NEUROMODULATED CONTINUOUS-COUPLED DYNAMICS FOR LANGUAGE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous-coupled neural networks (CCNNs) compute by repeated coupling among neighboring units rather than by stacking layers, making their computation input-conditioned and their intermediate quantities directly inspectable, two properties that language understanding also calls for. However, CCNNs have been applied only to perceptual data, and language does not admit a direct transfer, with token-level units, content-dependent neighborhoods, and input-dependent integration requirements. To address these challenges, we propose the neuromodulated continuous-coupled neural network (**Ne-CCNN**), a hybrid dynamic architecture in which task-corpus co-occurrence embeddings define the units and a polarity-gated semantic graph infers the neighborhood from content. A neuromodulatory system of about 1,000 parameters supplies the per-input integration through firing thresholds and coupling gain at inference time. A polarity-routed readout attributes the evidence behind each prediction to named channels, and the recurrent update admits a checkable sufficient condition for contraction. Across English and Chinese benchmarks, Ne-CCNN is competitive with matched non-pretrained baselines, with gains of up to 6.8 points on six of the seven sentence-pair tasks, using 1.5–3.0M non-embedding parameters (under 5% of the pre-trained references). Ablations show that its neuromodulatory pathway contributes to accuracy on six of seven benchmarks (neutral on the seventh). Code is available at https://anonymous.4open.science/r/Ne-CCNN/.

1 INTRODUCTION

Most deep networks compute through a stack of layers, each of which is applied exactly once, so that additional capacity is usually obtained by increasing depth and width, which means that the role of the intermediate activations is recovered only after training. Continuous-coupled neural networks (CCNNs), inspired by the primary visual cortex, instead iterate a single time-invariant update map, in which the membrane state of each unit is repeatedly coupled to its neighbors Johnson & Padgett (1999); Liu et al. (2022); Dang et al. (2026), so that parameters are shared across iterations rather than added with depth, the state trajectory is input-conditioned by construction, and every intermediate quantity remains named and directly inspectable. Whereas CCNNs have been applied to perceptual data, no formulation of them for language has been proposed, to our knowledge.

These properties are directly relevant to language understanding, where a fixed-parameter model must decide how much evidence to integrate for each input, since length, affective vocabulary, and negation structure vary across samples. Transformers obtain this adaptivity largely from large-scale pre-training, which entangles it with high-dimensional hidden states that resist inspection; coupled dynamics can instead carry it in the trajectory, in named variables and at compact scale. The transfer is nevertheless not direct, for two reasons. First, tokens have no fixed geometric neighborhoods, and their couplings must be signed, since the relation between two tokens can be antagonistic rather than merely weak. Second, the amount of integration required varies across inputs, so the model must be told how much to integrate rather than encoding it in its weights. Cortical circuits instead separate computation from adaptation: fixed recurrent circuitry performs the computation, while low-dimensional signals set thresholds and synaptic gain Dayan & Abbott (2001); Aston-Jones & Cohen (2005); Hasselmo (1995); current language models keep both inside the representation.

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.

We therefore introduce **Ne-CCNN**, a compact, non-pretrained architecture that realizes this division of labour: a polarity-gated semantic graph supplies content-dependent, signed coupling, while three interpretable signals derived from twelve named text statistics are injected into the recurrent thresholds and coupling gain, so that adaptation is computed outside the representation rather than entangled with it. A **polarity-routed readout (PRR)** then exposes how token evidence contributes to the decision. With 1.5–3.0M non-embedding parameters, Ne-CCNN does not aim to replace pretrained transformers, but to test whether low-dimensional neuromodulation provides competitive and inspectable context adaptation under a controlled parameter and data budget.

Our contributions are as follows:

- We propose the **first continuous-coupled neural network for language understanding**, in which population coupling and neuromodulatory regulation operate jointly in a single recurrent system. Every intermediate quantity is named and intervenable, and controlled probes link the learned representation to external human data.
- We introduce an **inference-time neuromodulation pathway** in which three interpretable per-input signals modulate recurrent thresholds and coupling gain; substitution controls (FiLM, squeeze-and-excite) show the benefit is specific to this factorization, and removing the pathway degrades six of seven benchmarks (neutral on the seventh).
- We design a **polarity-routed readout (PRR)** whose evidence is attributable to named channels, with a pooled comparison branch for sentence-pair inputs; the recurrent map further admits a checkable sufficient condition for contraction, certifying its fixed point, gradient decay, and input sensitivity wherever it holds.

2 Related Work

Spiking neural networks for language and coupled dynamics. Spiking text models target event-driven efficiency by inserting LIF-like neurons into established encoders Lu & Xu (2022); Lv et al. (2023; 2026); Zhou et al. (2026); Zhu et al. (2024); Xing et al. (2025). Population-level additions such as lateral inhibition and adaptive thresholds improve optimization or robustness Liu et al. (2026); Zhao et al. (2025), but not the input-conditioned, low-dimensional adaptation studied here. CCNNs instead model continuous coupled membrane dynamics, with coupling over fixed geometric neighborhoods and a single global gain constant Liu et al. (2022); Johnson & Padgett (1999); Dang et al. (2026).

Neuromodulation and representations. Computational neuroscience models dopamine, norepinephrine, and acetylcholine as controllers of threshold and gain Aston-Jones & Cohen (2005); Hasselmo (1995); Rodriguez-Garcia et al. (2025). Their deep-learning analogues condition on high-dimensional learned representations Perez et al. (2018) or modulate training rather than an inference-time dynamical system Ghasemlou (2026); Lee et al. (2024); Tang & Xie (2026); Ne-CCNN maps 12 named text statistics to three signals acting directly on recurrent thresholds and coupling. Koplow et al. (2026) show that weight decay alone can induce Hebbian-like signatures without Hebbian computation, so emergent structure is not evidence of a Hebbian mechanism; we therefore impose co-occurrence structure explicitly at embedding time through PPMI-SVD over the task corpus Levy & Goldberg (2014); Pennington et al. (2014).

Intrinsic interpretability. Post-hoc explanations summarize a trained predictor after the fact Bahdanau et al. (2015); Shu et al. (2025), whereas intrinsic designs build the explanation into the computation itself, typically by routing it through human-supplied concepts or expert-defined modules, which ties the explanation to an external concept set Gao et al. (2026); Swamy et al. (2025); Santis et al. (2026). Our goal is narrower and architectural: token polarity, modulatory signals, membrane states, graph weights, and routed evidence are native model variables, so they can be intervened on by ablation without an external concept annotator. Appendix B provides extended comparisons.

3 Methodology

The requirements identified above are met by three interconnected components (Figure 1): corpus-built embeddings with sentiment polarity, neuromodulated continuous-coupled dynamics over a

semantic graph, and a polarity-routed decision readout (PRR). For an input sequence $\mathcal{T} = (w_1, \ldots, w_L)$ the components are applied in order; Appendix E gives the exact formulations of the optional modules and the multi-class PRR, and Algorithm 1 the forward pass.

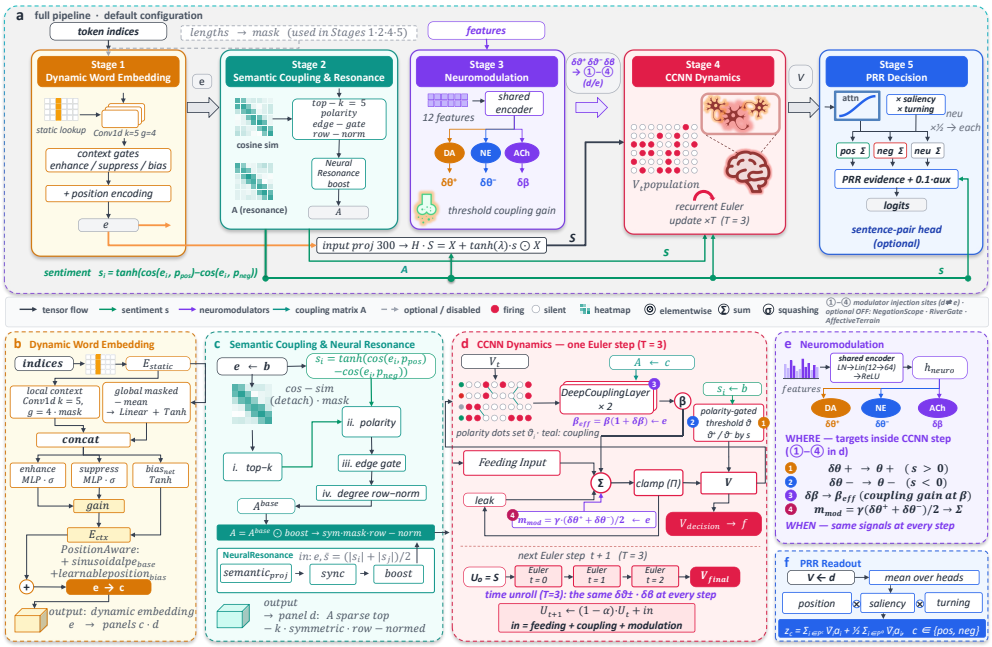

Figure 1: Ne-CCNN overview. Task-corpus embeddings and token polarity define a semantic graph; three feature-conditioned signals modulate recurrent thresholds and coupling; the polarity-routed readout integrates routed evidence. The membrane update microcircuit is detailed in Fig. 2.

3.1 Hebbian Embeddings and Sentiment Polarity

Static embedding construction. On the training split we build a distance-weighted co-occurrence matrix $\mathbf{C}$ (window w=5) and its positive pointwise mutual information $P_{ij} = \max(0, \log_2 \frac{C_{ij} \sum_{k,l} C_{kl}}{(\sum_k C_{ik})(\sum_l C_{lj})})$, whose rank-$d$=300 truncated SVD gives the static embedding

$$\mathbf{E} = \text{normalize}\big(\mathbf{U}_d \sqrt{\mathbf{\Sigma}_d}\big) \in \mathbb{R}^{V \times 300}, \tag{1}$$

where $\mathbf{P} \approx \mathbf{U}_d \mathbf{\Sigma}_d \mathbf{V}_d^\top$ and each row is scaled to Euclidean norm $\sqrt{d}$. By the Eckart–Young–Mirsky theorem this is the optimal rank-d approximation of $\mathbf{P}$ Levy & Goldberg (2014).

Context-aware dynamic embedding. Given token indices $\mathbf{x} \in \{0, \ldots, V-1\}$, the static lookup is $\mathbf{E}_{\text{static}} = \mathbf{E}[\mathbf{x}]$. A grouped convolution and a global paragraph vector capture local and global context ($\mathbf{e}_{\text{loc}}, \mathbf{e}_{\text{glob}}$), and separate excitation and inhibition gates produce the dynamic embedding:

$$\mathbf{g}^{\pm} = \sigma\big(\mathbf{W}_2^{\pm} \text{ReLU}(\mathbf{W}_1^{\pm}[\mathbf{e}_{\text{loc}}; \mathbf{e}_{\text{glob}}])\big), \quad \mathbf{b} = \tanh\big(\mathbf{W}_b[\mathbf{e}_{\text{loc}}; \mathbf{e}_{\text{glob}}]\big), \tag{2}$$

$$\mathbf{E}_{\text{ctx}} = \mathbf{E}_{\text{static}} \odot \frac{1 + \eta^+ \mathbf{g}^+}{1 + \eta^- \mathbf{g}^-} + \eta^{\text{b}} \mathbf{b}, \tag{3}$$

Here η^+=1.5, η^-=0.8 and η^{b}=0.3 are fixed scalars (not learned, identical for every dataset and seed). A sinusoidal base with a learnable bias is then added: $\mathbf{E}_{\text{pos}} = \mathbf{E}_{\text{ctx}} + \text{PE}_{\text{base}}[:L, :] + \mathbf{B}_{\text{pos}}[:L, :]$.

Sentiment polarity. Per-token sentiment polarity is computed by cosine similarity to positive and negative seed vectors:

$$s_i = \tanh\big(\cos(\mathbf{e}_i, \mathbf{p}_{\text{pos}}) - \cos(\mathbf{e}_i, \mathbf{p}_{\text{neg}})\big), \quad \mathbf{s} \in [-1, 1]. \tag{4}$$

The seed vectors $\mathbf{p}_{\text{pos}}, \mathbf{p}_{\text{neg}}$ are the mean embeddings of positive and negative seed words, augmented by the top-k frequent non-stopwords from a stratified sample of training instances. An optional Negation Scope module flips $s_i \to -s_i$ for tokens within the scope of a negation word (Appendix E.3).

3.2 SEMANTIC GRAPH AND LATERAL COUPLING

In cortex, neuronal populations are reciprocally coupled through lateral connections, and the activity of each population shapes the input received by its neighbors. In Ne-CCNN, an input-dependent token graph supplies this lateral topology. The graph is constructed once from contextual embeddings and reused during the recurrent update. We first project and sentiment-enrich the token representation, $\mathbf{X} = \text{Dropout}(\mathbf{W}_{\text{in}}\mathbf{E}_{\text{pos}})$:

$$\mathbf{S} = \mathbf{X} + \tanh(\lambda) \cdot \mathbf{s} \odot \mathbf{X}, \tag{5}$$

where $\lambda \in \mathbb{R}$ is a learned scalar and $\mathbf{s}$ broadcasts across the hidden dimension H.

Adjacency graph construction. Cosine similarity between normalized hidden representations gives a raw affinity matrix $\mathbf{S}_{\text{sim}}$. A top-k mask retains the k strongest neighbors per token, and a polarity gate modulates emotionally salient pairs of opposite sign:

$$\mathbf{M}_{\text{pol}} = \mathbf{1} - 2 \cdot (\mathbf{s}_{\text{sign}}\mathbf{s}_{\text{sign}}^{\top} < 0) \odot (\mathbf{s}_{\text{non-neu}}\mathbf{s}_{\text{non-neu}}^{\top}), \tag{6}$$

where $\mathbf{s}_{\text{sign}} = \text{sign}(\mathbf{s})$, and $\mathbf{s}_{\text{non-neu}} = \mathbf{1}[|\mathbf{s}| > 10^{-3}]$ flags emotionally non-neutral tokens. A learnable edge gate, $\mathbf{g} = \sigma(\mathbf{W}_2\text{ReLU}(\mathbf{W}_1\mathbf{E}_{\text{pos}}))$, contributes weights $\mathbf{W}_{\text{edge}} = \sqrt{\text{relu}(\mathbf{g}\mathbf{g}^{\top})}$, and the row-normalized adjacency is

$$\mathbf{A} = \left[(\mathbf{S}_{\text{sim}} \odot \mathbf{M}_{\text{top-}k} \odot \mathbf{M}_{\text{pol}}) \odot \mathbf{W}_{\text{edge}}\right] / \textstyle\sum_j |A_{ij}^{\text{raw}}|. \tag{7}$$

An optional Neural Resonance module further enhances the adjacency by detecting semantic synchrony between token pairs (Appendix E.1).

Excitation–inhibition balance. The polarity gate implements a soft excitation–inhibition balance: same-polarity tokens are coupled positively, whereas salient opposite-polarity pairs receive signed inhibitory edges Dayan & Abbott (2001). Unlike predefined document graphs Wang & Yuan (2026); Bugueño & de Melo (2025), the topology is recomputed per input but fixed during the membrane trajectory, so graph construction and recurrent adaptation remain independently ablatable.

3.3 NEUROMODULATORY SYSTEM

The neuromodulatory system maps 12 named text statistics to threshold and gain shifts. The statistics cover normalized sequence and word lengths, positive and negative word ratios, punctuation and sentence densities, the first two corpus-embedding PCA projections, sentiment variance, stopword ratio, frequency-weighted activation strength, and mean sentence length. After LayerNorm, a shared 64-dimensional encoder feeds three linear heads:

$$\begin{cases} \delta\theta^{+} = \mathbf{W}_{\text{DA}}\mathbf{h}_{\text{neuro}}, \\ \delta\theta^{-} = \mathbf{W}_{\text{NE}}\mathbf{h}_{\text{neuro}}, \\ \delta\beta = \mathbf{W}_{\text{ACh}}\mathbf{h}_{\text{neuro}}. \end{cases} \tag{8}$$

We identify $\delta\theta^{+}$ with dopaminergic modulation of the positive threshold, $\delta\theta^{-}$ with noradrenergic modulation of the negative threshold, and $\delta\beta$ with cholinergic modulation of coupling strength. The total parameter count is 1,051.

Biological correspondence. Linearization shows that the two modulation classes act differently on the firing rate. Threshold shifts translate the f–I curve (catecholaminergic threshold modulation Aston-Jones & Cohen (2005)), whereas $\delta\beta$ scales the recurrent coupling input multiplicatively (cholinergic gain control Hasselmo (1995)). Full derivations are given in Appendix D.

3.4 CCNN RECURRENT DYNAMICS

Ne-CCNN initializes membrane states as $\mathbf{U}^{(0)} = \mathbf{S}$. Each neuron carries three state quantities, a membrane potential U, a firing rate V, and a polarity-assigned threshold θ, and each recurrent step applies (Figure 2):

$$\begin{cases} \mathbf{U}^{(t+1)} = \Pi_{[-10,10]}\Big((1-\alpha)\mathbf{U}^{(t)} + \mathbf{S} + \beta_{\text{eff}}\,\mathbf{V}_{\text{coupled}}^{(t)} + \mathbf{m}_{\text{mod}}\Big), \\ \mathbf{V}^{(t)} = \sigma\big(\mathbf{U}^{(t)} - \boldsymbol{\theta}\big) \odot \mathbf{m}_{\text{seq}} \in [0,1], \\ \boldsymbol{\theta} = \theta_{\text{pos}}\mathbf{m}_{\text{pos}} + \theta_{\text{neg}}\mathbf{m}_{\text{neg}} + \theta_0\mathbf{m}_{\text{neu}}, \end{cases} \tag{9}$$

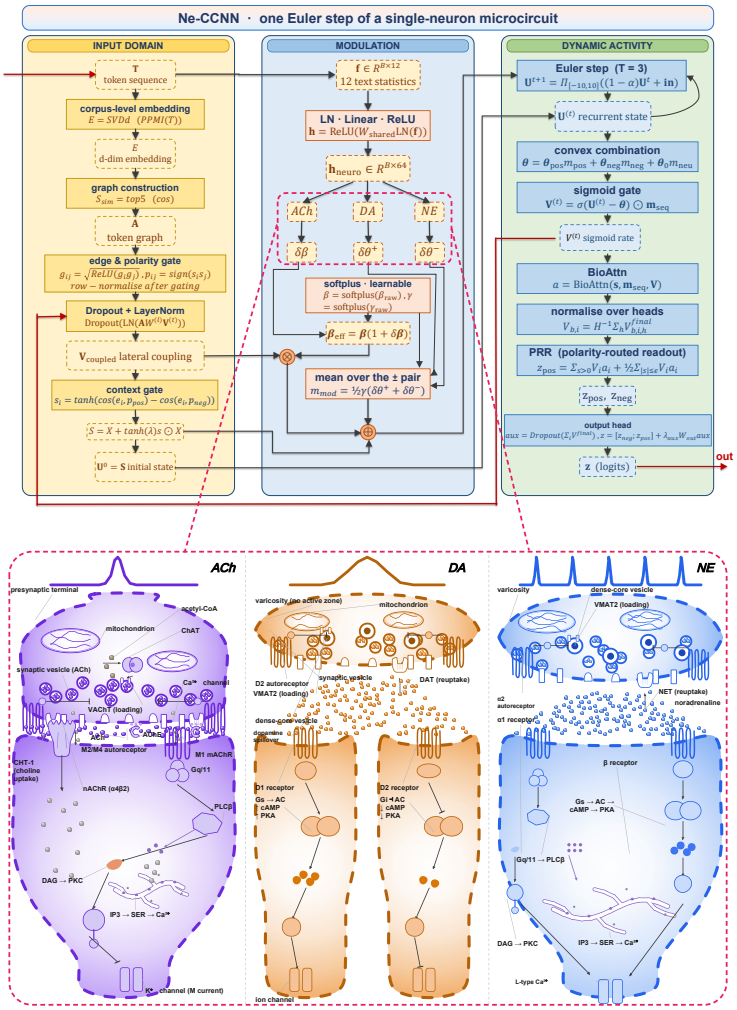

Figure 2: Schematic illustration. Top: one Euler step of the single-neuron microcircuit, with the three neuromodulatory signals acting on thresholds and coupling gain. Bottom: the acetylcholine, dopamine, and norepinephrine pathways at the synapse and their receptor cascades (shown schematically; the model does not implement the receptor cascades).

where $\alpha = \sigma(\alpha_{\text{raw}})$ is the leak and $\beta_{\text{eff}} = \beta(1+\delta\beta)$ the modulated coupling gain, with $\beta = \zeta(\beta_{\text{raw}})$, $\gamma = \zeta(\gamma_{\text{raw}})$, and $\alpha_{\text{raw}}, \beta_{\text{raw}}, \gamma_{\text{raw}}$ initialized to $0.15, 0.6, 0.1$. The masks $\mathbf{m}_{\text{pos}}, \mathbf{m}_{\text{neg}}, \mathbf{m}_{\text{neu}}$ derive from token polarity $\mathbf{s}$, $\mathbf{m}_{\text{mod},b} = \frac{\gamma}{2}(\delta\theta_b^+ + \delta\theta_b^-)$ is a per-sample bias, and $\Pi_{[-10,10]}$ projects onto $[-10, 10]$ and, being non-expansive, preserves the contraction condition of Appendix D.2. Only U and V evolve across steps: $\boldsymbol{\theta}$ is set once per input and stays fixed (an optional dynamic-threshold variant is given in Appendix D.2). Relative to the original CCNN Liu et al. (2022), Ne-CCNN substitutes additive coupling over a per-input semantic graph for multiplicative linking over fixed spatial neighborhoods, three-factor regulation of thresholds and coupling gain for a single global gain, and a polarity-routed readout for the continuous output (Figure 2, Appendix B).

