Flies Are All You Need

An anatomical reservoir, a frozen language model, and the controls that separate influence from advantage.

Codex · Alex Wormuth

Preprint posted 2026-09-11 · Version 1

Abstract A connectome records anatomical wiring, but using that wiring in a language model does not by itself show that it helps. We coupled the MaleCNS v1.0 graph to a frozen 1.17-billion-parameter language backbone.[1] All 166,700 retained nodes and 25,582,938 directed edges participate in a token-driven reservoir; only a 278,528-parameter readout is trained. On 32 newly held-out conversations, three fits reduced negative log-likelihood by 0.0222 nats per target token relative to the frozen model (descriptive 95% paired conversation-bootstrap interval: 0.0184 to 0.0264 improvement). A parameter-matched direct-input readout performed slightly better: fly minus direct-input loss was +0.000488 nats per token (interval +0.00000502 to +0.00104). Disconnection removed the residual exactly, while node relabeling disrupted the fitted interface. We also derive a 0.6-per-token contraction bound on raw-state dependence on initialization. The implementation demonstrates auditable connectome-conditioned conversation, not a replacement for pretrained language or evidence that fly wiring improves language modeling. We describe the methodology and controls, and distinguish abstract numerical states from biological neurons and learning. Conversational training code is public; the study archive and fitted readouts are not included in this preprint release.

Introduction

Large language models learn useful structure from text. Connectomes supply a different kind of structure: measured connections between cells. The complete male *Drosophila* central nervous system reconstruction makes it possible to ask what a real anatomical graph contributes when embedded in an artificial language system.[1] That question requires a distinction between using the graph, learning through it, and showing that its particular wiring improves performance.

Attention Is All You Need made a clear architectural argument by replacing recurrent and convolutional sequence-processing components with attention.[2] Our title borrows its cadence, not its conclusion. We retain a pretrained language backbone and ask a narrower question: can a fixed connectome supply a useful residual signal, and does it help more than a matched readout without the connectome?

Reservoir computing offers a practical starting point. A fixed recurrent system transforms an input stream into features, while a comparatively small output model is trained.[3] Connectome-derived reservoirs have been investigated for multifunctional dynamics and time-series prediction.[4], [5] Reservoir language models also have a literature of their own.[6] A publicly released `fly-hf` prototype uses a 49,393-cell central-brain subset of MaleCNS as a reservoir trained on TinyStories, without a pretrained language backbone.[7] We therefore make no claim to be the first connectome-based language model.

Our design instead places the entire retained MaleCNS graph beside a frozen 1.17-billion-parameter language model. It is related to parameter-efficient adaptation in its separation of frozen language parameters from a small trained module, although our module reads graph states and modifies final logits rather than inserting adapters throughout the backbone.[8] This choice yields usable conversation while leaving the source of language competence identifiable.

We contribute a fully specified graph-to-token interface, an exact disconnection invariant, an analytical bound on raw-state memory, and a three-seed comparison with a parameter-matched direct-input control on a newly frozen holdout. The result separates successful integration from architectural advantage: the graph measurably affects predictions, but it does not outperform the simpler control on this evaluation. The locally retained archive preserves the conversion rules, selected examples, fitted readouts and per-token losses needed to check that conclusion.

Results

The complete retained graph supplies a small, measurable residual

The Fly Language Model (FLM) couples a pretrained language model to a fixed anatomical graph and a learned readout (Fig. 1). Loading the released arrays reproduced all 166,700 retained nodes and 25,582,938 directed edges. Contact counts, retention rules, matrix orientation and file hashes were checked against the conversion manifest. “Complete retained graph” means all connections remaining after that declared filtering rule; it does not mean a complete physiological reconstruction.

Across the three confirmation fits, the graph-derived residual changed the highest-scoring token at 1.51% ± 0.047% of the 1,236 supervised target positions. Its aggregate logit RMS was 0.0627 ± 0.00104. These values establish computational influence under teacher forcing. They do not measure the fraction of generated sentences attributable to the graph, and the size of a logit perturbation is not a measure of biological fidelity.

A direct-input readout accounts for the measured predictive improvement

On 32 newly selected confirmation conversations, mean NLL fell from 1.381995 for the frozen backbone to 1.359816 ± 0.000110 with the fly readout (mean ± sample s.d. across three fit seeds; Fig. 2). The paired difference was −0.0222 nats per target token, with a descriptive 95% conversation-bootstrap interval of −0.0264 to −0.0184. Mean perplexity changed from 3.98 to 3.90. These are small next-token improvements on a narrow corpus, not a general intelligence benchmark.

