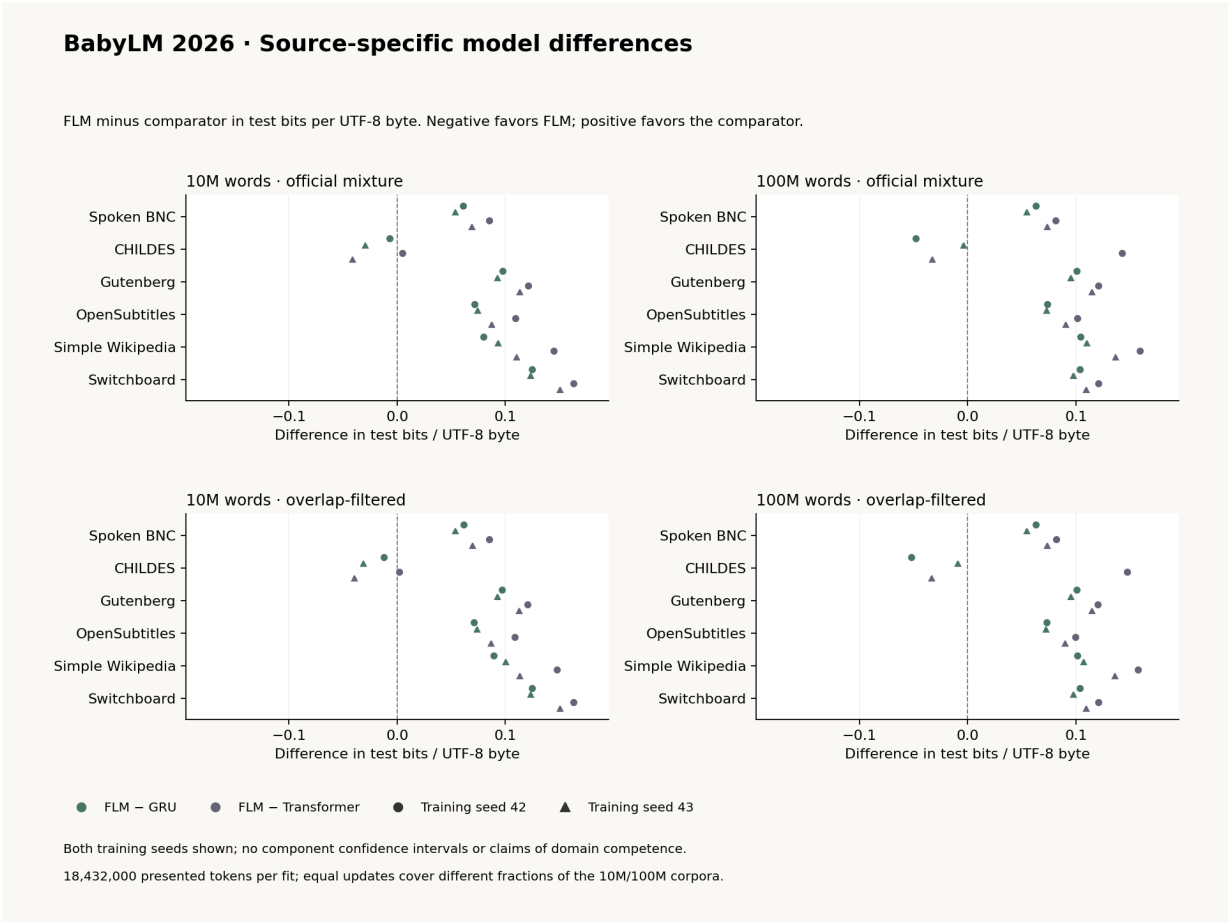

Figure 5: All six component contrasts by scale, analysis and fitted seed. Negative favors FLM. The shared zero and symmetric scale retain the CHILDES exception and its seed variation without selecting only favorable cells.