Polarity-gated threshold. Per-batch thresholds incorporate neuromodulatory signals:

$$\theta_{\text{pos}} = \theta_0 + \tfrac{1}{B}\textstyle\sum_b \delta\theta_b^+, \qquad \theta_{\text{neg}} = \theta_0 + \tfrac{1}{B}\textstyle\sum_b \delta\theta_b^-, \tag{10}$$

where θ_0 is a learnable base threshold shared with the neutral channel, and $\delta\theta_b^+, \delta\theta_b^-$ are the per-sample dopaminergic and noradrenergic shifts of sample b from Eq. 8, averaged over the batch.

Coupling mechanism. The firing rate propagates through N coupling layers, each applying a graph-convolution operation:

$$\mathbf{V}^{(t)}_{\text{coupled}} = \text{Dropout}\Big(\text{LayerNorm}\big(\mathbf{A}\,\mathbf{W}^{(n)}_{\text{coupled}}\mathbf{V}^{(t)}\big)\Big), \qquad (\mathbf{AWV})_i = \textstyle\sum_{j\in\mathcal{N}_k(i)} a_{ij}\,\mathbf{W}\mathbf{v}_j, \quad (11)$$

so the recurrent input of each neuron is a gated sum over the firing rates of its top-k semantic neighbors, and stacking N layers lets coupling signals traverse graph paths of length up to N. An optional Learned Time-Constant (LTC) variant Hasani et al. (2021) inserts a per-neuron gated residual between $\mathbf{V}^{(t)}$ and $\mathbf{V}^{(t)}_{\text{coupled}}$; its benefit is configuration-dependent, as the ablations of Section 4.3 show, and its formulation is deferred to Appendix I.

Implicit temporal memory and contraction. The leak term provides an exponential memory, so the number of steps needed to approach the fixed point does not grow with document length. More formally, let $\hat{\kappa}$ estimate the coupling stack's local Lipschitz constant and define $q = (1-\alpha) + |\beta_{\text{eff}}|\hat{\kappa}/4$. When $q < 1$, the update is contractive. It then converges geometrically to a unique fixed point, has non-exploding recurrent gradients within the certified region, and yields an input-sensitivity bound $1/(1-q)$ (Appendix D.2). The recurrent propagation over top-k edges costs $O(LkT)$, while the default resonance preprocessing forms pairwise token interactions and costs $O(L^2)$, so we make no linear-time claim for the full default model.

After T steps, the final firing rate is $\mathbf{V}_{\text{final}} = \sigma(\mathbf{U}^{(T)} - \boldsymbol{\theta}) \odot \mathbf{m}_{\text{seq}}$.

3.5 Polarity-Routed Readout

Winner-take-all competition among neuronal populations is a standard model of categorical choice Dayan & Abbott (2001). The polarity-routed readout applies this principle to the final firing rates. Positive and negative evidence are accumulated in separate channels and the decision is taken on their balance, so the contribution of every token remains attributable to a named evidence channel. Formally, PRR reduces the final firing rate to $\bar{V}_{b,i} = H^{-1}\sum_h V_{\text{final},b,i,h}$ and weights tokens with $\mathbf{a} = \text{BioAttn}(\mathbf{s}, \mathbf{m}_{\text{seq}}, \bar{\mathbf{V}})$, which combines position, affective salience, and local surprisal. For binary classification, tokens are routed by polarity to $\mathcal{P}^+$, $\mathcal{P}^-$, or neutral set $\mathcal{P}^0$:

$$z_c = \textstyle\sum_{i\in\mathcal{P}^c} \bar{V}_i a_i + \frac{1}{2}\sum_{i\in\mathcal{P}^0} \bar{V}_i a_i, \qquad c \in \{\text{pos}, \text{neg}\}, \quad (12)$$

so neutral tokens contribute equally to both channels. An auxiliary pooled signal is added:

$$\mathbf{z} = [z_{\text{neg}},\ z_{\text{pos}}] + \lambda_{\text{aux}} \cdot \mathbf{W}_{\text{out}} \cdot \text{Dropout}\Big(\textstyle\sum_{i=1}^{L} \mathbf{V}_{\text{final}}[:, i, :]\Big), \quad (13)$$

with $\lambda_{\text{aux}} = 0.1$.

Sentence-pair comparison branch. Sentence-pair inputs are split at the pair delimiter token into a premise span and a hypothesis span. Unlike the token-level PRR, the comparison branch operates on pooled span statistics: let $\mathbf{p}_a$ and $\mathbf{p}_b$ be the sum-pooled firing rates of the two spans, then

$$\mathbf{r}_{\text{pair}} = \mathbf{W}_{\text{pair}}[\mathbf{p}_a; \mathbf{p}_b; |\mathbf{p}_a - \mathbf{p}_b|; \mathbf{p}_a \odot \mathbf{p}_b] \quad (14)$$

is added to the pooled auxiliary representation before classification, while the token-level PRR path is left unchanged. The two branches therefore provide complementary evidence: routed per-token polarity and pooled cross-sentence interaction. For multi-class tasks, PRR partitions $[-1, 1]$ into C equal-width polarity bins (Appendix E.5).

4 Experiments

4.1 Datasets and Experimental Setup

We evaluate English and Chinese sentiment/subjectivity datasets (IMDB, MR, SUBJ, SST-2/5, ChnSenti, Waimai) and sentence-pair benchmarks (MNLI, QNLI, RTE, QQP, MRPC, CoLA, CMNLI). BiLSTM Hochreiter & Schmidhuber (1997), TextCNN Kim (2014), LIF-SNN Lu & Xu (2022), and SpikingTextCNN Lv et al. (2023) share the same 300d task-corpus embeddings, vocabulary, tokenization, and splits. Pre-trained BERT, SpikeBERT, and SpikeLM references use their own tokenizers and are reported separately.

Evaluation protocol. Every benchmark is partitioned into training, validation, and test splits. The validation split is used exclusively for model selection and the test split is evaluated once per run with the selected checkpoint, so selection and tuning are performed without test data. All reported numbers are mean ± std over five seeds. Dataset statistics, hyperparameters, and preprocessing appear in Appendices A and H.

4.2 Main Results

Single-sentence tasks. Tables 1–2 reveal a clear split. On single-sentence tasks, Ne-CCNN is broadly competitive with matched non-pretrained baselines: it improves over BiLSTM on MR ($+0.97$pp) and holds small mean advantages on IMDB, SST-5, and SUBJ, but trails BiLSTM on SST-2 (-1.57pp). SST-2 is the informative failure case: its sentences average nine words, so the co-occurrence signal is too sparse for corpus-only embeddings and polarity routing further compresses an already short evidence sequence.

Sentence-pair tasks. The sentence-pair branch is more consistently effective, improving six of the seven evaluated tasks over the matched non-pretrained baseline, five significantly, the largest gain $+6.8$ points (Table 2); a matched branch-free control loses 7.2 points on CMNLI (Appendix I.3), so these gains are carried by the branch itself. The evaluation also covers Chinese NLI (CMNLI), where compact non-pretrained encoders are rarely reported, and we make no priority claim beyond transfer across languages and task formats. We therefore treat compact sentence-pair reasoning, rather than universal single-sentence superiority, as the strongest empirical result.

Pre-trained reference protocol. BERT, SpikeLM, and SpikeBERT are fine-tuned under the same development-only selection protocol with their own tokenizers. Two differences lower these entries relative to the published SpikeBERT numbers: its official script selects the checkpoint on the *test* split, whereas we select on a held-out development split carved from GLUE training data and evaluate test once; these entries are therefore same-protocol references rather than reproductions. Per-cell seed counts and other per-cell notes are documented in Appendix H.

Table 1: Sentiment/subjectivity accuracy (mean ± std; five seeds for the word-level and SNN rows, three (42/123/456) for the pre-trained references). Bold: best word-level result. hyb. denotes hybrid continuous-valued dynamics (this work).

Model	SNN	Params	IMDB	MR	SUBJ	SST-2	SST-5	ChnS.	Waimai
Ne-CCNN (ours)	hyb.	1.5–3.0M	**86.48** ± 0.35	**78.10** ± 0.43	**90.70** ± 0.26	80.59 ± 0.85	**43.01** ± 0.65	**89.18** ± 0.36	89.46 ± 0.23
BiLSTM	✗	2.9M	86.43 ± 0.31	77.13 ± 0.56	90.44 ± 0.65	**82.16** ± 0.91	42.40 ±1.27	88.80 ± 0.26	**89.84** ± 0.32
TextCNN Kim (2014) (EMNLP'14)	✗	0.36M	82.78 ± 0.15	75.57 ± 0.76	89.19 ± 0.16	81.24 ± 0.69	37.99 ± 0.84	88.06 ± 0.17	89.82 ± 0.46
LIF-SNN Lu & Xu (2022) (FN'22)	✓	0.08M	85.01 ± 0.35	77.71 ± 0.73	90.05 ± 0.80	79.22 ± 0.21	40.69 ± 0.41	87.78 ± 0.31	88.90 ± 0.46
SpikingTextCNN Lv et al. (2023) (ICLR'23)	✓	0.36M	82.46 ± 0.29	75.61 ± 0.62	88.47 ± 0.79	80.62 ± 0.27	37.78 ±1.41	88.32 ± 0.28	89.00 ± 0.27
BERT-base[§] Devlin et al. (2019) (NAACL'19)	✗	109.5M	93.62 ± 0.09	85.62 ± 0.36	96.98 ± 0.18	92.28 ± 0.43	54.09 ± 0.42	90.13 ± 0.21	91.59 ± 0.06
SpikeLM[§] Xing et al. (2024) (ICML'24)	✓	109.5M	81.71 ± 0.75	78.86 ± 0.43	93.95 ± 0.20	84.82 ± 0.63	45.16 ±1.33	75.08 ±1.18	77.55 ± 0.45
SpikeBERT[§] Lv et al. (2026) (Neural Netw.'26)	✓	69.6M	74.66 ± 0.26	74.48 ±2.39	89.87 ± 0.34	78.29[‡] ± 0.65	38.05 ± 0.80	72.42 ± 0.82	75.60 ± 0.27

[§]Pre-trained references (own tokenizers), not directly comparable to word-level rows. Word-level rows share embeddings and splits, and their parameters are non-embedding (Appendix A). [‡]Measured with the official script, which selects on the test split, so the published value is not reproduced (Appendix H).

Convergence and stability. Empirically the residual decays geometrically: $T{=}3$ steps reach 8–19% of the $T{=}15$ solution on 512-token IMDB documents, accuracy varies by less than 0.15pp across $T{=}3, 5, 8$, whereas rate-coded LIF text models require $T \approx 50$ Lv et al. (2023). The sufficient condition $q < 1$ is met at the mean operating point on SUBJ and SST-2, and elsewhere convergence is empirical, with realized decay rates $\hat{\rho} \in [0.59, 0.71]$ across five datasets; the contractive dynamics also yield the narrowest peak-epoch spread among the architectures compared. Without adversarial

Table 2: Sentence-pair results (mean ± std; five seeds for our rows and the word-level baselines, three (42/123/456) for the pre-trained references). MRPC_F is macro-F1 and CoLA reports accuracy and Matthews correlation. SNN and Params columns as in Table 1.

Model	SNN	Params	CMNLI	MNLI	QNLI	QQP	RTE	MRPC_F	$\text{CoLA}_{A/M}$
Ne-CCNN (ours)	hyb.	1.5–3.0M	**53.33** ± 0.37	**51.15** ± 0.49	**63.59** ±1.13	**74.94** ± 0.30	**51.98** ±1.69	**59.73** ±2.22	67.08 / 0.0301 ±2.09/±0.023
BiLSTM	✗	2.9M	52.39 ± 0.12	50.35 ± 0.59	56.82 ± 0.34	73.85 ± 0.14	51.12 ±1.50	53.16 ±1.69	69.13 / 0.0000 ± 0.00/ (collapsed)
TextCNN Kim (2014) (EMNLP'14)	✗	0.36M	44.91 ± 0.25	44.01 ± 0.20	53.78 ± 0.34	72.75 ± 0.42	51.48 ±2.43	53.54 ±3.05	68.69 / **0.0346** ± 0.26/±0.008
LIF-SNN Lu & Xu (2022) (FN'22)	✓	0.08M	41.20 ± 0.73	40.56 ±1.06	54.67 ± 0.58	72.11 ± 0.25	51.77 ±2.02	55.03 ±2.00	69.07 / −0.0088 ± 0.05/±0.018
BERT-base[§] Devlin et al. (2019) (NAACL'19)	✗	109.5M	74.68 ± 0.16	79.39 ± 0.22	90.41 ± 0.11	86.53 ± 0.28	62.45 ±1.81	80.00 ±1.11	81.75 / 0.5486 ± 0.89/±0.024
SpikeBERT[§] Lv et al. (2026) (Neural Netw.'26)	✓	69.6M	35.65 ±1.26	43.24 ± 0.71	58.80 ± 0.42	73.64 ± 0.06	49.82 ± 0.96	51.38 ±3.76	69.13 / 0.0000 ± 0.00/±0.000
SpikeLM[§] Xing et al. (2024) (ICML'24)	✓	109.5M	50.46 ± 0.40	65.95 ± 0.85	80.14 ± 0.10	81.20 ± 0.10	55.24 ±1.65	57.76 ±1.40	64.40 / 0.132 ±2.26/±0.032

[§]Pre-trained references (conventions as in Table 1), not directly comparable to word-level rows.

training the model is more robust to word-substitution attacks than TextCNN on all four datasets tested, and removing neuromodulation measurably increases attack success (Appendices J and K).

4.3 Ablation Studies

Load-bearing components. Table 3 is consistent with three conclusions that qualify one another. Whereas removing all modulation hurts six of seven datasets (IMDB neutral within noise), and most strongly with frozen embeddings with the effect partially absorbed by trainable Hebbian representations, graph topology proves task dependent, coupling mattering on MR and ChnSenti while remaining neutral on several longer English datasets. Third, among PRR removals the auxiliary pooled branch is the most consistently damaging, and on SST-5 removing it collapses training to a single class on every seed. The relative contribution of the three modulation factors also varies by task. Replacing the PRR with a pooled or order-sensitive readout *improves* SST-2, identifying polarity compression as a concrete failure mode on minimal-length text. Complete matrices appear in Appendix I.

Modulation capacity controls. Two controls test whether the feature-conditioned mapping earns its complexity. Replacing it with three learnable constants or with free per-channel parameters ($3H$) degrades almost all benchmarks, and adding a fourth serotonin-like pathway that modulates the membrane time scale is redundant. Neither shrinking nor growing the modulation representation helps, so the interpretable 12D bottleneck is, within this architecture, the minimal sufficient formulation rather than an arbitrary design choice.

Graph topology and coupling. Whereas replacing the learned adjacency with random Erdős–Rényi edges and removing coupling both degrade ChnSenti, and removing coupling additionally degrades MR, the effects on IMDB, SUBJ, SST-2, and SST-5 are neutral or weakly negative (Table 3); shrinking the Hebbian co-occurrence window from $w{=}5$ to $w{=}3$, in contrast, benefits the short-text benchmarks, consistent with a shared dependence of representation and graph utility on corpus sparsity.

Generic feature modulation control. Replacing the three-factor pathway with FiLM conditioned on the same 12 features degrades almost all datasets; a larger squeeze-and-excite block degrades five (Table 3). Both substitutes are zero-initialized to identity. Together with the capacity controls, these results favor low-dimensional interpretable modulation over the tested generic alternatives, without implying that any modulation helps.

4.4 Cognitive-Neuroscience Correspondences

Controlled probes. We test whether the learned dynamics reproduce cognitive predictions without the optional rule-based modules, on five SUBJ seeds plus MR and IMDB head checkpoints as controls. The strongest evidence involves external human data (Appendix O): the Hebbian em-

Table 3: Consolidated ablation study: Δ accuracy (pp) from each dataset's full model, averaged over five seeds; "$-$X" removes component X and ChnSenti arms are measured against its no-NE variant. ¶Degenerate arm: all five seeds predict a single class, so the entry is that class's test proportion (17.60%). Remaining arms in Appendix I.

Dataset	$-$Nm	$-$DA	$-$NE	$-$ACh	Rand.adj	No coup.	Aux=0	sMod	fMod	FiLM	SE
ChnSenti	-0.75	-0.04	-0.55	$+0.23$	-1.03	-0.55	-1.16	-1.04	-0.81	-0.68	-1.17
MR	-0.26	$\mathbf{-1.09}$	-0.60	-0.36	$+0.06$	$\mathbf{-1.41}$	$\mathbf{-2.20}$	-0.30	-0.30	-0.37	-0.82
IMDB	$+0.05$	-0.29	-1.99	-0.07	$+0.06$	$+0.20$	-0.28	-0.24	$+0.21$	$+0.27$	$+0.15$
SUBJ	-0.17	-0.84	-0.28	-0.26	$+0.10$	-0.23	-0.28	-0.64	-0.59	-0.33	$+0.03$
SST-2	-0.76	$+0.23$	$\mathbf{-1.12}$	-0.25	$+0.04$	-0.23	-0.78	-0.80	-0.54	-0.82	-0.27
SST-5	-0.35	-0.45	-0.49	-0.03	-0.05	-0.18	$\mathbf{-25.41}$¶	-0.10	-0.07	-0.46	-0.53
Waimai	-0.25	-0.24	-0.11	-0.41	$+0.17$	-0.25	-0.24	-0.17	-0.21	-0.31	-0.32

bedding space aligns with human valence norms, although not as the single principal factor that Osgood's semantic differential predicts ($\Delta = +0.446$ over a random-embedding control, 95% CI $[0.407, 0.486]$); and a preregistered behavioral probe confirms an endpoint attention advantage on MR, SST-2, and SST-5, though the U-shape itself is inherited from initialization rather than learned. Three further probes, on semantic satiation, neuromodulatory baseline calibration, and negation scope, are weaker and we report them as such rather than as evidence for a mechanism Kounios et al. (2000); Dager et al. (2025); Aston-Jones & Cohen (2005); Venkatesh (2026); Kaup et al. (2006). These are controlled correspondences, not claims of neural equivalence; full stimuli, tests and numerics appear in Appendices L–O (Figure 5).