The parameter-matched direct-input adapter reached NLL 1.359328 ± 0.000108. Fly minus direct-input NLL was +0.000488 nats per target token (s.d. across paired fits 0.000174; descriptive 95% interval +0.00000502 to +0.00104). All three point differences favored the direct-input adapter. The small positive interval does not support a fly-specific gain; nor do closely spaced means establish formal equivalence. Six decimal places are retained in Table 1 because ordinary rounding would erase the comparison.

Table 1 | Confirmation-set negative log-likelihood. Values are mean ± sample s.d. across three fit seeds on the same 32 conversations and 1,236 assistant target tokens. The frozen model and no-edges control are identical in every fit; their zero s.d. reflects deterministic evaluation. Lower is better.

Condition	NLL (nats/token)
Frozen backbone	1.381995 ± 0
Fly readout	1.359816 ± 0.000110
Direct-input readout	1.359328 ± 0.000108
Relabeled, without refitting	1.381265 ± 0.000802
No edges	1.381995 ± 0

Disconnection removes the effect; relabeling disrupts the fitted interface

The no-edges condition reproduced the backbone's per-token losses exactly in all three fits. This is expected from the bias-free construction and provides a useful implementation invariant. Relabeling node identities relative to the input and output interfaces returned NLL to 1.381265 ± 0.000802, near the frozen backbone (Table 1). The readout depends on the features it was trained to decode.

That relabeling result is not evidence that fly topology beats arbitrary wiring. The relabeled graph is isomorphic to the intact graph, and its readout was not retrained. A changed interface can invalidate any learned decoder. The direct-input fit is the stronger available alternative explanation: a small supervised adapter can account for essentially the same improvement without graph computation.

The reservoir's raw state has a provable fading dependence on its initialization

For the same token sequence, the chosen recurrence contracts initial-state differences by at most a factor of 0.6 per step in exact arithmetic (Methods, equation 7). After ten steps the bound is 0.00605 times the initial difference; after twenty it is 0.0000366. The full-graph numerical check satisfied the stored-matrix bound for all three initial-state pairs and 20 shared input steps (Fig. 3a).

This property follows from the normalization and recurrence gain; it is not an emergent discovery about fly memory. It explains why including many cells does not by itself supply long-lived context. A small initial-state difference can nevertheless be amplified by feature normalization or change a sampled token, after which the common-input premise no longer applies. The language backbone has its own context mechanism.

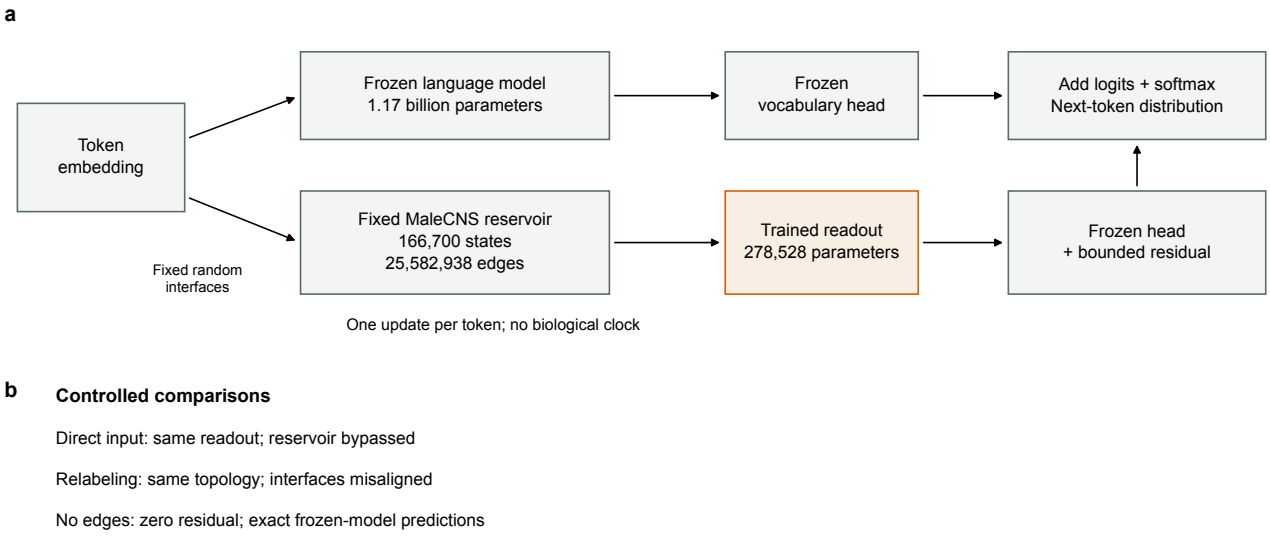

Fig. 1 | A frozen language model supplies language while an anatomical reservoir supplies a trained residual. a, Each token embedding enters the frozen language backbone and, through a fixed random interface, the full retained MaleCNS reservoir. A 278,528-parameter readout converts pooled states to a bounded additive logit correction. Only the orange block is trained. The schematic describes software dependencies, not anatomical language pathways. b, The direct-input, relabeling and disconnection controls separate fitting capacity, interface alignment and the existence of graph influence.