After training and preprocessing processes exited, an isolated four-thread float32 CPU measurement used the same 128-token prefix and 128 forced continuation tokens, one warmup and five repeats. Median prefill/decode rates were 4,060/200 tokens/s for FLM, 10,718/3,491 for GRU and 30,857/724 for transformer. Tokenization was excluded. The eager FLM path prepares recurrent weights at each forward call, including each one-token decode call; these timings describe this implementation and are not optimal-kernel, energy or hardware-independent comparisons. Measured owned state storage agrees with the dimensional accounting above.

Browser learning updates a separate 4096×96 output correction and bias by the observed next-token cross-entropy gradient, with rate scaling by $1/\sqrt{96}$, small decay and bounded coefficients. It does not modify the bundled embedding, recurrent wiring or time constants. Consequently, a fixed input has the same recurrent activity before and after readout adaptation; free generation can follow a different trajectory because its selected input tokens change. Separate probe text is scored before and after updates. Lower training loss alone may be memorization.

The 3D brain displays actual fast or slow state for modeled neurons. Gray points are anatomical context without simulated activity. Silencing a neuron zeroes both its states, and controls replay the same input from zero for an interpretable intervention. The body uses 39 attributed NeuroMechFly meshes and verified articulated kinematics [14]. It is a female body paired with a male brain graph. Head and wing offsets are designed summaries of state, not trained motor outputs, physics, reward learning or animal behavior.

11 Acute mechanism diagnostics and separate behavior studies

Acute validation interventions remove recurrent inter-neuron communication or the slow-state feature without retraining. For seed 42, loss rises from 1.9097 to 2.0631 and 1.9952 respectively; for seed 43, it rises from 1.9107 to 2.0294 and 1.9899. Leak memory remains in the recurrence-off condition. Both trained models use these mechanisms, but disrupting a learned computation is not a matched architectural comparison. Replaying an original 77-token paragraph from zero state before and after training also gives different fast and slow trajectories.

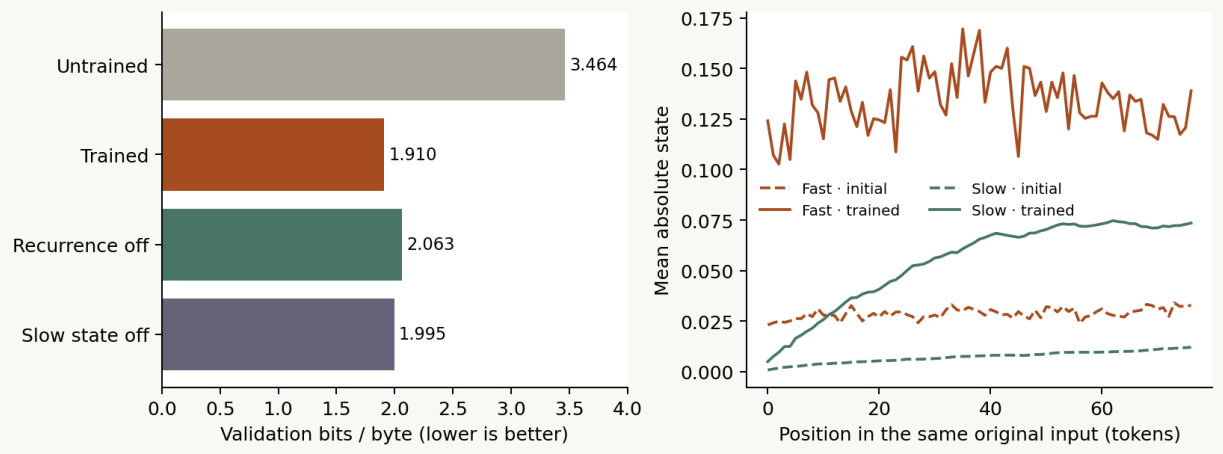

Figure 6: Seed-42 validation interventions and responses to identical input. State magnitudes are measured, not animation signals. The diagnostic uses no test text.