5 Discussion

Neuromodulation as a complementary pathway. Across the datasets where it was ablated, modulation helps in almost all conditions, but its effect shrinks with trainable Hebbian embeddings and grows again after pruning (pathway-level and per-channel switches are distinct interventions). This supports complementarity, not independence: plastic representations absorb part of the adaptation otherwise carried by threshold and gain control Yamada & Chao (2026).

Scope of interpretability. Ne-CCNN imposes an affective coordinate system rather than recovering one post hoc Sun et al. (2026); Radwan (2026). Its claim is therefore mechanistic and limited: named variables can be intervened on and traced, but that alone does not establish human-like semantics. The compact regime keeps component effects measurable rather than absorbed by redundant capacity.

Graph and capacity. The semantic graph is an inductive bias, not a universal source of gain: it helps MR and ChnSenti but is neutral elsewhere Mantri et al. (2026); Liao et al. (2025); Misra et al. (2026). The core uses 1.49–3.01M non-embedding parameters against the matched BiLSTM's 2.85M, while pruning gives only configuration-dependent gains (Appendix I).

Limitations. Ne-CCNN trails BiLSTM on minimal-length inputs such as SST-2, exposing the limits of corpus-only representations. The sentence-pair branch is untested on semantic textual similarity. The graph is fixed during recurrence, and the resonance module requires quadratic pairwise preprocessing. The contraction result is an operating-region certificate, not a worst-case guarantee, and all evidence is at compact mechanism scale.

6 Conclusion

Ne-CCNN adapts continuous coupled CCNN dynamics to language: corpus-built embeddings define token units, a polarity-gated graph supplies content-dependent coupling, and about 1,000 neuromodulatory parameters adjust thresholds and gain. Under matched representations it is competitive on single sentences and strongest on sentence pairs, with near-uniform neuromodulation benefits and task-dependent graph-coupling benefits. Its polarity-routed readout and contraction condition make the adaptation inspectable; limits remain short-text accuracy, scaling, and certification.

AI USE STATEMENT

We used large language models and automated LaTeX tools to convert this paper to the ICLR 2027 format, verify references, and polish wording. The authors reviewed and verified all technical claims, numbers, equations, figures, and tables.

REFERENCES

Gary Aston-Jones and Jonathan D. Cohen. An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, 28:403–450, 2005. doi: 10.1146/annurev.neuro.28.061604.135709.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1409.0473.

Margarita Bugueño and Gerard de Melo. Rethinking graph-based document classification: Learning data-driven structures beyond heuristic approaches. *arXiv preprint arXiv:2508.00864*, 2025. URL https://arxiv.org/abs/2508.00864.

Emily Cheng, Aditya R. Vaidya, and Richard Antonello. Abstraction induces the brain alignment of language and speech models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=n5Ds4qbtjM.

Daniel Dager, Robin Sobczyk, Emmanuel Chemla, and Yair Lakretz. A neural model for word repetition. In *Conference on Cognitive Computational Neuroscience (CCN)*, 2025. URL https://openreview.net/forum?id=V9p0CKrR4q.

Jisheng Dang, Delin Deng, Bimei Wang, Jingze Wu, Hui Zhang, Haijiang Li, Jingmei Jiao, Dengyue Pan, Mangang Xie, and Jizhao Liu. Primary visual cortex inspired point cloud analysis framework. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 3506–3514, 2026. doi: 10.1609/aaai.v40i5.37348.

Peter Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, Cambridge, MA, 2001. URL https://mitpress.mit.edu/9780262041997/theoretical-neuroscience/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.

Pascal Fries. A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence. *Trends in Cognitive Sciences*, 9(10):474–480, 2005. doi: 10.1016/j.tics.2005.08.011.

Yutong Gao, Qinglin Meng, Yuan Zhou, and Liangming Pan. Towards intrinsic interpretability of large language models: A survey of design principles and architectures. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 34741–34754. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1605.

Shervin Ghasemlou. Dopamine: Brain modes, not brains — activation-gated PEFT via threshold-and-gain tuning. *arXiv preprint arXiv:2602.11726*, 2026. URL https://arxiv.org/abs/2602.11726.

Murray Glanzer and Anita R. Cunitz. Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, 5(4):351–360, 1966. doi: 10.1016/S0022-5371(66)80044-0.

Ramin Hasani, Mathias Lechner, Alexander Amini, Daniela Rus, and Radu Grosu. Liquid time-constant networks. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 7657–7666, 2021. doi: 10.1609/aaai.v35i9.16936.

Michael E. Hasselmo. Neuromodulation and cortical function: Modeling the physiological basis of behavior. *Behavioural Brain Research*, 67(1):1–27, 1995. doi: 10.1016/0166-4328(94)00113-T.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

John L. Johnson and Mary Lou Padgett. PCNN models and applications. *IEEE Transactions on Neural Networks*, 10(3):480–498, 1999. doi: 10.1109/72.761706.

Barbara Kaup, Jana Lüdtke, and Rolf A. Zwaan. Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, 38(7):1033–1050, 2006. doi: 10.1016/j.pragma.2005.09.012.

Yoon Kim. Convolutional neural networks for sentence classification. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751, 2014. doi: 10.3115/v1/D14-1181.

David Aaron Koplow, Tomaso Poggio, and Liu Ziyin. Ubiquity of emergent Hebbian dynamics in regularized learning. In *International Conference on Machine Learning (ICML)*, Proceedings of Machine Learning Research, Seoul, South Korea, 2026. URL https://openreview.net/forum?id=fSRmJOzMA1.

John Kounios, Sonja A. Kotz, and Phillip J. Holcomb. On the locus of the semantic satiation effect: Evidence from event-related brain potentials. *Memory & Cognition*, 28(8):1366–1377, 2000. doi: 10.3758/BF03211837.

Sebastian Lee, Samuel Liebana Garcia, Claudia Clopath, and Will Dabney. Lifelong reinforcement learning via neuromodulation. *arXiv preprint arXiv:2408.08446*, 2024. URL https://arxiv.org/abs/2408.08446.

Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Conference on Computational Natural Language Learning (CoNLL)*, pp. 171–180. Association for Computational Linguistics, 2014. doi: 10.3115/v1/W14-1618.

Xinfeng Liao, Xuanqi Chen, Lianxi Wang, Jiahuan Yang, Zhuowei Chen, and Ziying Rong. OTESGN: Optimal transport-enhanced syntactic-semantic graph networks for aspect-based sentiment analysis. In *IEEE International Conference on Data Mining (ICDM)*, pp. 447–456, 2025. doi: 10.1109/ICDM65498.2025.00052.

Jizhao Liu, Jing Lian, Julien C. Sprott, Qidong Liu, and Yide Ma. The butterfly effect in primary visual cortex. *IEEE Transactions on Computers*, 71(11):2803–2815, 2022. doi: 10.1109/TC.2022.3173080.

Peiyu Liu, Jianhao Ding, and Zhaofei Yu. Training deep normalization-free spiking neural networks with lateral inhibition. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=U8preGvn5G.

Sijia Lu and Feng Xu. Linear leaky-integrate-and-fire neuron model based spiking neural networks and its mapping relationship to deep neural networks. *Frontiers in Neuroscience*, 16, 2022. doi: 10.3389/fnins.2022.857513.

Changze Lv, Jianhan Xu, and Xiaoqing Zheng. Spiking convolutional neural networks for text classification. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=pgU3k7QXuz0.

Changze Lv, Tianlong Li, Weiming Qiao, Xiaohua Wang, Muling Wu, Wenhao Liu, Shihan Dou, Xiaoqing Zheng, and Xuanjing Huang. SpikeBERT: A language spikformer learned from BERT with knowledge distillation. *Neural Networks*, 197:108482, 2026. doi: 10.1016/j.neunet.2025.108482.

Krishna Sri Ipsit Mantri, Carola-Bibiane Schönlieb, Zorah Lähner, and Moshe Eliasof. Towards improved sentence representations using token graphs. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=stMX9KBhUI.

Jie Mei, Alejandro Rodriguez-Garcia, Daigo Takeuchi, Gabriel Wainstein, Nina Hubig, Yalda Mohsenzadeh, and Srikanth Ramaswamy. Improving the adaptive and continuous learning capabilities of artificial neural networks: Lessons from multi-neuromodulatory dynamics. *arXiv preprint arXiv:2501.06762*, 2025. URL https://arxiv.org/abs/2501.06762.

Rajarshi Misra, Aditya Sharma, Vinti Agarwal, and Hari Om Aggrawal. ModTGCN: Modularity-aware graph neural networks for text classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp. 226–239, 2026. doi: 10.1007/978-981-92-1465-5_18.

Bennet B. Murdock. The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5):482–488, 1962. doi: 10.1037/h0045106.

Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. *The Measurement of Meaning*. University of Illinois Press, Urbana, IL, 1957. URL https://www.press.uillinois.edu/books/?id=p745393.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. Association for Computational Linguistics, 2014. doi: 10.3115/v1/D14-1162.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 32, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

Yousef Radwan. Nine emotion centroids: A label-free valence axis that transfers across four modalities. *arXiv preprint arXiv:2608.18090*, 2026. URL https://arxiv.org/abs/2608.18090.

Alejandro Rodriguez-Garcia, Christopher J. Whyte, Brandon R. Munn, Jie Mei, James M. Shine, and Srikanth Ramaswamy. The role of gain neuromodulation in layer-5 pyramidal neurons. In *Annual Computational Neuroscience Meeting (CNS)*, 2025. URL https://arxiv.org/abs/2507.03222.

Antonio De Santis, Schrasing Tong, Marco Brambilla, and Lalana Kagal. Learning concept bottleneck models from mechanistic explanations. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=gdEWoxhb70.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. In *Findings of the Association for Computational Linguistics (EMNLP)*, pp. 1690–1712. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.89.

Wolf Singer and Charles M. Gray. Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience*, 18:555–586, 1995. doi: 10.1146/annurev.ne.18.030195.003011.

Lihao Sun, Lewen Yan, Xiaoya Lu, Andrew Lee, Jie Zhang, and Jing Shao. Valence-arousal subspace in LLMs: Circular emotion geometry and multi-behavioral control. *arXiv preprint arXiv:2604.03147*, 2026. URL https://arxiv.org/abs/2604.03147.

Vinitra Swamy, Syrielle Montariol, Julian Blackwell, Jibril Frej, Martin Jaggi, and Tanja Käser. Intrinsic user-centric interpretability through global mixture of experts. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=wDcunIOAOk.

Bo Tang and Weiwei Xie. Scalable learning in structured recurrent spiking neural networks without backpropagation. *arXiv preprint arXiv:2605.00402*, 2026. URL https://arxiv.org/abs/2605.00402.

Sohan Venkatesh. Negative before positive: Asymmetric valence processing in large language models. In *Mechanistic Interpretability Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=3FqGwkuggE.

Zuo Wang and Ye Yuan. TextGSL: Graph-sequence learning for inductive text classification. In *International Conference on Knowledge Science, Engineering and Management (KSEM)*, pp. 353–362, 2026. doi: 10.1007/978-981-92-2868-3_30.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207, 2013. doi: 10.3758/s13428-012-0314-x.

Xingrun Xing, Zheng Zhang, Ziyi Ni, Shitao Xiao, Yiming Ju, Siqi Fan, Yequan Wang, Jiajun Zhang, and Guoqi Li. SpikeLM: Towards general spike-driven language modeling via elastic bi-spiking mechanisms. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54698–54714, 2024. URL https://openreview.net/forum?id=4PB1RMsUy4.

Xingrun Xing, Boyan Gao, Zheng Liu, David A. Clifton, Shitao Xiao, Wanpeng Zhang, Li Du, Zheng Zhang, Guoqi Li, and Jiajun Zhang. SpikeLLM: Scaling up spiking neural network to large language models via saliency-based spiking. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=ZadnlOHsHv.

Yohei Yamada and Zenas C. Chao. Joint encoding of "what" and "when" predictions through error-modulated plasticity in biologically-plausible spiking networks. *Communications Biology*, 9, 2026. doi: 10.1038/s42003-026-10836-2.

Kejie Zhao, Wenjia Hua, Aiersi Tuerhong, Luziwei Leng, Yuxin Ma, and Qinghai Guo. Threshold modulation for online test-time adaptation of spiking neural networks. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10, 2025. doi: 10.1109/IJCNN64981.2025.11228835.

Chenlin Zhou, Liutao Yu, Zhaokun Zhou, Han Zhang, Jiaqi Wang, Huihui Zhou, Zhengyu Ma, and Yonghong Tian. Spikingformer: A key foundation model for spiking neural networks. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 2236–2244, 2026. doi: 10.1609/aaai.v40i3.37207.

Rui-Jie Zhu, Qihang Zhao, Guoqi Li, and Jason K. Eshraghian. SpikeGPT: Generative pre-trained language model with spiking neural networks. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gcf1anBL9e.

A Dataset Statistics and Preprocessing

All datasets use a unified cleaning pipeline: HTML entity decoding, tag removal via a regular expression for HTML tags, URL stripping, and whitespace normalization. For IMDB, this removes `<br>` tags from ∼30% of reviews. Table 4 summarizes the datasets.

Table 4: Dataset statistics over all available splits. P95 = 95th percentile of sentence length in words (IMDB is truncated to 256 tokens during training, 512 in the pruning configuration, so that entry is the raw review length); PPMI-NNZ = non-zero entries of the Hebbian PPMI matrix (window 5).

Dataset	Train	Test	Vocab	Mean len	P95	PPMI-NNZ
IMDB	25,000	25,000	12,851	230	591	8,048,678
MR	8,530	1,066	9,321	21	38	518,393
SUBJ	8,000	2,000	10,823	25	43	549,422
SST-2	67,349	872	12,760	9	20	393,443
SST-5	8,544	2,210	8,858	19	35	415,822

Per-dataset defaults for maximum sequence length and batch size are: IMDB 256 and 8, extended to 512 in the final pruning configuration of Table 10; SST-2 128 and 16; MR 64 and 32; SUBJ 96 and 24; SST-5 64 and 32. In every case the vocabulary is capped at 20,000 types, and words occurring fewer than twice in the training split are dropped. The SST-2 evaluation set (872 examples) is the official development split, as test labels are not publicly available. MR additionally provides a 1,066-example validation split, which is used only for Hebbian embedding construction and never for evaluation. The 1,101-example SST-5 validation split is unused.

Table 5 provides a per-dataset parameter budget. The context-embedding block comprises the static lookup table (shared with all baselines under identical Hebbian embeddings) and the dynamic context-gating network (1.195M parameters, Ne-CCNN-specific, absent from the baselines). The Non-CE column deducts both. For a standard lookup-only comparison, add back the Dynamic column.

Parameter counting conventions. The Params column of Table 1 counts learnable non-embedding parameters for the word-level rows, so the static table is excluded; because that table is shared with every baseline and scales with the vocabulary of each benchmark, the total parameter counts of the fixed-configuration baselines range from 1.1M to 8.9M across benchmarks even though their architectures are unchanged. Ne-CCNN tunes the width of its recurrent head per dataset (128–512), which is why its non-embedding count is reported as a range, whereas the baselines use a single fixed configuration throughout; on the sentence-pair tasks of Table 2 the baselines are width-aligned with Ne-CCNN. Pre-trained rows report full-model totals, and SpikeLM inherits the parameter structure of BERT-base, hence the identical total.

Table 5: Per-dataset parameter budget measured from the trained models of Table 10. Non-CE = total minus context embedding (static lookup + dynamic gating). †The IMDB totals include 5 additional parameters from the optional bio modules (Affective Terrain, River Gate, Negation Scope) enabled in its final configuration. The IMDB row reflects the post-pruning 3-layer structure (Non-CE 1.29M; with dynamic gating 2.48M); the pre-pruning 5-layer structure, which produced four of the five IMDB seeds of Table 1, has a 3.01M core (with dynamic gating). Embedding columns use the full-corpus vocabularies of Table 4; the final checkpoints use train-split vocabularies (0.6–19% smaller), which slightly reduces their Static and Total entries.

	Dataset		Embedding		Core Dynamics					Readout			Non-CE
Dataset	Vocab	Total	Static	Dynamic	Pos	Edge	Input	Coupling	Resonance	Neuro	Attn	Output	Total
IMDB	12,851	6,336,763†	3,855,300	1,195,200	153,600	154,625	154,112	789,504	31,810	1,051	525	1,031	1,286,263†
MR	9,321	4,306,046	2,796,300	1,195,200	153,600	38,657	38,528	50,112	31,810	1,051	525	263	314,546
SUBJ	10,823	4,739,814	3,246,900	1,195,200	153,600	38,657	38,528	33,280	31,810	1,051	525	263	297,714
SST-2	12,760	5,563,218	3,828,000	1,195,200	153,600	77,313	77,056	198,144	31,810	1,051	525	519	540,018
SST-5	8,858	4,167,533	2,657,400	1,195,200	153,600	38,657	38,528	50,112	31,810	1,051	525	650	314,933

B Extended Related Work

Relation to the original CCNN (point-by-point). Ne-CCNN retains the two defining properties of the CCNN Liu et al. (2022); Johnson & Padgett (1999): shared transform weights within a layer and independent neurons coupled to their neighbors. Table 6 lists every substantive departure. In brief, the receptive-field geometry of image CCNN is replaced by a content-dependent semantic graph; the multiplicative linking term $U = F(1+\beta L)$ is replaced by additive coupling inside a leaky integrator; the single global gain and fixed threshold are replaced by three-factor neuromodulatory regulation driven by named text statistics; raw pixel inputs are replaced by corpus-internal embeddings with per-token polarity; and the continuous output is replaced by a polarity-routed readout. Only the iteration itself is unchanged: a time-invariant update map applied repeatedly, now with a certified contraction condition rather than an appeal to convergence. Figure 3 renders this mapping unit by unit.