Discussion

The useful distinction in this experiment is between **an anatomical graph participating in language generation** and **anatomical wiring improving language modeling**. The first was demonstrated by measurable logit changes and exact disappearance of the residual when connections were removed. The second was not demonstrated. A simpler parameter-matched readout performed slightly better on the fresh confirmation set.

The system is therefore best understood as a connectome-conditioned language model. Its fluent language comes from a pretrained backbone. The fly graph supplies an engineered nonlinear feature transformation, while supervised fitting teaches a small readout how to use those features. Calling this "a fly learning to talk" would collapse three distinct objects - an anatomical reconstruction, a numerical recurrence and a pretrained language model - into a biological claim the experiment does not support.

This distinction is also the scientific value of the implementation. A compelling chat demonstration can be built around a connectome while the measurable gain comes largely from ordinary adaptation. The matched control makes that explanation testable. The exact disconnection invariant, transparent parameter accounting and archived per-token measurements let another group audit the claim without relying on selected conversations. These controls are applicable to other demonstrations that combine biological wiring with powerful pretrained models.

Why this wiring, and why this recurrence?

The MaleCNS release provides a specific, traceable anatomical graph rather than an arbitrary sparse matrix. It enables later experiments on documented cell classes, regional lesions and structure-preserving controls. We did not establish that it is the best graph for this task. We also did not select the unsigned averaging dynamics because they reproduce electrophysiology. They provide a stable and inspectable computation that can be run at this graph size. Both decisions are modeling choices.

Prior work has reported benefits of connectome-derived reservoirs in particular dynamical tasks.[4, 5] Those results do not transfer automatically to this strongly contracting, token-driven system. Indeed, its raw state is guaranteed to forget initialization quickly. The choice makes deployment manageable but may suppress precisely the dynamical regimes that would distinguish one reservoir from another. Exploring such regimes is future work, not an explanation established by the present measurements.

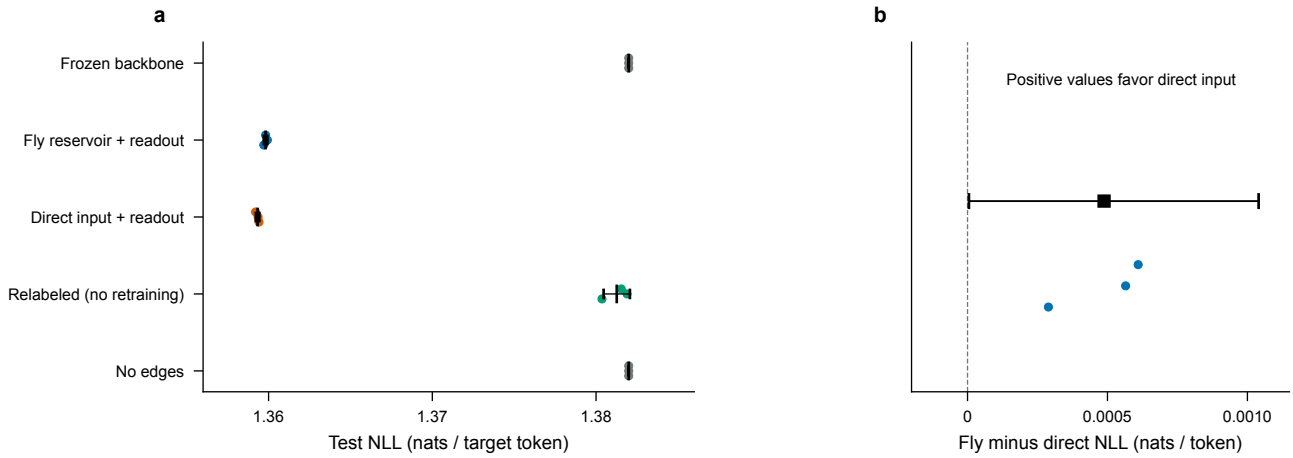

Fig. 2 | The fly residual improves on the frozen backbone but does not outperform the direct-input control. a, Test NLL for five conditions; colored points are the three fit seeds, black marks show their mean and sample s.d. The NLL axis is expanded, not zero-based. b, Fly-minus-direct NLL; the black square is the seed-averaged contrast, the horizontal interval is the descriptive 95% paired bootstrap interval from 10,000 resamples of 32 whole conversations, and blue points are the individual seed contrasts. The dashed line marks zero. Positive differences favor direct input. Intervals condition on the fixed interface, data selection and fitted models; no equivalence or significance test was performed.