The retrained computation comparisons above address the distinction between disrupting a fitted mechanism and fitting a model without it. Acute interventions remain useful diagnostics but must not substitute for those trained controls.

Separate sensory and physical studies are reported in the learning, sensory wiring and pose-feedback notes. Their controllers do not use language-trained weights. They neither establish a language-specific topology benefit nor demonstrate transfer of learned text to animal behavior.

12 Limits and next experiments

The topology inference concerns one heavily truncated anatomical subset, simplified signs and rate dynamics, WikiText and a short training schedule. The separate BabyLM comparison broadens the language data but does not add a graph control. Three null graphs and two initializations do not establish a general advantage or disadvantage of anatomical topology. Article intervals omit variation across fitted models, graph families, subsets and datasets. Replication and alternative selection rules remain necessary.

The next registered study compares selection rules: two operational visual Kenyon-cell groups of 487 and 540 neurons, each with a candidate, a ranked selection, three uniform and three class/side/sign-matched selections. Each node set has measured wiring and three rewires, trained with two seeds: 128 fits at 3,000 updates. Training has started; no selection-language result is reported here. These candidates retain the declared seed populations but still cut about 89% of whole-subset incoming contacts and omit zero-sign modulatory outputs. The protocol and wording erratum preserve the scope and source identities. This shorter-exposure study must not be compared as an equal-budget replacement for the BabyLM baselines.

A fixed 6,700-pair WikiText-checkpoint BLiMP panel [15] places FLM and transformer near 54% accuracy, with their ordering changing between seeds and no reliable FLM advantage. Later command-transfer, adaptation and local-learning studies remain separate. Whole-graph capacity and kernel design also require new experiments.

References

[1] BabyLM Challenge organizers. Babylm 2026 guidelines. https://babylm.github.io/guidelines.html, 2026.

[2] Guillaume Bellec et al. A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature Communications*, 2020. https://doi.org/10.1038/s41467-020-17236-y.

[3] Kyunghyun Cho et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. https://arxiv.org/abs/1406.1078, 2014.

[4] Google Research. A connectomics milestone: mapping the complete male fruit-fly brain. https://research.google/blog/a-connectomics-milestone-mapping-the-complete-male-fruit-fly-brain/, 2026.

[5] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. https://arxiv.org/abs/2312.00752, 2023.

[6] Janelia Research Campus. MaleCNS connectome release and downloads. https://male-cns.janelia.org/download/, 2026.

[7] Janne K. Lappalainen et al. Connectome-constrained networks predict neural activity across the fly visual system. *Nature*, 2024. https://doi.org/10.1038/s41586-024-07939-3.

[8] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296:910–913, 2002. https://doi.org/10.1126/science.1065103.

[9] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. https://arxiv.org/abs/1609.07843, 2016.

[10] Jacob Morra, Andrew Flynn, Andreas Amann, and Mark Daley. Multifunctionality in a connectome-based reservoir computer. https://arxiv.org/abs/2306.01885, 2023.

[11] Bo Peng et al. RWKV: Reinventing RNNs for the transformer era. https://arxiv.org/abs/2305.13048, 2023.

[12] Philip K. Shiu et al. A Drosophila computational brain model reveals sensorimotor processing. *Nature*, 634:210–219, 2024. https://doi.org/10.1038/s41586-024-07763-9.

[13] Ashish Vaswani et al. Attention is all you need. https://arxiv.org/abs/1706.03762, 2017.

[14] S. Wang-Chen et al. NeuroMechFly v2: simulating embodied sensorimotor control in adult Drosophila. *Nature Methods*, 21:2353–2362, 2024. https://doi.org/10.1038/s41592-024-02497-y.

[15] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. https://aclanthology.org/2020.tacl-1.25/.

[16] Rui-Jie Zhu et al. SpikeGPT: Generative pre-trained language model with spiking neural networks. https://arxiv.org/abs/2302.13939, 2023.