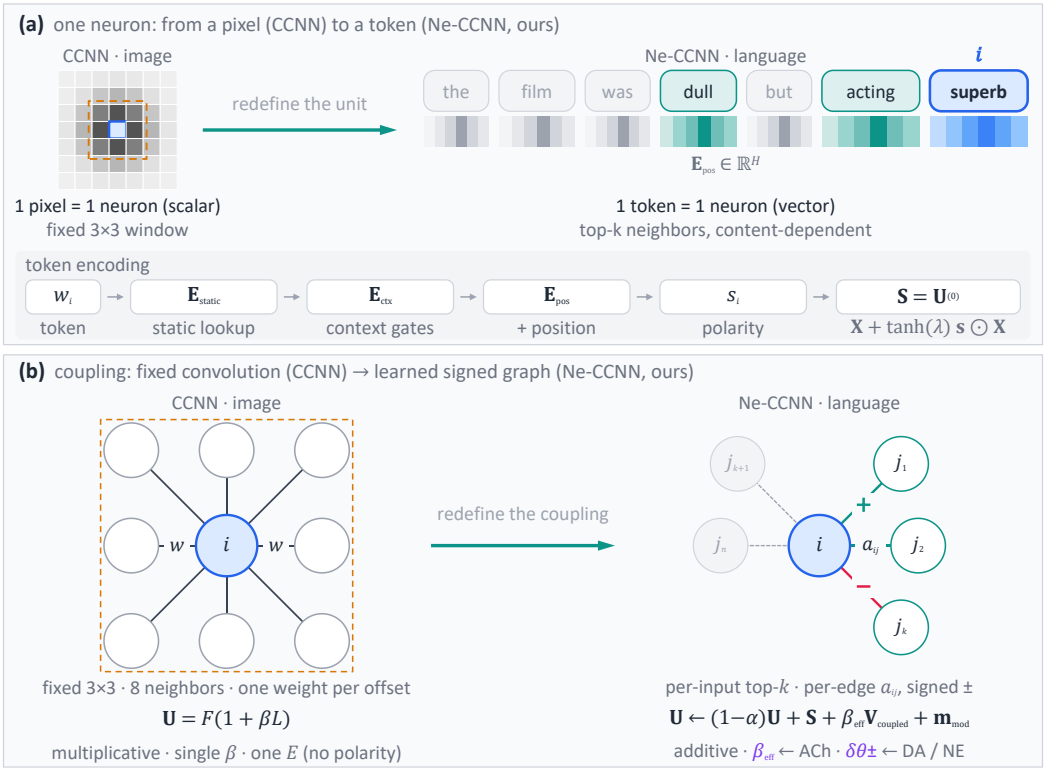

Figure 3: From the cited image CCNN Liu et al. (2022) (left) to Ne-CCNN (right, ours). (a) A unit: a CCNN pixel is a scalar-state neuron on a fixed 3×3 grid whose weight template has a zero center, coupling it to the eight surrounding neurons; a token is a neuron whose state is the H-dimensional $\mathbf{E}_{\text{pos}}$ produced by the corpus-only Hebbian chain below. (b) Coupling: the CCNN uses a position-shared kernel with a multiplicative linking term $U = F(1 + \beta L)$, one global gain β and one activity-dependent threshold E; Ne-CCNN uses per-edge signed weights a_{ij} on a per-input top-k graph, additive coupling in the same leaky membrane, and three-factor regulation of gain and threshold ($\beta_{\text{eff}} = \beta(1 + \delta\beta)$ from ACh, $\delta\theta^{\pm}$ from DA/NE). Table 6 lists the same comparison point by point.

LIF/PSN grafting degeneracy (detail). Directly grafting LIF or PSN spiking neurons onto a pre-trained BERT (the LIF-BERT and PSN-BERT variants of SpikeLM Xing et al. (2024)) degenerates to majority-class prediction: in our controlled replication, all LIF-BERT runs on MRPC predict the majority class (accuracy 66.49%, exactly the majority proportion $1147/1725$, with F1 79.87% matching the all-one-class formula $2\,\text{acc}/(1 + \text{acc})$), and LIF/PSN runs on RTE collapse identically ($52.71\% = 146/277$), across all seeds. The root cause is that the protocol patch re-spikes only the attention keys and values with binary spikes at a fixed threshold, discarding the ternary output range,

Table 6: Point-by-point comparison with the original CCNN Liu et al. (2022). "Shared" denotes parameters reused across positions or iterations, as in the original neuron model.

Aspect	Original CCNN	Ne-CCNN
Domain	image / video	text (token sequences)
Receptive field	fixed spatial neighborhood	per-input top-k semantic graph
Graph dynamics	static, geometry-determined	built once per input, fixed during recurrence
Coupling term	multiplicative, $U = F(1 + \beta L)$	additive, $(1-\alpha)U + \beta_{\text{eff}} V_{\text{coupled}}$
Coupling weights	kernel shared across positions	W_{coupled} shared across tokens; per-token edge gate
Gain	single global constant β	$\beta_{\text{eff}} = \beta(1 + \delta\beta)$, $\delta\beta$ from ACh
Threshold	fixed (or single activity-dependent)	polarity-assigned, $\delta\theta^{\pm}$ from DA/NE
Adaptation	none (raw pixel input)	12 named text features $\rightarrow$ three signals
Inputs	pixels from a fixed grid	PPMI–SVD embeddings + token polarity
Readout	continuous output Y	polarity-routed readout (+ pair comparison)
Iteration	converge to steady state	fixed T Euler steps, contraction certified
Theory	periodic / chaotic regimes analyzed	unique input-dependent fixed point, Lipschitz bound
Weight sharing	shared across positions & iterations	shared across tokens & iterations (retained)
Neuron independence	independent membrane states	independent $[B, L, H]$ membrane states (retained)

learnable step size, and per-time-step quantization that the original Elastic Bi-Spiking mechanism pairs with spiking; a further defect is that PSN states are never reset. This both reproduces the near-zero LIF/PSN results in the SpikeLM paper and motivates our design choice of continuous (non-spiking) membrane dynamics.

Hebbian-signature identifiability (detail). Koplow, Poggio, and Ziyin Koplow et al. (2026) prove that near stationarity, L_2 weight decay drives the learning signal of broad classes of update rules into Hebbian-aligned directions, with alignment increasing monotonically with decay strength, while strong stochastic noise induces anti-Hebbian alignment, yielding a decoupled trade-off with a phase boundary in regression settings. Consequently, observing Hebbian structure in trained-network updates does not uniquely imply mechanistic Hebbian computation. Ne-CCNN is unaffected by this identifiability ambiguity because it imposes Hebbian structure mechanistically at embedding construction time (explicit PPMI factorization) rather than relying on emergent signatures from regularization.

Full census arms. The 21-arm component census additionally includes threshold-exit arms ($-\delta\theta_{\text{pos}}$ exit: $-0.32/-0.08/-0.30$, $-\delta\theta_{\text{neg}}$ exit: $-0.45/0.00/-0.16$ on MR/ChnSenti/SST-2), PRR internal arms (neutral drop: $-0.71/+0.03/+0.07$, neutral full: $-0.23/+0.01/-0.18$, uniform attention: $-0.69/+0.08/+0.18$), replacement arms (max-pool: $-0.47/+0.26/+0.30$, vector-wise max: $-1.73/-0.35/-0.39$, LCA$\times$2: $-0.41/+0.17/-0.05$), and the LTC variant formulation with per-neuron gated residual $\mathbf{g}_{\text{ltc}}^{(n)} = \tau\text{-net}(\text{mean}_l(\mathbf{V}^{(t)}))$, $\mathbf{V}_{\text{ltc}}^{(t)} = \mathbf{g}_{\text{ltc}}^{(n)} \odot \mathbf{V}^{(t)} + (1 - \mathbf{g}_{\text{ltc}}^{(n)}) \odot \mathbf{V}_{\text{coupled}}^{(t)}$ (τ-net: Linear($H \rightarrow H/4$)$\rightarrow$LayerNorm$\rightarrow$Tanh$\rightarrow$Linear($H/4 \rightarrow H$)$\rightarrow$Sigmoid, $\approx H^2/2$ parameters per layer).

Graph-topology seven-dataset grid. Full-model accuracy with Random-adjacency and No-coupling ablations across all seven sentiment datasets: ChnSenti 89.18/88.15/88.63, MR 78.10/78.16/76.69, IMDB 86.48/86.54/86.68, SUBJ 90.70/90.80/90.47, SST-2 80.59/80.63/80.36, SST-5 43.01/42.96/42.83, Waimai 89.46/89.63/89.21 (clean evaluation protocol, 5 seeds).

C Negative Results and Architecture Validation

Neuromodulation Ablation (Hebbian Trainable Embeddings). Table 7 reports the effect of disabling DA/NE/ACh ($\delta\theta^{+} = \delta\theta^{-} = \delta\beta = 0$) on the current architecture. The contribution shrinks to 0.19–1.35% compared with 6.31% under static GloVe, supporting the complementarity hypothesis that trainable Hebbian embeddings absorb part of the adaptive load. On the final IMDB configuration the pathway-level switch of Table 3 is statistically indistinguishable from zero (mean change $+0.05$ pp; two-sample Welch $p = 0.80$), while the per-channel ablations remain negative. The two switches intervene at different stages: the pathway-level arm builds the model without the modulation modules (a construction-time removal), whereas the per-channel arms zero a single factor's output inside a network that retains the full pathway (a generation-time intervention applied iden-

Table 7: Neuromodulation ablation on the current architecture (Hebbian embeddings). Δ denotes the accuracy drop when disabling DA/NE/ACh, each relative to that row's own ON baseline. These configurations (256/5/3; pre- and post-prune SOTA) differ from the final-configuration census of Table 3, so the two tables' deltas are not directly comparable.

Configuration	ON Acc	OFF Acc	Δ
IMDB 256/5/3 baseline	87.21%	86.20%	−1.01%
MR 128/3/2 baseline	77.77%	77.58%	−0.19%
IMDB SOTA (pre-prune)	87.03%	86.48%	−0.55%
IMDB SOTA (post-prune FT)	87.55%	86.20%	−1.35%

tically during training and inference). Their magnitudes therefore need not agree, and the ordering $|\Delta(\text{all-off})| \geq |\Delta(\text{single channel})|$ is not implied by the design; a single factor's removal can also disrupt the balance among the three signals, which no single scalar can predict.

Sparse Coupling. A sparse adjacency-list variant did not reduce GPU memory, so activation checkpointing was adopted instead.

Hippocampal Buffer / Synaptic Consolidation / Replay. All produced neutral or negative results ($\leq$0.1% change) and are not used in final configurations.

D Biological and Mathematical Foundations

This appendix formalizes the correspondence between Ne-CCNN's components and standard models in computational neuroscience. The derivations show that the architecture is not merely bio-inspired terminology but implements well-defined dynamical approximations.

D.1 CCNN Dynamics as a Discretized Leaky Integrator

The membrane update (9 of the main text) is

$$\mathbf{U}^{(t+1)} = \Pi_{[-10,10]}\Big((1-\alpha)\mathbf{U}^{(t)} + \mathbf{S} + \beta_{\text{eff}}\,\mathbf{V}^{(t)}_{\text{coupled}} + \mathbf{m}_{\text{mod}}\Big), \tag{15}$$

with $\beta_{\text{eff}} = \beta(1+\delta\beta)$ and $\mathbf{m}_{\text{mod},b} = \frac{\gamma}{2}(\delta\theta_b^+ + \delta\theta_b^-)$. Writing $\alpha = 1-\Delta t/\tau$ for discrete time step Δt and membrane time constant τ, and collecting the external and recurrent inputs into the autonomous forcing term

$$\mathbf{F}\big(\mathbf{U}^{(t)}\big) = \mathbf{S} + \beta_{\text{eff}}\,\mathbf{V}^{(t)}_{\text{coupled}}\big(\mathbf{U}^{(t)}\big) + \mathbf{m}_{\text{mod}}, \tag{16}$$

where $\mathbf{V}_{\text{coupled}}$ depends on the current state through the firing rate $\mathbf{V} = \sigma(\mathbf{U}-\boldsymbol{\theta})$ and $\mathbf{m}_{\text{mod}}$ is an input-conditioned, state-independent bias, the update is a unit-step Euler discretization ($\Delta t = 1$) of the autonomous ODE followed by the elementwise projection $\Pi_{[-10,10]}$

$$\tau\frac{d\mathbf{U}}{dt} = -\mathbf{U} + \tau\,\mathbf{F}(\mathbf{U}). \tag{17}$$

The leaky-integrator form in (17) is standard for the subthreshold dynamics of conductance-based and LIF models Dayan & Abbott (2001). In Ne-CCNN the discrete step is $\Delta t = 1$, so $\alpha = 1 - \Delta t/\tau = 1-1/\tau$, equivalently $\tau = 1/(1-\alpha) > 1$; τ is task-adapted via the learned decay parameter $\alpha = \sigma(\alpha_{\text{raw}}) \in (0,1)$.

Convergence. Because the firing nonlinearity is the smooth sigmoid $\mathbf{V} = \sigma(\mathbf{U}-\boldsymbol{\theta})$ rather than the Dirac pulse $\Theta(\mathbf{U}-\boldsymbol{\theta})$, gradients flow directly through the membrane update, enabling end-to-end backpropagation through time without surrogate gradients. At a fixed point $\mathbf{U}^*$ we require $\mathbf{U}^* = \tau\,\mathbf{F}(\mathbf{U}^*)$. §D.2 turns the convergence of the discrete iteration into a rigorous and measurable contraction analysis. Empirically, Table 13 confirms that $T = 3$ suffices on 512-token documents, whereas rate-coded LIF SNNs for text typically require $T \approx 50$ steps Lv et al. (2023).

D.2 Contraction Certificates and Convergence Rate

The convergence discussion above is conditional. This subsection makes it rigorous and measurable. Write the single-step update as $\mathbf{U}^{(t+1)} = \mathbf{T}(\mathbf{U}^{(t)})$ with $\mathbf{T} = \Pi \circ \mathbf{A}$, where Π is the elementwise

projection onto $[-10, 10]$ and

$$\mathbf{A}(\mathbf{U}) = (1-\alpha)\,\mathbf{U} + \mathbf{S} + \beta_{\text{eff}}\, C\big(\sigma(\mathbf{U} - \boldsymbol{\theta})\big) + \mathbf{m}_{\text{mod}}, \tag{18}$$

with C the stacked coupling subnetwork (the LayerNorm-wrapped, adjacency-weighted linear maps of Section 3.4 in the main text).

Status of the bound. The bound is not vacuous in the released models: at the final step, $|\mathbf{U}|$ exceeds 10 at 90.8% of real-token units on SST-2 and 98.7% on IMDB (reaching 32.3 and 31.9), with $E|\mathbf{U}| \approx 17.5$ in both cases. It is nevertheless not a functional component at the trained solution. The realized thresholds are strongly negative ($\bar{\theta} = -3.75$ on SST-2, -4.29 on IMDB), so $\sigma(\mathbf{U} - \boldsymbol{\theta})$ is already saturated at these magnitudes and is largely insensitive to the exact value of $\mathbf{U}$: removing the clamp displaces $\mathbf{U}$ by a mean of 16.5–20.0 but changes $\mathbf{V}$ by only 1.9×10^{-3}–3.4×10^{-3} on average, altering $0/872$ SST-2 and $11/1024$ IMDB predictions (Δacc 0.00 pp and $+0.29$ pp, the latter within seed noise). We therefore claim no accuracy contribution for Π. It is retained so that (9) of the main text matches the released implementation and so that the state space is compact, which is what makes the fixed-point argument below apply. Its effect on the optimization trajectory is not measured here, as we did not retrain without it.

Scope of the map: batch composition and thresholds. The thresholds $\theta_{\text{pos}}, \theta_{\text{neg}}$ are computed from batch-level means of the per-sample neuromodulatory signals, while β_{eff} and $\mathbf{m}_{\text{mod}}$ are per-sample constants, and none of these quantities depends on the membrane state. For a *fixed batch composition* they are therefore constants of the iteration, and $\mathbf{T}$ is a well-defined map on the batched state. The contraction argument then yields a unique fixed point for the whole batch, which is thus input-*and batch-composition*-dependent. Under per-sample thresholds (equivalently, deployment with batch size 1, where the batch mean degenerates to the sample's own signal), $\mathbf{T}$ is a per-sample autonomous map, the fixed point is a function of the single input, and Corollary 2 applies verbatim. Under batched inference, a perturbation of one input additionally shifts the shared thresholds by $O(\Delta/B)$, a cross-sample coupling attenuated by $1/B$ that vanishes at $B = 1$. The measurements in Table 8 are taken on single sentences (batch size 1), the per-sample regime. (The optional dynamic-threshold rule, disabled in all reported configurations, would add a state-dependent component $\theta^{(t+1)} = 0.9\,\theta^{(t)} + 0.1\,\mathbf{V}^{(t)}$. This affine-in-$\mathbf{V}$ update is itself contractive with factor $0.9 + 0.1/4 < 1$ and folds into q.)

Two facts make the analysis tractable. First, the projection Π onto the convex box $[-10, 10]$ is nonexpansive in ℓ_2: $\|\Pi(\mathbf{u}) - \Pi(\mathbf{v})\|_2 \le \|\mathbf{u} - \mathbf{v}\|_2$, so the clamp cannot destroy contractivity. Second, the sigmoid is $1/4$-Lipschitz because $\|\sigma'\|_\infty = 1/4$. LayerNorm is not globally 1-Lipschitz, so instead of an unverifiable worst-case bound we estimate the Lipschitz constant of the coupling map empirically on the state manifold $\mathcal{M}$ of activations realized along actual trajectories:

$$\hat{\kappa} = \sup_{\substack{\mathbf{x},\mathbf{y}\in\mathcal{M} \\ \mathbf{x}\neq\mathbf{y}}} \frac{\|C(\mathbf{x}) - C(\mathbf{y})\|_2}{\|\mathbf{x} - \mathbf{y}\|_2}, \qquad q := (1-\alpha) + \frac{|\beta_{\text{eff}}|\,\hat{\kappa}}{4}. \tag{19}$$

For the LTC coupling variant, each layer outputs a convex combination of its input and the coupled state, whose Lipschitz constant does not exceed that of the coupled branch. In both variants $\hat{\kappa}$ is measured end-to-end over the full stack. The absolute value in q accounts for signed gains: β_{eff} is negative when neuromodulation suppresses the coupling term below sign reversal (observed on IMDB), and the coupling contribution to the bound scales with $|\beta_{\text{eff}}|$. Because β_{eff} is per-sample, the condition is required for every realized gain; $\hat{q}$ in Table 8 evaluates it at the mean operating point, $\hat{q}_{\text{worst}} = (1-\alpha) + \max_b |\beta_{\text{eff},b}|\,\hat{\kappa}_{\max}/4$ at the realized worst case. Both $\hat{\kappa}$ (a 99th-percentile estimate over realized perturbations) and the realized $|\beta_{\text{eff}}|$ are empirical quantities, so $\hat{q}$ is an operating-point diagnostic rather than a worst-case guarantee. Proposition 1 certifies the dynamics only insofar as these estimates are representative.