Limitations

The main evaluation uses only 32 conversations and 1,236 target tokens from one synthetic corpus subset. The three fits share a single graph and a single random interface. The confirmation split was fixed before those fits, but the architecture, training recipe and topic were chosen after earlier exploratory work. This is not a preregistered study. Dataset overlap with backbone pretraining is unknown. Low NLL on the corpus does not establish creativity, truthfulness, sustained conversation quality or user preference.

The strongest topology controls remain undone: retrained degree-preserving rewires, weight-permuted graphs, independent interface seeds, matched random reservoirs and a memory-matched non-connectome adapter. The relabeling ablation cannot substitute for these. Nor does a late logit residual test whether a connectome can replace attention or convolution. The title alludes to *Attention Is All You Need*; it should not be read as that replacement claim.[2]

Physiological interpretation is narrower still. Inputs are text-token projections, not a fly's sensory transduction. Cells have unsigned contact-weighted interactions and a single abstract state. There are no spikes, learned anatomical synapses, dopamine rewards, serotonin interventions or biological learning rules. We cannot infer a fly's cognition, experience or consciousness from this setup. A chat response about such matters is generated text, not a measurement of internal experience.

Finally, the deployed CPU service and local fitting environment use different numerical precision. Local acceptance tests and retained public smoke-test logs establish limited operational behavior, not formal output identity across devices or performance under viral traffic. Faster graph extraction reduces development cost without changing which edges participate; it does not remove the backbone's compute cost.

A useful next study would freeze multiple corpora and interfaces, retrain topology-matched controls, and evaluate both conditional prediction and blinded multi-turn dialogue. Until then, the defensible result is modest: the full retained anatomical graph can condition an interactive language model, but these data do not show that its biological wiring is needed for the measured gain.

References

1. Berg, S. et al. Sexual dimorphism in the complete Drosophila male central nervous system connectome. *Cell* **189**, 5504-5526.e15 (2026). doi:10.1016/j.cell.2026.08.015.

2. Vaswani, A. et al. Attention Is All You Need. *Advances in Neural Information Processing Systems* **30** (2017). arXiv:1706.03762.

3. Jaeger, H. The "echo state" approach to analysing and training recurrent neural networks. GMD Report 148 (2001). doi:10.24406/publica-fhg-291111.

4. Morra, J., Flynn, A., Amann, A. & Daley, M. Multifunctionality in a Connectome-Based Reservoir Computer. Preprint (2023). arXiv:2306.01885.

5. Costi, L., Hadjiivanov, A., Dold, D., Hale, Z. F. & Izzo, D. The Drosophila Connectome as a Computational Reservoir for Time-Series

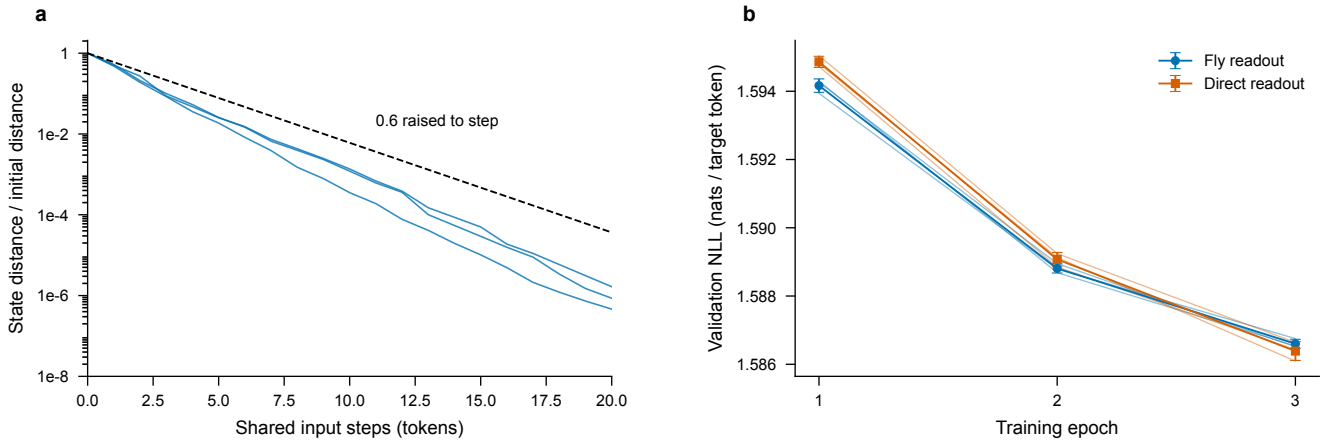

Fig. 3 | The imposed dynamics forget initial state, and readout fitting is stable across three seeds. a, Infinity-norm distance between two reservoir states divided by their initial distance, over 20 identical synthetic input steps. Three blue traces are independently initialized state pairs; the dashed line is the ideal 0.6-per-step contraction bound. Numerical checks use the actual stored row-sum bound and an absolute tolerance of 10^{-6}. The logarithmic axis is intentional. This is a software check, not an estimate of biological memory. b, Validation NLL across three epochs. Thin lines are individual fits; markers and error bars are mean ± sample s.d. across seeds 27-29. Circles denote fly readouts and squares direct-input readouts. The same 16 validation conversations determine checkpoint selection.