Proposition 1 (Contraction condition). *If $q < 1$, then $\mathbf{T}$ is a contraction on the operating region with factor at most q. By the Banach fixed-point theorem, $\mathbf{U}^{(t)}$ converges to a unique fixed point $\mathbf{U}^*$ and $\|\mathbf{U}^{(t)} - \mathbf{U}^*\|_2 \le q^t\,\|\mathbf{U}^{(0)} - \mathbf{U}^*\|_2$.*

Proof sketch. For $\mathbf{u}, \mathbf{v} \in \mathcal{M}$,

$$\begin{aligned}\|\mathbf{A}(\mathbf{u}) - \mathbf{A}(\mathbf{v})\|_2 &\le (1-\alpha)\,\|\mathbf{u} - \mathbf{v}\|_2 \\ &\quad + |\beta_{\text{eff}}|\,\hat{\kappa}\,\big\|\sigma(\mathbf{u} - \boldsymbol{\theta}) - \sigma(\mathbf{v} - \boldsymbol{\theta})\big\|_2 \\ &\le q\,\|\mathbf{u} - \mathbf{v}\|_2,\end{aligned}$$

using the triangle inequality with $|\beta_{\text{eff}}|$, the $1/4$-Lipschitz sigmoid, and the definition of $\hat{\kappa}$; Π preserves the bound by nonexpansiveness. □

Corollary 1 (Geometric gradient decay). *Within the certified region, the single-step Jacobian* $\mathbf{J}_t = (1-\alpha)\mathbf{I} + \beta_{\text{eff}}\mathbf{J}_C \operatorname{diag}\big(\sigma'(\mathbf{U}^{(t)} - \boldsymbol{\theta})\big)$*, with* $\mathbf{J}_C$ *the Jacobian of* C*, satisfies*

$$\|\mathbf{J}_t\|_2 \le q \implies \big\|\partial\mathcal{L}/\partial\mathbf{U}^{(t)}\big\|_2 \le q^{T-t}\big\|\partial\mathcal{L}/\partial\mathbf{U}^{(T)}\big\|_2 :$$

gradients decay geometrically with an effective memory horizon of $\approx 1/(1-q)$ *steps, and exploding gradients are ruled out by construction.*

Corollary 2 (Lipschitz input-to-fixed-point map). *Let* $\mathbf{F}$ *and* $\tilde{\mathbf{F}}$ *be two forcing terms, e.g., those induced by an input and a perturbed version of it, with* $\sup_{\mathbf{u}} \|\mathbf{F}(\mathbf{u}) - \tilde{\mathbf{F}}(\mathbf{u})\|_2 \le \Delta$*, and let* $\mathbf{U}^*$ *and* $\tilde{\mathbf{U}}^*$ *denote the corresponding fixed points within a certified region with factor* $q < 1$*. Then*

$$\|\mathbf{U}^* - \tilde{\mathbf{U}}^*\|_2 \le \frac{\Delta}{1-q},$$

so the input-to-fixed-point inference map is Lipschitz with constant $1/(1-q)$*.*

Proof sketch. Since $\mathbf{U}^* = \mathbf{T}(\mathbf{U}^*)$ and $\tilde{\mathbf{U}}^* = \tilde{\mathbf{T}}(\tilde{\mathbf{U}}^*)$,

$$\|\mathbf{U}^* - \tilde{\mathbf{U}}^*\|_2 \le q\,\|\mathbf{U}^* - \tilde{\mathbf{U}}^*\|_2 + \sup_{\mathbf{u}} \|\mathbf{T}(\mathbf{u}) - \tilde{\mathbf{T}}(\mathbf{u})\|_2,$$

and the last term is at most Δ because Π is nonexpansive and $\mathbf{T}, \tilde{\mathbf{T}}$ differ only in their forcing terms. Rearranging gives the bound. □

Corollary 2 ties contractivity to the intrinsic robustness evaluated in the main text: a word-substitution attack perturbs the forcing term through the input injection $\mathbf{S}$, the adjacency, and the neuromodulatory bias, and wherever the condition holds, every downstream quantity (the fixed point, the firing rates, and the PRR evidence) can move by at most $\Delta/(1-q)$. The condition also yields a principled stopping rule for the number of recurrent steps: tolerance ε is guaranteed after $T \ge \log\big(\varepsilon/\|\mathbf{U}^{(0)} - \mathbf{U}^*\|_2\big)/\log q$ steps, and the realized rate $\hat{\rho}$ is consistently faster than this certified bound (Table 8). On configurations with $\hat{q} > 1$ both bounds are vacuous, and convergence and robustness there remain empirical properties of the learned operating region.

Table 8 reports the measured quantities. At the mean operating point the sufficient condition is estimated to hold on SUBJ and SST-2 ($\hat{q} = 0.92$ and 0.82); on IMDB, MR, and SST-5 it fails even at the mean, because the coupling stack is locally expansive along a small fraction of realized directions. Under the realized worst-case gain and Lipschitz estimate the condition fails on all five benchmarks ($\hat{q}_{\text{worst}} \in [1.6, 5.4]$). Proposition 1 must therefore be read as an interpretable sufficient condition evaluated at the operating point, a diagnostic that becomes a guarantee only insofar as the empirical estimates are representative, not as a worst-case certificate. What is consistent across all benchmarks is the *realized* behavior: the fitted decay rate $\hat{\rho}$ of $\|\mathbf{U}^{(t+1)} - \mathbf{U}^{(t)}\|_2$ lies in $[0.59, 0.71]$ (95% CIs within ± 0.12), the residual $\|\mathbf{U}^{(3)} - \mathbf{U}^{(15)}\|_2/\|\mathbf{U}^{(0)} - \mathbf{U}^{(15)}\|_2$ falls to 8–19% after three steps, and gradient norms decay backward in time at ≈ 0.6 per step, matching $\hat{\rho}$. Convergence and input sensitivity are thus guaranteed only where the estimates certify them, and are empirical properties of the learned operating region elsewhere.

Sequence length and computational cost. The expression $q = (1-\alpha) + |\beta_{\text{eff}}|\hat{\kappa}/4$ has no explicit L term, but $\hat{\kappa}$ is measured end to end and may depend on sequence length, width, normalization, and the realized adjacency. We therefore treat the observed sufficiency of $T = 3$ on 512-token IMDB documents as empirical evidence, not a length-independence theorem. With an already constructed top-k graph, recurrent propagation costs $O(LkT)$. However, the default resonance module evaluates pairwise token interactions and requires $O(L^2)$ preprocessing. Sparse non-resonant configurations avoid this pairwise stage, but the default model should not be described as linear-time.

Why training converges in a few epochs. Epoch-level convergence of nonconvex training does not admit a rigorous proof. We provide three converging lines of evidence. (i) Exploding gradients, a primary source of optimization instability in recurrent architectures, are ruled out within the certified

Table 8: Measured contraction quantities (200 test sentences per dataset, seed 42, batch size 1). $\bar{\beta}_{\text{eff}}$: mean neuromodulated coupling gain over test sentences ($\pm$ std, negative on IMDB, where neuromodulation suppresses the coupling term); $\max_b |\beta_{\text{eff}}|$: realized worst-case gain; $\hat{\kappa}$: empirical Lipschitz estimate of the coupling stack (99th percentile over perturbations along 15-step trajectories); $\hat{\kappa}_{\max}$: maximal realized ratio; $\hat{q} = (1-\alpha) + |\bar{\beta}_{\text{eff}}|\,\hat{\kappa}/4$ (mean operating point; < 1, in bold, means the sufficient condition is estimated to hold at the mean); $\hat{q}_{\text{worst}} = (1-\alpha) + \max_b |\beta_{\text{eff},b}|\,\hat{\kappa}_{\max}/4$ (worst-case estimate); $\hat{\rho}$: fitted decay rate of $\|\mathbf{U}^{(t+1)} - \mathbf{U}^{(t)}\|_2$.

Dataset	α	$\bar{\beta}_{\text{eff}}$	$\max_b \lvert\beta_{\text{eff}}\rvert$	$\hat{\kappa}$	$\hat{\kappa}_{\max}$	$\hat{q}$	$\hat{q}_{\text{worst}}$	$\hat{\rho}$
IMDB	0.365	-0.247 ± 0.12	0.236	34.44	50.60	2.759	3.619	0.594
MR	0.424	2.203 ± 0.16	2.602	1.20	2.28	1.239	2.059	0.699
SUBJ	0.444	0.839 ± 0.56	2.090	1.74	3.99	**0.920**	2.639	0.618
SST-2	0.220	0.063 ± 0.21	0.450	2.81	7.69	**0.824**	1.645	0.708
SST-5	0.374	0.698 ± 0.42	1.913	3.13	10.07	1.172	5.443	0.633

region (Corollary 1) and are never observed in practice: measured gradient norms decay backward in time at $\approx$0.6 per step on all five datasets, consistent with the narrow peak-epoch spread across seeds reported in the main text. (ii) The Hebbian PPMI+SVD initialization is the Eckart–Young-optimal rank-d solution of the co-occurrence objective (§D), so fine-tuning starts near a well-fitted solution and requires few effective updates. (iii) The recurrent core contains only $\sim$105K parameters (SUBJ configuration), so the effective capacity is small. The observed peak-then-decline pattern is the expected signature of rapid convergence followed by overfitting rather than instability.

D.3 Neuromodulatory Gain and Threshold Modulation

The three neuromodulatory signals alter the firing function $V = \sigma(U - \theta)$ through threshold shifts and a coupling gain:

$$\theta_{\text{pos}} = \theta_0 + \tfrac{1}{B} \sum_b \delta\theta_b^+, \tag{20}$$

$$\theta_{\text{neg}} = \theta_0 + \tfrac{1}{B} \sum_b \delta\theta_b^-, \tag{21}$$

$$\beta_{\text{eff}} = \beta(1 + \delta\beta). \tag{22}$$

Linearizing the firing rate around the base threshold gives

$$\Delta V \approx -\sigma'(U - \theta_0)\,\Delta\theta, \tag{23}$$

so increasing θ (positive $\delta\theta^+$ or $\delta\theta^-$) suppresses activation, shifting the f–I curve rightward. This matches the effect of tonic dopamine and norepinephrine on pyramidal neurons: by modulating voltage-gated conductances, catecholamines effectively change the current threshold for firing Aston-Jones & Cohen (2005).

The β term scales recurrent coupling. Linearizing $\beta_{\text{eff}} = \beta(1 + \delta\beta)$ around $\delta\beta = 0$ gives

$$\Delta V_{\text{coupled}} \approx \beta\,\sigma'(U - \theta_0)\,\delta\beta\,V_{\text{coupled}}, \tag{24}$$

which is multiplicative gain modulation of the recurrent input, mapping onto cholinergic modulation of cortical synaptic gain Hasselmo (1995). Thus the compact 1,051-parameter neuromodulatory system implements two distinct biological operations, threshold shifting (DA/NE) and gain scaling (ACh), computed from 12 interpretable text features through a shared 64-dimensional representation.

D.4 Polarity Gating and Excitation–Inhibition Balance

The polarity-gated adjacency matrix (7 of the main text) is

$$\mathbf{A} = (\mathbf{S}_{\text{sim}} \odot \mathbf{M}_{\text{top-}k} \odot \mathbf{M}_{\text{pol}}) \odot \mathbf{W}_{\text{edge}}, \tag{25}$$

where the polarity mask $\mathbf{M}_{\text{pol}} = \mathbf{1} - 2 \cdot (\mathbf{s}_{\text{sign}}\mathbf{s}_{\text{sign}}^\top < 0) \odot (\mathbf{s}_{\text{non-neu}}\mathbf{s}_{\text{non-neu}}^\top)$ suppresses edges between tokens of opposite sentiment when both are emotionally salient. This is a soft implementation of

excitation–inhibition (E–I) balance: same-polarity tokens form mutually excitatory neighborhoods, while oppositely polarized salient tokens are inhibited. The edge gate $\mathbf{W}_{\text{edge}}$ learns a data-dependent scaling of these signed connections, resembling the balancing of recurrent excitation and lateral inhibition in local cortical circuits Dayan & Abbott (2001).

D.5 Hebbian PPMI and Low-Rank Representation

The co-occurrence count

$$C_{ij} = \sum_{t} \sum_{\substack{k=-w \\ k\neq 0}}^{w} \frac{1}{|k|} \mathbf{1}[w_t = i]\, \mathbf{1}[w_{t+k} = j] \tag{26}$$

accumulates weighted co-activations of words i and j within a context window, a discrete approximation to the Hebbian rule $\Delta W_{ij} \propto x_i x_j$ averaged over the corpus. The PPMI transform

$$P_{ij} = \max\left(0,\ \log_2 \frac{C_{ij} \cdot N}{C_{i\cdot}\, C_{\cdot j}}\right) \tag{27}$$

renormalizes these correlations by marginal word frequencies. It is the positive part of an empirical estimate of the pointwise mutual information $I(i;j)$, measuring how much more often two words co-occur than expected under independence.

Truncated SVD of rank d gives the low-rank approximation $\mathbf{P} \approx \mathbf{U}_d \mathbf{\Sigma}_d \mathbf{V}_d^\top$. By the Eckart–Young–Mirsky theorem, this is the optimal rank-d approximation of $\mathbf{P}$ in Frobenius norm. The word vectors therefore form the d-dimensional subspace that best preserves pairwise associations in a least-squares sense, providing a rigorous justification for the embedding step independent of informal biological analogies. Compared to GloVe, which is trained on $\sim$6B tokens, our Hebbian embeddings use only the dataset's training data, yet the embedding-source ablation in the main text shows they provide a measurable and synergistic advantage when combined with CCNN dynamics and bio components.

E Module Formulations

This appendix gives the exact formulations of modules that are described at a higher level in the main text: Neural Resonance and River Gate (Stage 2), Negation Scope (Stage 1), the Bio-Syntactic Attention readout (Stage 5), and the multi-class PRR construction. All equations follow the implementation, and parameter counts match Table 5.

E.1 Neural Resonance (default on; 31,810 parameters)

Neural Resonance enhances the base adjacency matrix $\mathbf{A}^{\text{base}}$ (the row-normalized output of (7) of the main text, after top-k masking, polarity gating, edge gating, and optional River Gate attenuation) by detecting synchronized token pairs. Token embeddings are projected into a 64-dimensional resonance space,

$$\mathbf{h}_i = \mathbf{W}_{\text{res}}^{(2)}\, \text{Dropout}\big(\text{ReLU}(\mathbf{W}_{\text{res}}^{(1)} \mathbf{e}_i)\big) \in \mathbb{R}^{64}, \tag{28}$$

with $\mathbf{W}_{\text{res}}^{(1)} \in \mathbb{R}^{64\times 300}$ and $\mathbf{W}_{\text{res}}^{(2)} \in \mathbb{R}^{64\times 64}$. For each pair (i,j), a symmetrized pair feature and the mean absolute sentiment form the detector input:

$$\mathbf{p}_{ij} = \tfrac{1}{2}\big([\mathbf{h}_i;\mathbf{h}_j] + [\mathbf{h}_j;\mathbf{h}_i]\big), \qquad \bar{s}_{ij} = \tfrac{1}{2}\big(|s_i| + |s_j|\big), \tag{29}$$

$$\text{sync}_{ij} = \sigma\Big(\mathbf{w}_{\text{sync}}^{(2)}\, \text{ReLU}\big(\mathbf{W}_{\text{sync}}^{(1)}[\mathbf{p}_{ij};\bar{s}_{ij}]\big)\Big) \in (0,1), \tag{30}$$

where $\mathbf{W}_{\text{sync}}^{(1)} \in \mathbb{R}^{64\times 129}$ and $\mathbf{w}_{\text{sync}}^{(2)} \in \mathbb{R}^{1\times 64}$. The base adjacency is boosted multiplicatively,

$$\tilde{\mathbf{A}} = \mathbf{A}^{\text{base}} \odot \text{clip}\big(\mathbf{1} + \rho\, \text{sync} \cdot \sigma(b),\ 0.5,\ 3.0\big), \tag{31}$$

with resonance strength ρ=0.3 and a learnable scalar b initialized at 0.5. The result is symmetrized, masked, row-normalized in L1, and symmetrized once more to produce $\mathbf{A}$. Consequently, the base graph is exactly row-normalized, whereas the final resonance-enhanced graph is symmetric but not guaranteed to retain unit row norm. Parameter count: 19,264 + 4,160 (projection) + 8,320 + 65 (synchrony detector) + 1 (b) = 31,810.

E.2 River Gate (optional; 1 parameter)

A boundary mask marks the first token after each punctuation mark in $\{., !, ?, ;\}$, and the sequence start. Cumulative summation assigns each token a basin index $\beta_i \in \mathbb{N}$ (punctuation-delimited spans). Before row normalization in (7) of the main text, cross-basin edges are attenuated by a learned decay:

$$A_{ij}^{\text{raw}} \leftarrow A_{ij}^{\text{raw}} \cdot \big(\mathbf{1}[\beta_i = \beta_j] + d \cdot \mathbf{1}[\beta_i \neq \beta_j]\big), \quad d = \sigma(d_{\text{raw}}), \tag{32}$$

with d_{raw} a learnable scalar initialized so that d=0.1.

E.3 Negation Scope (optional; 0 parameters)

Negation Scope is a rule-based module with no learnable parameters. It scans the token sequence with three fixed lexica: 47 negation forms (e.g., *not*, *never*, *n't*, *without*, *hardly*), 11 adversative turn words (*but*, *although*, *though*, *however*, *nevertheless*, *nonetheless*, *yet*, *still*, *despite*, *except*, *apart*), and closing punctuation $\{., !, ?, :, ;\}$. A scope opens at any negation form (including the negation token itself) and closes after the next turn word or punctuation mark. A river boundary (above) also resets the scope. The polarity of tokens inside the scope is flipped: $s_i \leftarrow -s_i$.

E.4 Bio-Syntactic Attention Readout (525 parameters)

The per-token attention weights used in the decision readout of the main text (Section 3.5) are a normalized product of three terms (all applied multiplicatively, then masked and L1-normalized over valid positions):

$$\tilde{a}_i = w_i^{\text{pos}} \cdot w_i^{\text{emo}} \cdot w_i^{\text{turn}}, \qquad a_i = \frac{m_{\text{seq},i}\, \tilde{a}_i}{\sum_j m_{\text{seq},j}\, \tilde{a}_j}. \tag{33}$$

Primacy–recency position bias.

$$w_i^{\text{pos}} = u_i \cdot \sigma(\pi_i), \tag{34}$$

$$u_i = 1.5 - \tfrac{1}{2}\Big(1 + \cos\tfrac{\pi i}{L-1}\Big) \in [0.5, 1.5], \tag{35}$$

where $\boldsymbol{\pi} \in \mathbb{R}^{512}$ is a learnable positional gate initialized to a U-shape (linearly decaying from 1.3 to 0.9 over the first half and rising back to 1.3). **Emotional saliency.** $w_i^{\text{emo}} = 0.5 + 0.5\,|s_i| \in [0.5, 1.0]$. **Turning-point (surprisal) detection.**

$$w_i^{\text{turn}} = 1 + \sigma\Big(\mathbf{w}_{\text{nov}}^{(2)} \text{ReLU}\big(\mathbf{w}_{\text{nov}}^{(1)} v_i\big)\Big) \in [1, 2], \tag{36}$$

where v_i is the masked local variance of the sentiment sequence over a window of 3 tokens, and the novelty MLP has hidden size 4 ($\mathbf{w}_{\text{nov}}^{(1)} \in \mathbb{R}^{4\times1}$, $\mathbf{w}_{\text{nov}}^{(2)} \in \mathbb{R}^{1\times4}$; 13 parameters). Parameter count: 512 (position gate) + 13 (novelty MLP) = 525.

E.5 Multi-class PRR Construction

For C-class tasks, the binary positive/negative evidence split of the main text's decision readout generalizes to equal-width bins of the sentiment range: class $c \in \{0, \ldots, C-1\}$ (most negative to most positive) integrates

$$z_c = \sum_{i:\, s_i \in B_c} \bar{V}_i\, a_i, \qquad B_c = \big[-1 + \tfrac{2c}{C},\, -1 + \tfrac{2(c+1)}{C}\big), \tag{37}$$

where the last bin is closed at 1.0. The auxiliary pooled logit is added exactly as in the binary case: $\mathbf{z} = [z_0, \ldots, z_{C-1}] + \lambda_{\text{aux}} \mathbf{W}_{\text{out}}\, \text{Dropout}\big(\sum_{i=1}^{L} \mathbf{V}_{\text{final}}[:, i, :]\big)$ with λ_{aux}=0.1. For binary tasks (C=2) the model uses the binary PRR of the main text (with neutral tokens split equally between the two channels) rather than this bin construction.