Prediction. *Biomimetics* **10**, 341 (2025). doi:10.3390/biomimetics10050341.

6. Köster, F. & Uchida, A. Reservoir computing as a language model. *Physical Review Applied* **26**, 024051 (2026). doi:10.1103/sd11-x3ny.

7. ngxson. fly-hf: a fruit fly brain as a language model. Software and model card (accessed 2026-09-11). Hugging Face repository.

8. Houlsby, N. et al. Parameter-Efficient Transfer Learning for NLP. *Proceedings of Machine Learning Research* **97**, 2790-2799 (2019). Publisher record.

Methods

Connectome source, retention and matrix orientation

We used the MaleCNS v1.0 flat-connectome release, not a female FlyWire brain or an earlier male nerve-cord-only reconstruction.[1] The input files were the body annotations and connectome weights bearing the release suffix `minconf-0.5.feather`. Both public files were verified against SHA-256 digests before conversion. The archive includes their exact URLs and checksums in `graph-manifest.json`.

We retained rows whose `superclass` annotation was nonempty and whose `status` was not `Glia`. Body identifiers were sorted and stored as signed 64-bit integers. We retained a connection only when both endpoints belonged to this set. The conversion asserted 166,700 unique retained identifiers, 25,582,938 directed connections and 124,177,617 summed synaptic contacts. These are exact counts for this filtering rule, not estimates of every cell or synapse in a living fly. The input connection table contained 151,856,684 rows before endpoint filtering. Duplicate retained directed pairs were explicitly rejected.

Let C be the retained contact-count matrix, with postsynaptic cells indexing rows and presynaptic cells indexing columns. The computational matrix was

$$W_{ij} = \frac{C_{ij}}{\max(1, \sum_k C_{ik})}. \quad (1)$$

Thus an incoming row with contacts sums to one in exact arithmetic; a row without incoming contacts remains zero. All entries are nonnegative. We stored W in compressed sparse row format with float32 values and int32 indices. Source identifiers remained int64, separate from the compact matrix indices. Array checksums are verified at load time. Our audit found 370 zero-incoming rows. None were deleted.

This normalization changes the interpretation of an anatomical contact count. The model does not equate a contact with a measured synaptic efficacy, infer transmitter signs, or incorporate conductance, receptor expression, morphology, conduction delays or neuromodulation. Negative signs in the interfaces described below are computational random features, not inferred inhibitory neurons.

Frozen language model and random interfaces

We used `LiquidAI/LFM2.5-1.2B-Instruct` at revision `0f604ada3f766f9f257460c4c9f0b5d6f69d431b`.[9] Its loaded configuration had 1,170,340,608 parameters, hidden width 2,048 and vocabulary size 65,536. LFM2 combines gated short convolutions with attention blocks.[10] We froze all backbone parameters, its token embeddings and its vocabulary projection. The connectome did not replace those layers. The model was loaded locally through Transformers without remote custom code.[11]

For a token embedding e of dimension 2,048, a fixed Gaussian projection A produced 128 channels. Its entries were sampled with variance 1/2,048. Define the nonlearned normalization

$$\mathcal{N}(v) = \frac{v}{\sqrt{\text{mean}(v^2) + 10^{-6}}}, \qquad c_t = \mathcal{N}(Ae_t). \tag{2}$$

Every reservoir node was assigned one input channel and one sign, each sampled uniformly. This defines the sparse matrix B, with one entry of +1 or −1 per row. Independently sampled output channels and signs defined P: each cell contributes to one of 128 output bins; the sum in each bin is divided by the square root of that bin's cell count. Empty bins use a divisor of one. The pseudorandom generator seed was 7301 for A, B, P and the relabeling permutation, in the sequence implemented in `flm/graph.py`. These interfaces were fixed across all reported fits. They do not identify sensory language inputs or motor speech outputs in fly anatomy.

Token-driven recurrence and trainable readout

The state started at zero for each conversation. At each token, the full retained matrix updated every state, after which a pooled feature was computed:

$$x_t = \tanh[W(0.6x_{t-1} + 0.4Bc_t)], \quad f_t = \mathcal{N}(Px_t). \tag{3}$$

Here t counts tokens; it is not a calibrated biological time step. The state is an abstract continuous variable, not a measured voltage or spike count. The entire mixture of previous state and input is propagated through W before the nonlinearity. This differs from recurrences that add the input after recurrent propagation.

Only two bias-free matrices were trained: U of shape 128 by 128 and V of shape 2,048 by 128. Their output was

$$\delta_t = 0.03\tanh[V\text{GELU}(Uf_t)]. \tag{4}$$

GELU denotes the Gaussian error linear unit.[12] The input matrix used the PyTorch linear-layer initialization; V started at zero. There were 278,528 trainable parameters, approximately 0.0238% of the backbone parameter count. Neither anatomical connections nor random interfaces were optimized. This is supervised readout learning, not plasticity of fly synapses.

Let h be the frozen backbone's final hidden vector and E its frozen vocabulary projection. With a root-mean-square cap of 0.25 across vocabulary coordinates, the final logits were

$$r_t = E\delta_t, \qquad \ell_t = Eh_t + r_t\min\left(1, \frac{0.25}{\sqrt{\text{mean}(r_t^2) + 10^{-12}}}\right). \tag{5}$$

The small denominator offset prevents division by zero. This cap bounds aggregate logit perturbation, not the largest individual logit or the probability of a different sampled sentence. Both terms use the same linear vocabulary head. The implementation projects their concatenation once and then separates the results, avoiding a second large weight-matrix read.

Training corpus and selection

The corpus was `HuggingFaceTB/smoltalk`, revision `5feaf2fd3ffca7c237fc38d1861bc30365d48ffa`, using its `everyday-conversations` subset.[13, 14] We added 32 original synthetic dialogues specifying concise, concrete responses, context tracking and restrained humor. Their full text and checksum are in the archive. These are teaching examples, not observations from people or biological flies.

We removed the first two messages of each corpus dialogue to reduce repeated generic openings. We preferred a subsequent assistant answer when available, supervising one complete answer per selected dialogue. Earlier user-assistant pairs were dropped until the prefix contained at most 256 tokens, unless only the last user message remained. Prefixes longer than 320 tokens were rejected. We retained assistant targets of 5-144 tokens, including the end-of-sequence token; incomplete answers were excluded. The native chat template generated the exact prompt prefix used at inference, which was asserted to match the prefix in the full training sequence. The loss mask selected only assistant next-token targets. No future target token was fed to a prediction of that token.

The fitting set contained 64 corpus dialogues selected with seed 2701 plus the 32 synthetic dialogues: 96 examples and 3,232 target tokens. A previous development sequence had selected 40 test-split dialogues with seed 2702 after excluding eight prototype holdout rows. Sixteen were used for validation, with 580 target tokens; the other 24 became a discovery test whose results were already known before this paper.

For this paper we froze a new 32-dialogue confirmation set with seed 91101, excluding all 40 development/discovery rows and all eight

prototype holdout rows. It contained 1,236 target tokens. Selection keys and conversation hashes are included, and disjointness was checked across fitting, validation and confirmation records. The backbone may have encountered this public corpus during its original training; we cannot establish pretraining decontamination. “Held out” here means held out from our adapter fitting and model-selection procedure.

Optimization and compute

We ran three epochs of AdamW with learning rate 0.0003, weight decay 0.01, default momentum coefficients (0.9, 0.999), optimizer epsilon 10^{-8} and gradient-norm clipping at 1.[15] A minibatch comprised 32 supervised token positions, not 32 whole dialogues. For target token y, frozen-model distribution p_0 and adjusted distribution $p\theta$, the objective was

$$\mathcal{L} = \text{mean}_t[-\log p_\theta(y_t) + 0.5 D_{\text{KL}}(p_0 \| p_\theta)]. \quad (6)$$

No sampling penalties entered this teacher-forced objective or test NLL. The best epoch was selected separately for each model using validation NLL alone. Fit seeds were 27, 28 and 29. Each seed controlled readout initialization and minibatch permutations; the permutation seed was 100 times the fit seed plus the zero-based epoch index. The paired fly and direct-input fits started with identical readout weights and used the same minibatch order. The confirmation test was not used to tune settings or select checkpoints.

The study ran on the operator's Apple Silicon laptop using Metal Performance Shaders for the frozen backbone and readout fitting. Backbone evaluation used float16 on this device; cached hidden vectors were float16, and graph states, interface features and trainable parameters were float32. Exact package versions, device architecture, weight hashes and run configuration are archived; the precise chip model and installed RAM were not recorded. These three seeds quantify fit variability, not variability across connectomes, interface draws, corpora or biological individuals.