E.6 Sentence-Pair Comparison Branch

For inputs separated into premise and hypothesis spans, let $\mathbf{p}_a$ and $\mathbf{p}_b$ be their sum-pooled firing-rate vectors. The implemented comparison feature is

$$\mathbf{r}_{\text{pair}} = \mathbf{W}_{\text{pair}}[\mathbf{p}_a; \mathbf{p}_b; |\mathbf{p}_a - \mathbf{p}_b|; \mathbf{p}_a \odot \mathbf{p}_b] + \mathbf{b}_{\text{pair}}. \tag{38}$$

This feature augments the pooled representation used by the auxiliary classifier. The token-level PRR branch still receives the original sequence firing rates, so the final logits combine routed token evidence with pooled cross-sentence interactions.

F Algorithmic Summary of the Forward Pass

Algorithm 1 states the computation of Section 3 of the main text end to end in the notation used there, making explicit the order in which quantities are computed and which of them are input-conditioned. It describes the default configuration; optional modules are marked as such and are formalized in Appendix E.

Algorithm 1 Forward pass of Ne-CCNN for one batch. Comments give the corresponding equation of the main text. Optional modules (Neural Resonance, Learned Time-Constant coupling, Negation Scope) are marked where they enter and are omitted from the body for brevity.

Require: token indices $x \in \{0, \dots, V-1\}^{B \times L}$; statistics $\mathbf{f} \in \mathbb{R}^{B \times 12}$; sequence mask $\mathbf{m}_{\text{seq}} \in \{0,1\}^{B \times L}$
Ensure: logits $\mathbf{z} \in \mathbb{R}^{B \times C}$

$\mathbf{E}_{\text{static}} \leftarrow \mathbf{E}[x]$ ▷ Eq. (1)
$\mathbf{E}_{\text{ctx}} \leftarrow \mathbf{E}_{\text{static}} \odot \frac{1+\eta^{+}\mathbf{g}^{+}}{1+\eta^{-}\mathbf{g}^{-}} + \eta^{\text{b}}\mathbf{b}$ ▷ Eq. (3); fixed scales η^{+}=1.5, η^{-}=0.8, η^{b}=0.3
$\mathbf{E}_{\text{pos}} \leftarrow \mathbf{E}_{\text{ctx}} + \text{PE} + \mathbf{B}_{\text{pos}}$
$s_i \leftarrow \tanh\big(\cos(\mathbf{e}_i, \mathbf{p}_{\text{pos}}) - \cos(\mathbf{e}_i, \mathbf{p}_{\text{neg}})\big)$ ▷ Eq. (4)
$s_i \leftarrow -s_i$ within a negation scope ▷ optional
$\mathbf{h}_{\text{neuro}} \leftarrow \text{ReLU}\big(\mathbf{W}_{\text{nm}}^{(1)}\,\text{LN}(\mathbf{f})\big)$ ▷ shared encoder
$\delta\theta^{+} \leftarrow \mathbf{W}_{\text{DA}}\mathbf{h}_{\text{neuro}};\ \delta\theta^{-} \leftarrow \mathbf{W}_{\text{NE}}\mathbf{h}_{\text{neuro}};\ \delta\beta \leftarrow \mathbf{W}_{\text{ACh}}\mathbf{h}_{\text{neuro}}$ ▷ Eq. (8)
$\mathbf{S} \leftarrow \text{Dropout}(\mathbf{W}_{\text{in}}\mathbf{E}_{\text{pos}})$ ▷ Eq. (5)
$\mathbf{S} \leftarrow \mathbf{S} + \tanh(\lambda)\,\mathbf{s} \odot \mathbf{S}$
$\mathbf{A} \leftarrow$ top-k mask $\odot$ polarity gate $\odot$ edge gate on $\cos(\mathbf{E}_{\text{pos}}, \mathbf{E}_{\text{pos}})$, L1 row-normalized ▷ Eqs. (6), (7)
$\mathbf{A} \leftarrow \text{NeuralResonance}(\mathbf{A}, \mathbf{E}_{\text{pos}}, \mathbf{s})$ ▷ optional, App. E.1
$\theta_{\text{pos}} \leftarrow \theta_0 + \frac{1}{B}\sum_b \delta\theta_b^{+};\ \theta_{\text{neg}} \leftarrow \theta_0 + \frac{1}{B}\sum_b \delta\theta_b^{-};\ \beta_{\text{eff}} \leftarrow \beta(1+\delta\beta);\ \mathbf{m}_{\text{mod}} \leftarrow \frac{\gamma}{2}(\delta\theta^{+} + \delta\theta^{-})$
$\mathbf{U} \leftarrow \mathbf{S}$
for $t = 1$ to T **do**
  $\boldsymbol{\theta} \leftarrow \theta_{\text{pos}}\mathbf{m}_{\text{pos}} + \theta_{\text{neg}}\mathbf{m}_{\text{neg}} + \theta_0\mathbf{m}_{\text{neu}};\ \mathbf{V} \leftarrow \sigma(\mathbf{U} - \boldsymbol{\theta}) \odot \mathbf{m}_{\text{seq}}$
  $\mathbf{Z} \leftarrow \mathbf{V}$
  for $n = 1$ to N **do**
    $\mathbf{Z} \leftarrow \text{Dropout}\big(\text{LN}(\mathbf{A}\,\mathbf{W}_{\text{coupled}}^{(n)}\mathbf{Z})\big)$ ▷ Eq. (11)
  end for
  $\mathbf{U} \leftarrow \Pi_{[-10,10]}\big((1-\alpha)\mathbf{U} + \mathbf{S} + \beta_{\text{eff}}\mathbf{Z} + \mathbf{m}_{\text{mod}}\big)$ ▷ Eq. (9)
end for
$\mathbf{V}_{\text{final}} \leftarrow \sigma(\mathbf{U} - \boldsymbol{\theta}) \odot \mathbf{m}_{\text{seq}};\ \bar{\mathbf{V}} \leftarrow \frac{1}{H}\sum_h \mathbf{V}_{\text{final}}[:, :, h]$
$a_i \leftarrow \text{BioAttn}(\mathbf{s}, \mathbf{m}_{\text{seq}}, \bar{\mathbf{V}})_i$ ▷ App. E.4
$z_c \leftarrow \sum_{i \in \mathcal{P}^c} \bar{V}_i a_i + \frac{1}{2}\sum_{i \in \mathcal{P}^0} \bar{V}_i a_i,\ c \in \{\text{pos}, \text{neg}\}$ ▷ Eq. (12); C bins if $C > 2$
$\mathbf{h}_{\text{pool}} \leftarrow \text{Dropout}\big(\sum_i \mathbf{V}_{\text{final}}[i, :]\big)$
for pair inputs, $\mathbf{h}_{\text{pool}} \leftarrow \mathbf{h}_{\text{pool}} + \mathbf{W}_{\text{pair}}[\mathbf{p}_a; \mathbf{p}_b; |\mathbf{p}_a - \mathbf{p}_b|; \mathbf{p}_a \odot \mathbf{p}_b]$ ▷ App. E.6
return $\mathbf{z} = [z_{\text{neg}}, z_{\text{pos}}] + \lambda_{\text{aux}}\mathbf{W}_{\text{out}}\mathbf{h}_{\text{pool}}$

G Decision Trace of a Single Instance

Figure 4 and Table 9 show the complete decision path of one MR test instance, obtained from the headline checkpoint of Table 1 in a single forward pass with no retraining. The instance is not hand-picked for its appearance: among correctly classified test examples whose token length lies in $[10, 18]$, we take the one whose decision margin $|z_{\text{pos}} - z_{\text{neg}}|$ is closest to the median of that subset (266 candidates, median margin 0.164), so the trace shows a typical rather than an extreme case. The

trace is regenerated by one command in the released repository, which also asserts that the routed evidence it records reproduces the logits the model returns.

Three quantities make the path attributable. Each token carries a signed polarity s_i and is routed to the positive or negative evidence channel by that sign; each entry of Table 9 is the product of the token's firing rate, its readout weight, and that channel assignment, so the two channel totals are exactly the sums of the entries above them. The coupling graph in panel (b) is signed and sparse: 116 of the 17×17 possible edges are non-zero and 20 of them are negative, that is, flipped by the polarity gate on emotionally salient pairs of opposite sign. The readout weights are close to uniform with a mild recency tilt, consistent with the position-bias measurements of Appendix O.

One caveat is visible in the same panel, and we report it rather than the routed term alone. The routed evidence accounts for about 5% of this instance's logit margin ($z_{\text{pos}} = 0.027$ against $z_{\text{neg}} = 0.192$, final logits $[1.48, -2.06]$); the pooled auxiliary branch of Eq. (12) supplies the remainder. The trace therefore localizes the polarity-routed part of the decision rather than the whole decision, which is consistent with the census finding that removing the auxiliary branch is the most damaging single removal (Table 3). What the artefact supports is inspectability, not the claim that every logit is carried by the polarity channels.

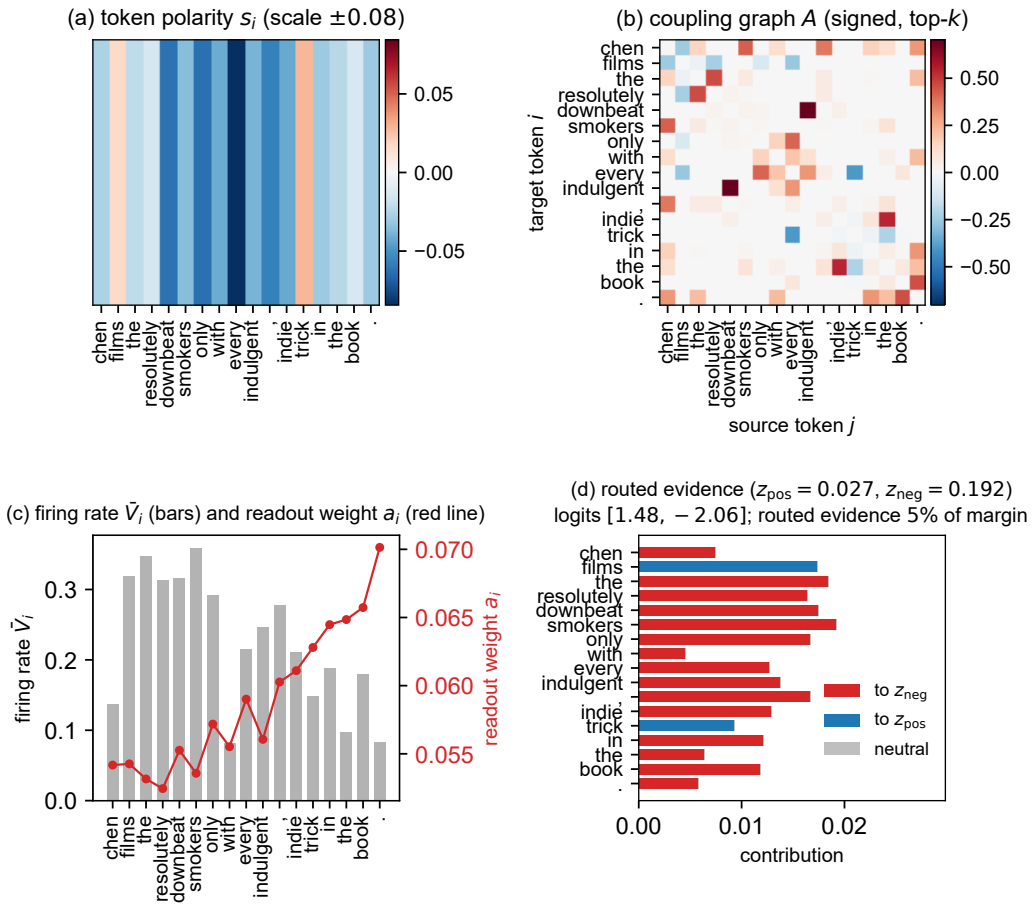

Figure 4: Decision trace of one MR test instance (length 17, labeled and predicted negative). (a) Token polarity s_i; the color scale is symmetric about zero and set to the realized maximum, since the polarity scorer never approaches its nominal ± 1 range on this dataset. (b) The coupling graph used by the recurrence, row-normalized and signed, where blue entries are excitatory and red entries inhibitory; white entries are removed by the top-k mask. (c) Firing rate $\bar{V}_i$ and readout weight a_i. (d) Per-token contributions to the two routed evidence channels, with the logits and the share of the margin they account for.

Table 9: The same instance in tabular form (mr, headline seed 42 checkpoint). s_i: token polarity; $\bar{V}_i$: firing rate; a_i: readout weight; the last two columns are the terms entering the two evidence channels, whose totals are $z_{\text{pos}} = 0.0267$ and $z_{\text{neg}} = 0.1925$. The pooled auxiliary branch adds $+0.0297/-3.543$ of its own to the final logits, so the routed evidence accounts for 5% of the margin.

pos.	token	s_i	$\bar{V}_i$	a_i	to z_{pos}	to z_{neg}
0	*chen*	-0.03	0.137	0.054	$+0.0000$	$+0.0074$
1	*films*	$+0.02$	0.319	0.054	$+0.0173$	$+0.0000$
2	*the*	-0.02	0.347	0.053	$+0.0000$	$+0.0184$
3	*resolutely*	-0.01	0.313	0.052	$+0.0000$	$+0.0164$
4	*downbeat*	-0.06	0.315	0.055	$+0.0000$	$+0.0174$
5	*smokers*	-0.04	0.359	0.054	$+0.0000$	$+0.0192$
6	*only*	-0.06	0.292	0.057	$+0.0000$	$+0.0167$
7	*with*	-0.04	0.082	0.056	$+0.0000$	$+0.0045$
8	*every*	-0.08	0.216	0.059	$+0.0000$	$+0.0127$
9	*indulgent*	-0.04	0.247	0.056	$+0.0000$	$+0.0138$
10	,	-0.06	0.278	0.060	$+0.0000$	$+0.0167$
11	*indie*	-0.04	0.211	0.061	$+0.0000$	$+0.0129$
12	*trick*	$+0.03$	0.149	0.063	$+0.0093$	$+0.0000$
13	*in*	-0.03	0.189	0.064	$+0.0000$	$+0.0122$
14	*the*	-0.02	0.098	0.065	$+0.0000$	$+0.0063$
15	*book*	-0.01	0.180	0.066	$+0.0000$	$+0.0118$
16	.	-0.03	0.083	0.070	$+0.0000$	$+0.0058$
				total	$+0.0267$	$+0.1925$

H Hyperparameters and Training Details

Table 10 reports the hyperparameters and architectural choices for the best Ne-CCNN configuration on each benchmark. Unless otherwise noted, all experiments use AdamW optimization with warmup-cosine scheduling, Hebbian PPMI+SVD embeddings of dimension 300, a k-nearest-neighbor adjacency graph with k=5, the enhanced dynamic-embedding mode, gradient clipping at 5.0, label smoothing with ϵ=0.1, and seed 42. Multi-seed results use the seed sets noted in each table (5 seeds: 42, 123, 456, 789, 1024). For the pre-trained reference rows in the main-text tables, every cell averages the same three runs (seeds 42, 123, 456), so no reference cell mixes seed counts: BERT, SpikeLM, and SpikeBERT are all averaged over that triple. Reference cells that had accumulated further seeds were reduced to the common triple so that every entry in a table shares one convention. The SpikeBERT SST-2 entry reads 78.29 ± 0.70 test accuracy while development accuracy sits at 90.4–91.8, and that gap of 12.7–13.4 points reproduces in all three completed runs (development/test: 91.80/78.44, 90.38/77.52, 91.61/78.90). One run on this benchmark costs 65–83 GPU-hours at the official configuration (40–49 early-stopped epochs), roughly an order of magnitude above a full three-seed sweep of Ne-CCNN on the same task. The split layout explains the direction of such gaps: for the GLUE tasks the development split is a 10% carve of the official training split (SST-2: 60,614 train / 6,735 dev / 872 test), so checkpoint selection is made on an in-distribution subset of the training data, whereas the test split is the official development set; the same pattern appears on RTE (7.6 points) and MRPC (2.7 points). We therefore report these cells as same-protocol references only, and the published SpikeBERT accuracies are not comparable because the official script selects on the test split. Four SpikeBERT cells carry additional notes: CMNLI and QQP were warm-started from the best checkpoints of interrupted earlier runs and then early-stopped normally, MNLI's seed 456 is the test accuracy of its best-validation checkpoint (the run was interrupted during epoch 11; its best epoch was 4 and none of the six epochs logged after it improved on that value, so this equals the number the protocol's single final test would have produced), MRPC_F re-evaluates the saved best-validation checkpoints under macro-F1, and on CoLA every seed collapses to the majority class (MCC=0), the same failure mode as LIF/PSN-BERT (Appendix B).

Training objective. All reported models are trained with cross-entropy loss with label smoothing (ϵ=0.1) on the final logits $\mathbf{z}$ of Stage 5, which already include the auxiliary pooled logit with weight λ_{aux}=0.1. We use AdamW with default momentum parameters (β_1=0.9, β_2=0.999) and a per-step warmup-cosine schedule (warmup of one epoch, two for the IMDB pruning configuration), with global-norm gradient clipping at 5.0. Post-pruning fine-tuning restarts the optimizer at half the learning rate without warmup. Focal and key-region losses are implemented but disabled in all reported configurations.

Table 10: Final training configurations. The accuracy column repeats the five-seed means of Table 1 for the corresponding configuration, under the clean evaluation protocol. "Bio" denotes Affective Terrain, River Gate, and Negation Scope. "Corpus" indicates the source of Hebbian co-occurrence statistics (same = target dataset training split); SUBJ additionally uses a learnable auxiliary weight.

Dataset	Best Acc	H	T	N	Max len	Batch	Epochs	LR	WD	Drop	w	LTC	Corpus	Bio	Prune
IMDB	86.48%	512	8	5→3	512	8	15 + 5 FT	4e−4	1e−5	0.3	5	No	same	Yes	0.4
MR	78.10%	128	3	2	64	32	25	5e−4	5e−5	0.5	3	Yes	same	No	0.0
SUBJ	90.70%	128	3	2	96	24	25	5e−4	5e−5	0.5	3	No	same	No	0.0
SST-2	80.59%	256	5	3	128	16	15	4e−4	5e−5	0.5	3	No	same	No	0.0
SST-5	43.01%	128	3	2	64	32	25	3e−4	5e−5	0.5	3	Yes	same	No	0.0

The IMDB configuration follows a train-then-prune pipeline: an oversized network with H=512, T=8, and N=5 is trained for 15 epochs, the bottom 40% of coupling layers (ranked by L2 norm of their weight matrices) are removed, and the resulting 3-layer network is fine-tuned for one epoch (Table 15). For MR, w=3 and LTC coupling are synergistic: the window reduction helps on its own at w=3, LTC adds a further gain at w=3 whereas it is neutral at w=5, and their combination gives the best MR configuration of Table 1; the heterogeneous temporal dynamics help capture sentiment reversals in medium-length texts. For SUBJ and SST-2, reducing the Hebbian window to w=3 is the dominant improvement, reflecting the importance of local co-occurrence statistics in short texts. For SST-5, w=3 and LTC coupling combined with a reduced learning rate (3×10^{-4}) give the best configuration of Table 1: the smaller step size turns noisy single-epoch peaks into a stable mid-training plateau.

Reproducibility. The implementation retains switches that were introduced while the architecture was under development; all of them are disabled in every configuration reported here, and the components that are enabled for each dataset are the ones listed in Table 10. Every reported number is produced by a single command line or probe script, and the released repository indexes, for each table entry, the command that generates it.

I Extended Ablation Studies

This appendix collects the complete ablation matrices underlying the quantitative claims made in the main paper. All deltas are computed against the corresponding baseline configuration reported in the same row group.