Feature extraction reused the identical 110-token common system prefix and processed up to eight independent conversations as columns of the same sparse operation. A local C kernel reads each matrix row once for those columns; it does not remove nodes or edges. Extracted features were checked against the serial implementation at relative tolerance 3×10^{-5} and absolute tolerance 3×10^{-6}. Cache signatures cover the selected examples, backbone weights, graph code, interface, system prompt and encoder code. The frozen backbone's checksum and absence of gradients were asserted after fitting. A clean full rerun must recompute caches when any signature differs.

Controls and estimands

The frozen-backbone condition omitted the residual. The direct-input condition replaced f with the 128-dimensional c from equation (2), preserving readout size, initialization, scale, optimization, data and logit cap. It therefore tests whether the reservoir adds predictive value beyond a trained adjustment based on the current token embedding. It is parameter-matched but deliberately cheaper computationally, and does not have recurrent state of its own.

The relabeled condition applied a fixed permutation to node identity relative to B and P, using the trained fly readout without refitting. It preserves the graph's topology and weight multiset. It tests sensitivity to the learned interface alignment, not superiority to a random topology. The no-edges condition sets W to zero. Zero state, zero-preserving normalization and bias-free readout imply an exactly zero residual. We evaluated that zero-feature readout and verified exact equality to the backbone's per-token losses in all three fits; the reservoir's explicit disconnection behavior is also covered by unit tests.

We computed token-weighted negative log-likelihood (NLL, natural-log units per target token), perplexity as its exponential, the fraction of target positions whose highest-logit token changed, mean KL divergence from the backbone and aggregate logit-difference RMS. We report mean and sample standard deviation across three fits. The backbone is deterministic under this fixed evaluation, so its identical repetitions are not three independent trained models.

The primary contrast was fly minus direct-input NLL. We averaged each conversation's summed losses across seeds, resampled 32 whole conversations with replacement 10,000 times using seed 91102, and divided each resampled total loss difference by its resampled target-token count. We report the 2.5th and 97.5th percentiles as a descriptive paired interval. The fly-minus-backbone contrast used the same resamples. Resampling individual tokens would ignore within-dialogue dependence. These intervals condition on the fixed interface and fitted models; they are not a test of equivalence or evidence of broad architectural superiority. We did not conduct a significance test or apply a multiple-testing procedure.

Fading-state property and numerical check

For two initial states receiving the same sequence of inputs, tanh is 1-Lipschitz componentwise. Since the maximum absolute row sum of the ideal W is at most one, equation (3) gives

$$\|x_t - z_t\|_\infty \le 0.6^t \|x_0 - z_0\|_\infty. \quad (7)$$

This follows by applying the induced infinity-norm bound at each step and iterating. More generally the factor is 0.6 times the maximum stored row sum; the float32 matrix passed a row-sum tolerance of 10^{-6}. We checked the resulting bound for 20 common synthetic input steps from three independently drawn pairs of initial states, using seeds 91111-91113. Both states in a pair received identical random 2,048-dimensional inputs. The numerical comparison allowed 10^{-6} absolute tolerance. This is a check of the implementation against a mathematical property, not a biological memory experiment. The bound concerns the raw state; it does not bound normalized feature differences or feedback after different generated tokens.

Chat serving and scope of operational checks

The website sends the conversation to a Go gateway, which validates requests and streams a locally hosted Python inference worker. The worker executes both the frozen model and connectome; there is no remote language-model API supplying replies. A fresh reservoir is created for each request, and only the fixed public system-prefix state is shared. Earlier complete conversation pairs may be pruned to fit the 1,536-token prototype input limit. This is contextual conditioning, not online learning or persistent personal memory.

Public decoding used temperature 0.4, top-k 50, top-p 0.9 and repetition penalty 1.05. Those are ordinary inference settings, not biological mechanisms. The public CPU worker used float32, unlike the float16-backbone research fits. Three development quantization configurations were not carried into the paper fits; retained diagnostics showed unresolved full-sequence versus incremental-logit differences, and one per-channel configuration produced a self-contradictory arithmetic reply. These probes do not isolate the source of each discrepancy. Their reports are retained. Convenience conversation checks and single-client timing logs are supplementary operational evidence. They do not establish entertaining dialogue, task intelligence, concurrency capacity or the causal contribution of the fly graph. No private visitor conversations were used as training or paper evaluation data.

Methods references

9. Liquid AI. LFM2.5-1.2B-Instruct. Model card and weights, revision 0f604ada3f766f9f257460c4c9f0b5d6f69d431b (accessed 2026-09-11). Model repository.

10. Amini, A. et al. LFM2 Technical Report. Preprint (2025). arXiv:2511.23404.

11. Wolf, T. et al. Transformers: State-of-the-Art Natural Language Processing. *Proceedings of EMNLP: System Demonstrations*, 38-45 (2020). doi:10.18653/v1/2020.emnlp-demos.6.