I.1 Hebbian Co-occurrence Window

Table 11 varies the co-occurrence window w on the short-text datasets. Reducing w from the default 5 to 3 yields the largest accuracy gains; w=2 is either not tested or degrades performance, indicating an optimal local context granularity.

Table 11: Hebbian co-occurrence window ablation. PPMI-NNZ = number of non-zero entries in the PPMI matrix.

Dataset	w	PPMI-NNZ	Accuracy
SST-2	5	393,443	81.08%
	3	271,996	**82.34%**
	2	194,051	82.00%
SUBJ	5	549,422	90.25%
	3	377,129	**91.50%**
MR	5	518,393	77.77%
	3	362,206	**77.95%**

Table 12: LTC coupling effect across datasets. Δ is versus the identical configuration without LTC.

Dataset	Configuration	LTC	Accuracy	Δ
MR	same-corpus 128/3/2	No	77.77%	—
		Yes	**77.95%**	+0.18%
IMDB	terrain+river+negscope	No	87.36%	—
		Yes	87.07%	−0.29%
SST-2	same-corpus 256/5/3, w=3	No	**82.34%**	—
		Yes	81.31%	−1.03%
SUBJ	baseline 128/3/2	No	**91.40%**	—
		Yes	89.75%	−1.65%

Table 13: Time-step ablation on IMDB (256/T/3, Hebbian, lr=4×10^{-4}, no bio components).

T	Accuracy	Peak Epoch
3	**87.32%**	E3
5	87.21%	E3
8	87.25%	E3

I.2 Learned Time-Constant Coupling

Table 12 reports the effect of LTC coupling across datasets and configurations. LTC is beneficial on MR, where heterogeneous temporal integration helps capture sentiment reversals, but it is detrimental on SUBJ and SST-2, where the default temporal dynamics already suffice. On SST-5 it is accuracy-neutral (4-seed mean 43.01% versus 43.00% without LTC) but reduces seed variance (σ 0.58%→0.31%), and is therefore retained in the final SST-5 configuration.

I.3 Sentence-Pair Readout Isolation

The sentence-pair results (Table 2) use a comparison branch $W_{\text{pair}}[\mathbf{p}_a; \mathbf{p}_b; |\mathbf{p}_a - \mathbf{p}_b|; \mathbf{p}_a \odot \mathbf{p}_b]$. To isolate that branch we compare it with its matched control ($A0$: the identical recipe with the pair readout removed) on CMNLI, the pair task for which both sides have complete five-seed coverage. The branch is worth $+7.20$ pp (53.33 versus 46.13; per-seed gains $+7.73$, $+6.60$, $+8.12$, $+7.07$, $+6.47$; all five positive; paired $t = 22.5$), so the pair-task gains are carried by this branch rather than by the pooled representation alone. We do not claim the branch as a novel component: it is the standard comparison-feature combination, and the contribution we measure is its effect inside Ne-CCNN. A second matched sweep re-runs both arms from scratch on the English pair task RTE (five seeds, fresh environment): the branch is worth $+2.02$ pp there (51.99 versus 49.96; four of five seeds positive, not significant), which agrees with the marginal RTE effect already reported, and the with-branch mean reproduces the main-run RTE value almost exactly (51.99 versus 51.98), cross-checking the environment. Single-seed controls on QNLI and QQP point the same way ($+9.6$ and $+1.6$ pp).

I.4 Time-Step Efficiency

Table 13 shows that the CCNN membrane EMA reaches steady state with few recurrent steps on IMDB. Accuracy is nearly flat across $T \in \{3, 5, 8\}$, confirming that T=3 is sufficient even for 512-token documents.

I.5 Embedding Source

Table 14 quantifies how much performance depends on Hebbian pre-training versus the CCNN dynamics themselves. Random Xavier initialization recovers 98.1% of Hebbian performance, but the bio-component pipeline only helps when embeddings carry meaningful semantic structure.

Table 14: Embedding-source ablation on IMDB. "+bio+prune" adds Affective Terrain, River Gate, Negation Scope, and synaptic pruning.

Embedding	Baseline	+bio+prune	Δ bio
Hebbian PPMI+SVD	87.21%	**87.55%**	+0.34%
Random Xavier	85.51%	85.56%	+0.05%
Δ (Hebbian − Random)	+1.70%	+1.99%	

Table 15: Synaptic pruning pipeline on IMDB. Upper: extended training (15 epochs, seed 42). Lower: standard training (8 epochs, 4 seeds mean ± std). PP = pre-prune. FT-k = post-prune fine-tuned for k epochs.

Stage	Layers	Params	Accuracy	Notes
Extended training (15 epochs, seed 42)				
PP	5	6.86M	87.03%	peak E3; overfit by E15
FT-1	3	6.34M	**87.55%**	best; +0.52% over PP
FT-5	3	6.34M	86.82%	re-overfitting
Base	3	5.59M	87.36%	hand-tuned 3-layer
Standard training (8 epochs, 4 seeds)				
PP	5	6.86M	$87.32 \pm 0.11\%$	peak E3–E4; mild overfit
FT-3	3	6.34M	$87.32 \pm 0.11\%$	Δ ≈0% (no overgrowth)

Table 16: Cross-corpus Hebbian augmentation. "Same" means embeddings built only from the target dataset's training split.

Dataset	Corpus	Accuracy	Same-corpus baseline	Δ
MR	imdb+mr	77.95%	77.77%	+0.18%
SUBJ	imdb+subj	90.85%	91.40%	−0.55%
SST-2	imdb+sst2	81.77%	81.08%	+0.69%

I.6 Synaptic Pruning

Table 15 details the train-then-prune pipeline. Under extended training (15 epochs, seed 42), the overgrown 5-layer network is pruned by 40% and fine-tuned, recovering +0.52% via overgrowth-elimination. Under standard training (8 epochs, 4 seeds 123/456/789/1024), the pre-prune model attains $87.32 \pm 0.11\%$ without overgrowth: pruning is neutral (Δ ≈0%). Both regimes are reported for completeness: the 15-epoch result demonstrates the pruning principle, while the 8-epoch multi-seed result provides the statistical benchmark.

I.7 Cross-Corpus Hebbian Augmentation

Table 16 shows that cross-corpus augmentation is domain-sensitive. It helps on MR (movie reviews, domain-adjacent to IMDB) and SST-2, but hurts on SUBJ because IMDB reviews are inherently subjective and dilute the objective/subjective distinction. Since all baselines build embeddings from the target corpus only, cross-corpus variants are excluded from every reported main result and appear here solely as analysis.

I.8 Optional Bio Components

Table 17 confirms that Affective Terrain, River Gate, and Negation Scope are only weakly beneficial on large-data IMDB and become detrimental on datasets with fewer than 10K training examples, validating their default-off design for small corpora.

Table 17: Effect of optional bio components. IMDB adds Affective Terrain + River Gate + Negation Scope. MR and SUBJ add Affective Terrain + River Gate.

Dataset	Train size	Baseline	+Bio modules
IMDB	25,000	87.21%	87.36%
MR	8,530	77.77%	77.11%
SUBJ	8,000	91.40%	91.25%

Table 18: Skip-attention ablation. A firing-rate EMA is fused into the PRR decision as $\mathbf{V}_{\text{decision}} = \mathbf{V}_{\text{final}} + \mathbf{V}_{\text{ema}}$.

Dataset	Baseline	+Skip-attention	Δ
IMDB	87.36%	87.09%	−0.27%
SST-2	82.34%	82.45%	+0.11%

I.9 SKIP-ATTENTION

Table 18 shows that explicitly fusing a firing-rate EMA into the PRR decision layer is either neutral or harmful, because the membrane potential $\mathbf{U}$ already implements sufficient temporal memory.

J MULTI-SEED STABILITY

All main-text results are means over five seeds (42, 123, 456, 789, 1024) with the per-seed values recorded in the project's experiment log; they are not duplicated here. Standard deviations are tightest on the large IMDB dataset ($\sigma = 0.35\%$) and wider on the smaller benchmarks (MR $\sigma = 0.43\%$, SUBJ $\sigma = 0.49\%$, SST-5 $\sigma = 0.65\%$, SST-2 $\sigma = 0.85\%$), consistent with the expected inverse relationship between test-set size and variance. Peak-epoch consistency (Table 19) reflects the contractive dynamics, which reach their best validation epoch within a narrow window across seeds.

Table 19 compares peak-epoch spreads across architectures on MR and SUBJ (5 seeds). The CCNN's contractive dynamics yield the most stable convergence (see the main text, Discussion).

Table 19: Peak-epoch spread (range of best-test epoch across 5 seeds) on MR and SUBJ. Lower spread = more stable convergence.

Model	MR	SUBJ	Avg
Ne-CCNN	8	5	6.5
LIF-SNN	12	6	9.0
TextCNN	15	9	12.0
BiLSTM	4	21	12.5
SpikingTextCNN	19	15	17.0

K INTRINSIC ROBUSTNESS

Following the SpikingTextCNN protocol Lv et al. (2023), we evaluate n=1000 test samples per dataset under TextFooler and PWWS word-substitution attacks, without adversarial training. Ne-CCNN is more robust than TextCNN on all four datasets. Relative to BiLSTM, it is more robust on SUBJ and ChnSenti, approximately tied on MR, and less robust on SST-2, mirroring the clean-accuracy pattern. The full BERT-Attack battery gives the same qualitative ordering on the three English datasets. Under the strongest SUBJ attack, removing neuromodulation increases mean attack success by 1.47 points, and removing coupling is also detrimental.

Table 20: Intrinsic robustness under word-substitution attacks, reported here in full: mean attack success rate (%, lower is better) over TextFooler and PWWS, without adversarial training. Bold: best per dataset. The ChnSenti column uses a custom Chinese character-level noise attack (deletion and character substitution at several rates) instead of TextFooler/PWWS, since TextAttack does not support Chinese and its substitution operator requires Chinese word vectors.

Model	MR	SUBJ	SST-2	ChnSenti
Ne-CCNN	89.86	**85.05**	83.03	**3.59**
BiLSTM	**89.05**	86.03	**78.32**	4.44
TextCNN Kim (2014)	90.28	87.71	87.85	5.81

L Controlled Probe Results

Tables 21, 22, and 23 report the full results of the semantic-satiation, neuromodulatory-margin, and negation-propagation probes described in the main text (Cognitive-Neuroscience Correspondences), summarized in Fig. 5. These probes follow an established program in which internal model quantities are compared against predictions from cognitive science, a program with precedent in both directions: shared meaning abstraction, rather than next-word prediction, has been shown to drive the alignment between language models and measured brain responses Cheng et al. (2026), and neuromodulatory principles have been used to improve the adaptivity and continual learning of artificial networks Mei et al. (2025). We report the positive and null outcomes of our own probes with equal prominence.

Provenance and multi-seed recomputation. An earlier single-seed version of these probes (seed 42) loaded checkpoints with the model's default *original* gating branch rather than the *enhanced* branch used during training; because both gating branches are constructed for every model and the loader was non-strict, this mismatch was silent, and the earlier numbers therefore measured an untrained gating pathway. All results below were recomputed on the trained enhanced branch over five SUBJ seeds plus the MR and IMDB seed-42 head checkpoints, and supersede the earlier values. The probes of Appendix O are unaffected: their loading chain already selected the trained branch.

Table 21: Semantic-satiation probe on the trained enhanced branch: change in target-word firing rate $\Delta V = V(T{=}8) - V(T{=}1)$, and the number of target words (of *good*, *bad*, *film*) whose $V(T)$ is non-increasing over $T{=}1{\to}8$. MR and IMDB use the seed-42 head checkpoints.

Checkpoint	Mono./3	ΔV *good*	ΔV *bad*	ΔV *film*
SUBJ seed 42	3/3	-0.130	-0.117	-0.156
SUBJ seed 123	2/3	-0.060	-0.031	-0.071
SUBJ seed 456	3/3	-0.013	-0.020	-0.036
SUBJ seed 789	3/3	-0.035	-0.013	-0.054
SUBJ seed 1024	3/3	-0.140	-0.081	-0.161
SUBJ (5 seeds)	14/15	$\Delta V = -0.075 \pm 0.053$ over all word–seed pairs		
MR seed 42	1/3	$+0.001$	$+0.002$	-0.044
IMDB seed 42	3/3	-0.029	-0.021	-0.017

M Readout Ordering and Modulation Contribution

Both experiments in this section are computed on the headline five-seed checkpoints used by the main comparison, over the *full* test split of each dataset. Neither retrains anything: the ordering probe only reads the polarity-routed evidence that the model's own decision function computes (the wrapper re-derives it from the same inputs and asserts numerically that its reconstruction of the logits matches the returned logits on every batch, aborting otherwise), and the modulation probe reverses the model's own inference-time switches.

A. Does the readout order inputs by human affective intensity? Table 24 reports two tests. On SST-5, whose labels are five human intensity levels, the model's graded output tracks the hu-

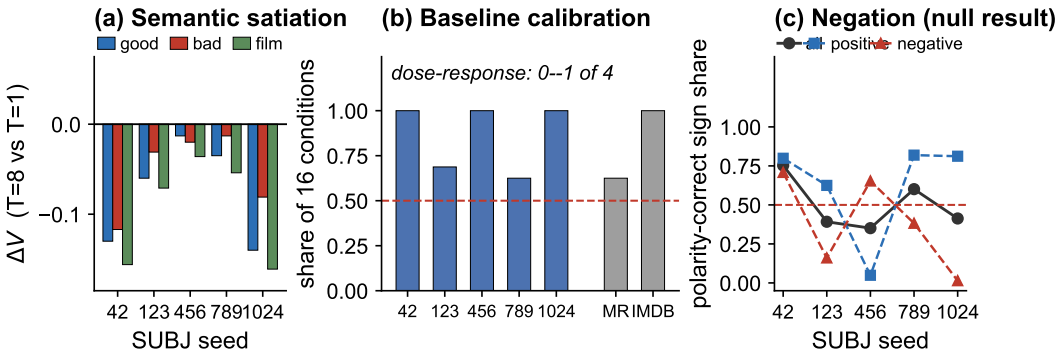

Figure 5: Controlled probes on the trained enhanced branch (five SUBJ seeds, plus MR and IMDB seed-42 head checkpoints as controls; dashed red lines mark chance). (a) Semantic satiation: change in target firing rate between $T{=}1$ and $T{=}8$ for three probe words. All 15 SUBJ word–seed pairs are non-positive except one ($14/15$ monotone), with a mean of $\Delta V = -0.075$; the IMDB control reproduces the direction on all three words ($3/3$) while the MR control does not ($1/3$). The effect is a direct consequence of the leaky membrane approaching its asymptote and carries no repetition-specific machinery. (b) Neuromodulatory baseline calibration: share of the 16 probe conditions (2 targets $\times$ 2 context polarities $\times$ 4 emotional densities) in which enabling neuromodulation suppresses target activation. The suppression is broad (0.86 ± 0.19 of conditions across SUBJ seeds; $16/16$ on IMDB) but a monotone dose–response in emotional density is absent (0–1 of the 4 density groups on every checkpoint). (c) Negation probe: share of cells whose induced change has the polarity-correct sign. Accuracy is at chance on average across SUBJ seeds and splits by target polarity in a seed-dependent way; magnitude shows no reliable decay with negator–target distance, so we make no claim of scope-free negation handling. These panels report mechanism-consistency checks, not claims of neural equivalence.

Table 22: Neuromodulatory baseline-calibration probe on the trained enhanced branch. Of 16 conditions (2 targets $\times$ 2 context polarities $\times$ 4 emotional densities), the number in which enabling neuromodulation suppresses target activation ($\Delta V_{\text{neuro}} < 0$), and the number of density groups (of 4) in which $|\Delta V|$ grows monotonically with emotional density. The suppression effect is broad but a monotone dose–response is not supported.

Checkpoint	**Suppressed /16**	**Dose–response /4**
SUBJ seed 42	16/16	1/4
SUBJ seed 123	11/16	0/4
SUBJ seed 456	16/16	0/4
SUBJ seed 789	10/16	0/4
SUBJ seed 1024	16/16	0/4
SUBJ (5 seeds)	0.863 ± 0.190	—
MR seed 42	10/16	0/4
IMDB seed 42	16/16	1/4

man label closely: the expected grade $\sum_c c\,\mathrm{softmax}(\mathbf{z})_c$ correlates with the label at Spearman $\rho = 0.613 \pm 0.003$ across seeds ($n{=}2210$ examples, $p < 10^{-220}$ in every seed), ordinal MAE is 0.82 on a 0–4 scale, and the per-label grade means increase monotonically in all five seeds (e.g., $1.38, 1.56, 1.97, 2.50, \ldots$ for seed 42). On the binary benchmarks, the polarity-route evidence correlates with a human norm-based sentence valence (mean Warriner V over lexicon-covered tokens) at $\rho = 0.448{\pm}0.005$ (IMDB, $n{=}25{,}000$), $0.400{\pm}0.008$ (SST-2), $0.328{\pm}0.007$ (MR) and $0.155{\pm}0.005$ (SUBJ), each seed significant at $p < 10^{-11}$. Two caveats are reported rather than hidden. First, the human valence proxy covers only 26–27% of tokens, and it ignores composition, so it is a noisy reference, not a gold polarity label. Second, on SST-5 the *auxiliary-free* binned evidence alone does not by itself reach the human label (argmax accuracy 0.176 against 0.20 chance); it is the full readout, in which the auxiliary pooled branch cooperates with the polarity-routed evidence, that is ordinally competent. This is consistent with the census finding that the auxiliary branch is the single most

Table 23: Expanded negation probe on the trained enhanced branch (12 targets $\times$ 3 negators $\times$ distances d=1..8). Sign: fraction of word$\times$negator$\times$distance cells whose $\Delta V = V_{\text{neg}} - V_{\text{base}}$ has the polarity-correct sign (pos/neg: within positive/negative targets). ρ<0: words whose $|\Delta V|$ decreases with distance (Spearman); sig.: p<0.05. Pooled: Spearman ρ of the word-averaged curve. No checkpoint shows significant decay in a majority of words; sign accuracy is near chance on SUBJ (0.502 ± 0.170 across five seeds) and splits by polarity in a seed-dependent manner. Four words (*bad*, *terrible*, *awful*, *perfect*) show ρ<0 on all five SUBJ seeds; the single significant pooled decay (seed 42) is driven by a d=2 peak rather than a clean monotone decay, and the MR pooled trend is positive.

Checkpoint	Sign	Sign (pos)	Sign (neg)	ρ<0	sig.	Pooled ρ (p)
SUBJ seed 42	0.753	0.799	0.708	9/12	2/12	−0.833 (0.010)
SUBJ seed 123	0.392	0.625	0.160	9/12	4/12	−0.476 (0.233)
SUBJ seed 456	0.351	0.049	0.653	9/12	0/12	−0.214 (0.610)
SUBJ seed 789	0.601	0.819	0.382	7/12	0/12	+0.429 (0.289)
SUBJ seed 1024	0.413	0.812	0.014	9/12	2/12	−0.452 (0.260)
MR seed 42	0.622	0.257	0.986	3/12	0/12	+0.690 (0.058)

damaging removal (§4.3) and it argues against interpreting the polarity route as a self-sufficient intensity estimator.