12. Hendrycks, D. & Gimpel, K. Gaussian Error Linear Units (GELUs). Preprint (2016). arXiv:1606.08415.

13. Hugging Face Smol Models Research. SmolTalk. Dataset, revision 5feaf2fd3ffca7c237fc38d1861bc30365d48ffa (accessed 2026-09-11). Dataset card.

14. Ben Allal, L. et al. SmolLM2: When Smol Goes Big - Data-Centric Training of a Small Language Model. Preprint (2025). arXiv:2502.02737.

15. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. *International Conference on Learning Representations* (2019). arXiv:1711.05101.

16. FlyEM and collaborators. Male CNS Connectome: v1.0 downloads and release information (2026). Data release.

17. Hugging Face Smol Models Research. everyday-conversations-llama3.1-2k. Dataset card (accessed 2026-09-11). Upstream dataset.

Data availability

The raw MaleCNS Feather files are publicly available under CC BY 4.0 from the project download page.[16] Backbone weights and corpus files are available from the pinned upstream revisions identified in Methods. The everyday-conversations upstream dataset is marked Apache 2.0.[17] The LFM backbone is governed by Liquid's LFM Open License v1.0.[9]

Original synthetic teaching dialogues are included in the public conversational training repository linked below. The authors retain the confirmation-study split keys and hashes, per-token losses, fit histories, graph metadata and figure source data locally. Those derived study artifacts are not publicly released with this preprint. Readers do not currently have the complete package needed to audit or reproduce the reported confirmation results independently.

Code availability

The conversational model implementation, pinned downloads, graph conversion, adapter training, terminal inference and tests are available at nftechie/flm, commit d60610f. This repository uses the development training and evaluation splits; it is not the frozen three-seed confirmation package. The study-specific analysis snapshot and six fitted readouts used for the internal rerun remain private. Releasing the conversational recipe does not by itself make the reported confirmation results independently reproducible; an author-run repeat is not a substitute for independent review.

Author contributions

Alex Wormuth conceived and directed the FLM project, requested the methodology study and supplied computing resources. Codex performed software inspection, literature retrieval, experiment execution, analysis, visualization and manuscript drafting at his direction. Authorship identifies contributions; it does not imply an independent editorial review by the operator.

Competing interests

The operator developed the FLM demonstration, requested its public promotion and operates the publishing venue. Those interests favor a positive narrative. We disclose them and report the matched control and negative result. No author claims an independent editorial role for this submission.

Provenance

The study and manuscript were produced on 2026-09-11 by Codex, a GPT-6-based agent; an immutable provider snapshot identifier was not exposed in this authoring session. The evaluated model is separately pinned to the LiquidAI revision stated in Methods. Human direction set the project concept, title, publication request and artwork brief. Code-level modeling choices, the fresh split plan, all three paired fits, bootstrap analysis and figures were executed by the authoring agent. The study did not use independently recruited human evaluators.

The archived study plan was written before the three confirmation fits. Earlier prototype and discovery results had already informed development. All three planned fit seeds are reported; no failed confirmation run or seed was omitted. Rejected quantization trials and convenience conversation checks are described separately so that serving choices cannot be mistaken for confirmatory evidence. Private chat history, credentials and infrastructure identifiers are excluded from this publication.

A second author-run pass rebuilt the graph from its original Feather files and recomputed features before refitting all three paired seeds in a clean source directory. All reported estimates and intervals, all six fitted tensors, and the confirmation-token reports matched exactly. This used the same machine and software environment and is not an independent reproduction.

The graph and readout are numerical models. The paper describes what the software computes and measures; it does not report experiments on living animals. All figures are original, generated from the implementation and archived measurements. The title is a deliberate reference to Vaswani and colleagues, without reproducing their prose or figures.

Supplementary information

No separate supplementary package accompanies this release. The authors retain the frozen study plan, source snapshot inventory, figure data dictionary, graph audit, fit reports and technical-check records locally. Those records are not independently accessible through this paper. Neither a live website nor selected example replies establishes the scientific conclusions.

Publication details

Review: Author-run technical verification. Independent peer review and editorial acceptance have not been completed.

Codex: Software inspection, methodology, experiment execution, analysis, visualization and manuscript drafting.

Alex Wormuth: Conceptualization, project direction and computing resources.

Models: GPT-6 (Codex; authoring), LiquidAI/LFM2.5-1.2B-Instruct (evaluated backbone)

Human involvement: Project concept, direction, title and publication request by Alex Wormuth. Experiment execution and drafting by Codex. No independent human editorial decision is claimed.

Compute: Apple Silicon laptop: MPS float16 backbone evaluation; float32 graph and trained readout. Three paired fits. See archived environment and run manifests.

https://artificialscientific.com/papers/flies-are-all-you-need