B. What does neuromodulation contribute at inference time? Table 25 zeroes the modulatory signals of the trained checkpoint through the model's own switches and re-evaluates on the identical examples. On MR the contribution is unambiguous and factor-specific: removing dopamine costs 1.76 ± 0.38 points (5/5 seeds positive), while removing norepinephrine ($+0.19$) or acetylcholine (-0.08) has no measurable effect; removing all three costs 0.90 ± 0.53. This independently reproduces, at inference time and with the same weights, the census result that DA carries MR among the three factors. On SST-2, SST-5 and SUBJ all factors are within noise of zero, again matching the census picture in which modulation is load-bearing on MR and Chinese sentiment rather than on every benchmark; the one further signal is a weakly positive norepinephrine effect on SUBJ ($+0.40 \pm 0.68$, positive in 4 of 5 seeds). Finally, binning MR test examples by model-scored emotional-word density, the modulation benefit is largest in the lowest-density quartile ($+1.44$ points, McNemar 42 versus 23 discordant pairs, $p \approx 0.02$) and smallest in the second and fourth quartiles, which is the ordering predicted if modulation substitutes for affective context that the input does not itself supply; the same binning is uninformative on SST-5 because 75% of its test sentences contain no token above the salience threshold.

Table 24: Readout ordering. SST-5 rows report Spearman correlation of the graded output with the human 5-level label and ordinal MAE, over the full test split (2210 examples); binary rows report Spearman correlation of the polarity-route evidence with a human norm-based sentence valence (Warriner), over the full test splits of IMDB (25,000), SST-2 (872), MR (1066) and SUBJ (2000) examples. All entries are means over five seeds $\pm$ standard deviation; every individual seed is significant at $p < 10^{-11}$. Checkpoints are the five-seed configurations of Table 1 (the SUBJ entry uses the w=3 variant, whose measured accuracy is 90.24%, rather than the learnable-aux headline configuration).

Dataset	Quantity	Statistic	Value (5 seeds)
SST-5	graded output against human labels	Spearman ρ	0.613 ± 0.003
SST-5	graded output against human labels	ordinal MAE	0.817 ± 0.006
SST-5	argmax against human labels	Spearman ρ	0.576 ± 0.003
SST-5	aux-free evidence argmax	accuracy (chance 0.20)	0.176
IMDB	polarity evidence against human valence	Spearman ρ	0.448 ± 0.005
SST-2	polarity evidence against human valence	Spearman ρ	0.400 ± 0.008
MR	polarity evidence against human valence	Spearman ρ	0.328 ± 0.007
SUBJ	polarity evidence against human valence	Spearman ρ	0.155 ± 0.005

Table 25: Inference-time modulation contribution: accuracy drop caused by zeroing each modulatory channel of the trained checkpoint through the model's own switches (positive = modulation helps), over the identical test examples and the five seeds. Values are percentage points, mean ± standard deviation; the number of seeds with a positive contribution is in parentheses.

Dataset	All off	DA off	NE off	ACh off
MR	$+0.90 \pm 0.53$ (5/5)	$+1.76 \pm 0.38$ (5/5)	$+0.19 \pm 0.36$ (4/5)	-0.08 ± 0.32 (2/5)
SST-2	-0.14 ± 0.74 (2/5)	-0.48 ± 0.73 (1/5)	$+0.14 \pm 0.60$ (3/5)	$+0.05 \pm 0.54$ (2/5)
SST-5	$+0.31 \pm 0.91$ (3/5)	$+0.02 \pm 0.13$ (2/5)	$+0.50 \pm 0.48$ (4/5)	-0.01 ± 1.03 (1/5)
SUBJ	-0.35 ± 0.48 (2/5)	-0.29 ± 0.47 (1/5)	$+0.40 \pm 0.68$ (4/5)	-0.09 ± 0.47 (2/5)

N TRAINING-FREE READOUT AND UNTRAINED BACKBONES

This appendix asks how much of the pipeline must be learned. None of the experiments retrain anything: each one loads a headline checkpoint, then either replaces the learned readout with a rule that contains no free parameters, re-initializes a group of parameters in place, or removes the coupled dynamics altogether. All numbers are computed on the full test split of each dataset, and prototype features are drawn class-balanced with a fixed seed, because several of these splits are ordered by label.

Readout ladder. Table 26 replaces the Stage 5 decision with four rules that contain no learned parameters. The rules read only quantities the network already produces at inference: the per-token sentiment estimate $\hat{s}_t$, the membrane magnitude $|U_t|$, and a three-token negation window σ_t. The pure-lexicon row uses no network at all and shows how much of the task the hand lexicon carries on its own.

Table 26: Accuracy with a parameter-free readout substituted for the learned Stage 5 decision, full test splits, headline checkpoints (seed 42). $s(w_t)$ is the hand lexicon, $|U_t|$ the membrane magnitude, σ_t the negation window, and $\hat{s}_t$ the model's own per-token sentiment estimate. IMDB is the $T{=}8$, width 512, prune-40 configuration.

Dataset	n	Learned readout	$\sum_t s(w_t)$	$\times \lvert U_t \rvert$	$\times \sigma_t$	$\times \hat{s}_t$
IMDB	25,000	**86.66**	70.03	66.22	66.19	50.01
MR	1,066	**77.77**	54.03	54.22	53.94	74.11
SST-2	872	**81.31**	54.82	54.82	54.70	67.66
SUBJ	2,000	**89.70**	51.40	51.60	51.75	88.85

Three readings follow. First, on short text a parameter-free rule recovers nearly the whole of the learned readout (MR -3.7 points, SUBJ -0.9), so the trained decision layer is a thin mapping over evidence the network already holds, while on SST-2 the gap widens to 13.7 points. Second, multiplying by the membrane magnitude changes nothing on any dataset (± 0.2 points against the lexicon row), so token magnitude is not a lexicon detector. Third, the rule does not transfer to long documents: on IMDB it falls to chance while the lexicon row alone reaches 70.03. We report the last point as a limitation of the aggregation rather than as evidence about the model.

Which parameter groups carry the signal. Table 27 re-initializes parameter groups in place and repeats the two extreme readouts. The coupled dynamics remain input-driven in every row, because the coupling graph is built from the Hebbian PPMI–SVD embedding, which is corpus-level preprocessing rather than supervision.

Replacing the readout collapses the learned decision and leaves the parameter-free row numerically unchanged, which confirms that this row never reads the decision layer. On SST-2 the coupling parameters are dispensable at inference: their re-initialization changes neither readout, consistent with the census arms in which coupling is neutral on that benchmark. Removing every non-embedding parameter collapses both readouts, so the mechanism is not usable without training on this benchmark. On MR the same condition retains 66.6 against the lexicon row's 54.0 while the trained model reaches 74.1: part of the per-token sentiment estimate survives a randomised backbone because it is

Table 27: Parameter groups re-initialized in place on the SST-2 headline checkpoint, full test split (n=872). The last row removes every non-embedding parameter, which is the closest text analogue of a training-free image CCNN.

Re-initialized group	Learned readout	Parameter-free ($\times \hat{s}_t$)
none (trained checkpoint)	**81.31**	**67.66**
Stage 5 readout only	51.15	67.66
lateral coupling only	81.42	67.55
all non-embedding parameters	50.00	52.18

computed from the embedding geometry, but the remainder, which is what brings the rule close to the learned readout, is placed there by training.

Text laid out as a 1D image. A token sequence with embedding-valued tokens is an image with one spatial axis, and the token coupling graph is then an image coupling graph, which is the construction used throughout this work. Table 28 tests that construction with nothing trained: only the Hebbian embedding is kept, every other parameter is re-initialized, and a nearest-centroid classifier is fitted on labeled prototypes drawn from the training split. The lexicon-alignment column reports the AUC of $|U_t|$ at predicting whether a token belongs to the hand lexicon, a label-free measure of whether saliency lands on semantically loaded tokens.

Table 28: Classification and saliency alignment with an untrained backbone, full test splits. Features are mean vectors of the Hebbian embedding (no coupled dynamics), of U, or of V, evaluated with n_{lab}=20 prototypes per class. Lexicon AUC uses $|U_t|$ as the score.

Dataset	Backbone	Embedding	U	V	Lexicon AUC
MR	trained	61.56	**77.44**	**77.11**	0.397
MR	untrained	61.56	55.78	56.22	0.669
SST-2	trained	49.31	**62.27**	**81.77**	0.454
SST-2	untrained	49.31	61.35	63.19	0.506

Two conclusions follow. First, the benefit is small. Embeddings with labeled prototypes stay at chance on SST-2 and reach only 61.6 on MR, and the untrained dynamics add little, whereas the trained states reach 77.4 and 81.8. Second, training changes where saliency lands: on MR the association between $|U_t|$ and lexicon tokens reverses from positive in the untrained network to below chance in the trained one, so training moves the network away from spotlighting surface sentiment words and toward distributing the evidence. On SST-2 the same reversal is only marginal. This is the behavior that the ordering analysis above reports from the readout side, and it argues against reading per-token magnitude as a lexicon detector.

O Additional Cognitive Probes

This appendix reports three preregistered probes (predictions were written down before any run, and all executed conditions are reported) targeting components not covered by the probes in the main text: the learned position bias of the readout, the selectivity of the resonance module, and the valence geometry of the Hebbian embedding space. Unless stated otherwise, statistics use 10,000 permutation or bootstrap samples, with Holm correction within each family of comparisons. Figure 6 summarises the two results that survive their controls most robustly; the remaining comparisons are reported with their limitations in place.

O.1 Serial-Position Weighting

Method. The readout's position bias (512 learned scalars passed through a sigmoid) is initialized as a shallow U; we ask whether training preserves, amplifies, or erases this shape, and whether a *functional analogue* of the human serial-position curve Murdock (1962); Glanzer & Cunitz (1966) emerges in behavior. (i) *Structural*: the U-index (mean sigmoid bias at the first and last 10% of positions minus the middle 20%) is compared before and after training on all five checkpoints. (ii) *Be-*

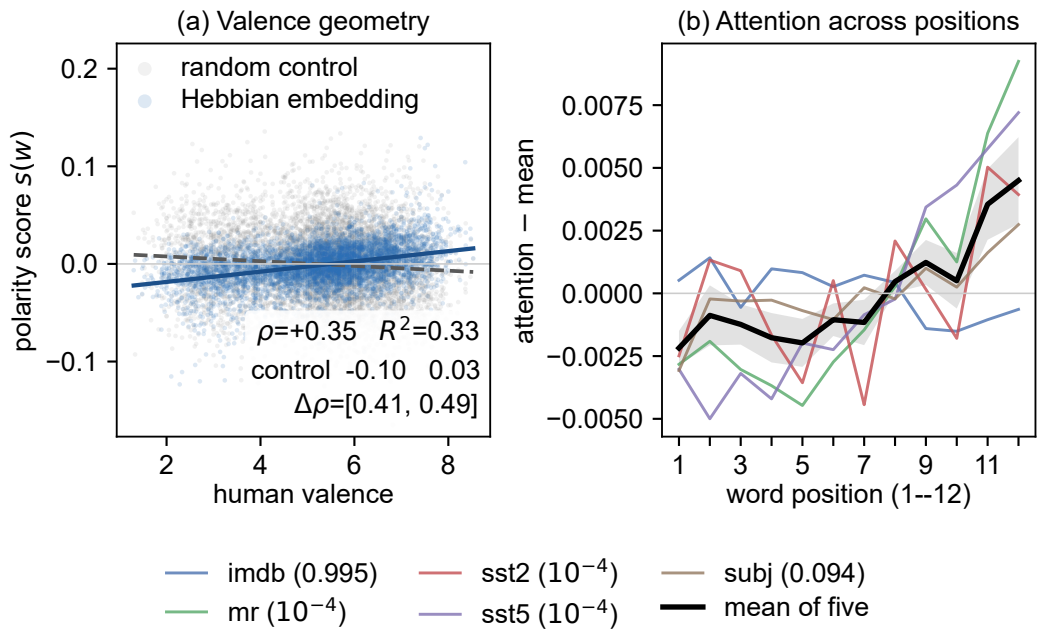

Figure 6: The two best-supported cognitive probes, drawn from the runs reported in this appendix. (a) Model polarity score $s(w)$ against human valence Warriner et al. (2013) for the n=5,782 shared words, for the Hebbian embedding and for the random-embedding control; lines are ordinary least-squares fits and the annotated statistics are Spearman ρ and 5-fold cross-validated ridge R^2, given for the Hebbian embedding and for the control in that order (difference in ρ: $+0.446$, bootstrap 95% CI $[0.407, 0.486]$). (b) Readout attention as a single emotional word is swept across the 12 positions of a neutral template, one curve per checkpoint; the value in parentheses is that checkpoint's paired-permutation p-value for the endpoint advantage; each curve is centered on its own mean so that the endpoint structure is comparable across checkpoints, and the ribbon is ± 1 standard error across the five checkpoints. The endpoint advantage is significant on MR, SST-2 and SST-5 (p=10^{-4}), marginal on SUBJ (p=0.094), and absent on IMDB (p=0.99); we report the negative checkpoint alongside the positive ones rather than discarding it.

havioral: a single emotional target (6 words) is placed at each position p of a length-12 sentence of neutral fillers (3 filler sets; 216 runs per model), and the readout attention $\mathrm{att}(p)$ is recorded. We test the endpoint advantage $\mathrm{mean}(\mathrm{att}(0), \mathrm{att}(11)) - \mathrm{mean}(\mathrm{att}(4..7))$ with a paired permutation test (pairing unit: target $\times$ filler set), and the quadratic coefficient a of the normalized curve.

Results. Training preserves but does not amplify the U-shape: the trained U-index stays within ± 0.004 of the initialization (0.0601) on all five checkpoints (Table 29). Behaviorally, the preregistered endpoint advantage is confirmed on MR ($+0.0061$, p=10^{-4}, a= $+\,0.022$), SST-2 ($+0.0032$, p=10^{-4}, a= $+\,0.013$), and SST-5 ($+0.0038$, p=10^{-4}, a= $+\,0.013$), with the MR curve recency-dominated. SUBJ shows a marginal trend (p=0.094), and IMDB is null (p=0.99). The position bias thus yields a dataset-dependent, recency-weighted readout on short texts. We claim only a functional analogue of the serial-position curve, not a model of human memory storage.

O.2 Resonance Synchronization Selectivity

Method. The resonance module is inspired by synchronization-based feature binding Singer & Gray (1995); Fries (2005): its sync detector receives a symmetrized semantic representation of a word pair together with the pair's mean emotional magnitude $(|s_i|+|s_j|)/2$, so it has no access to polarity *sign*. Predictions were registered accordingly. (i) *Controlled pairs*: five categories: same-polarity emotional (A), opposite-polarity emotional (B), emotional–neutral (C), neutral–neutral (D), and semantically related neutral (E), in a fixed template (6–8 pairs per category after vocabulary filter-

Table 29: Serial-position probe. U-index: U-shapedness of the learned position bias before/after training (positive = U-shaped). Δatt: endpoint-minus-middle attention advantage in the behavioral probe, with paired-permutation p and quadratic coefficient a of the position curve. Behavioral runs on IMDB/MR/SUBJ use the phase-2 registry checkpoints; those on SST-2/SST-5 use the final head checkpoints of Table 10.

Dataset	U-index (init)	U-index (trained)	Δatt (end.−mid.)	p	a
IMDB	0.0601	0.0608	−0.0007	0.995	−0.002
MR	0.0601	0.0606	+0.0061	10^{-4}	+0.022
SUBJ	0.0601	0.0600	+0.0003	0.094	+0.002
SST-2	0.0601	0.0640	+0.0032	10^{-4}	+0.013
SST-5	0.0601	0.0596	+0.0038	10^{-4}	+0.013

Table 30: Controlled-pair probe: median sync by pair category. A: same-polarity emotional, B: opposite-polarity emotional, C: emotional–neutral, D: neutral–neutral, E: semantically related neutral. No pairwise difference survives Holm correction (SUBJ: all p_{Holm}=1, IMDB: minimum p_{Holm}=0.59).

Dataset	A	B	C	D	E
SUBJ	0.0218	0.0215	0.0259	0.0250	0.0218
IMDB	0.2259	0.2103	0.1824	0.1992	0.1821

Table 31: In-vivo resonance selectivity (200 test sentences per dataset): median sync by edge category (EE/EN/NN = both/one/no endpoint emotionally salient; n: 1,302/2,973/8,549 edges on SUBJ and 3,112/8,189/96,979 on IMDB) and pairwise Cohen's d. All six pairwise permutation tests give $p_{\text{Holm}}=3\times10^{-4}$. The ordering EE>EN>NN is monotone on both datasets.

	SUBJ	IMDB
median sync, EE	0.0189	0.1782
median sync, EN	0.0150	0.1675
median sync, NN	0.0043	0.1581
d, EE versus EN	+0.35	+0.21
d, EE versus NN	+1.93	+0.64
d, EN versus NN	+1.31	+0.33

ing). (ii) *In-vivo edges*: 200 test sentences per dataset. Every realized coupling edge is classified EE (both endpoints emotionally salient), EN, or NN by an adaptive per-sentence salience threshold (top-quartile $|s|$, floored at 0.05), and sync values are compared across categories with permutation tests.

Results. On controlled pairs (Table 30), no category difference survives Holm correction on SUBJ (all p_{Holm}=1). On IMDB the ordering matches the magnitude hypothesis (A>B>D>C) but the probe is underpowered (n=8 pairs, $p_{\text{Holm}}\geq0.59$). On real test-set edges (Table 31), the resonance map is strongly selective: sync decreases monotonically EE>EN>NN on both datasets, with all six pairwise tests at $p_{\text{Holm}}=3\times10^{-4}$ and large effect sizes on SUBJ (Cohen's d up to 1.93). Combined with the polarity insensitivity observed in the controlled probe (A≈B), the module implements magnitude-based, synchronization-inspired selectivity: it amplifies coupling between emotionally salient words irrespective of polarity agreement, consistent with its magnitude-only input. We do not claim polarity-sensitive binding.

O.3 Valence Geometry of Hebbian Embeddings

Method. Osgood's semantic differential identifies valence as the primary factor of semantic space Osgood et al. (1957). We correlate the trained static embedding matrix (IMDB configuration) with human valence norms for 13,915 lemmas Warriner et al. (2013). The intersection covers n=5,782 vocabulary words. Three proxies are used: (i) Spearman's ρ between valence and the first five

Table 32: Valence geometry of the IMDB embedding space (n=5,782 words). $\rho(s, \text{valence})$: Spearman correlation of the model's polarity scorer with human valence; partial ρ controls log corpus frequency; ridge R^2: 5-fold CV supervised bound. Difference in $\rho(s, \text{valence})$: $+0.446$, bootstrap 95% CI [0.407, 0.486]; permutation p=10^{-4} for both embeddings.

	Hebbian	**Random-init**
$\rho(s, \text{valence})$	$+0.351$	-0.095
partial ρ (freq.)	$+0.351$	-0.100
$\rho(\text{PC1}, \text{valence})$	-0.197	$+0.165$
max $\lvert\rho(\text{PC}_k, \text{valence})\rvert$	0.321 (PC5)	0.165 (PC1)
ridge CV R^2	0.325	0.032

principal-component scores (component signs are arbitrary). (ii) The model's own polarity scorer $s(w) = \tanh(\cos(e_w, \text{seed}^+) - \cos(e_w, \text{seed}^-))$, the sentiment signal used at inference time, also as a partial correlation controlling log corpus frequency. (iii) Ridge-regression 5-fold CV R^2 as a supervised upper bound. An identical architecture trained from random embeddings serves as the control. The difference in ρ is bootstrap-tested over shared words.

Results. The Hebbian space is strongly aligned with human valence along the model's own polarity direction (Table 32): $\rho = +0.351$ (partial $\rho = +0.351$ controlling frequency; permutation p=10^{-4}), versus -0.095 for the random-embedding control; the difference ($+0.446$, bootstrap 95% CI [0.407, 0.486]) excludes zero by a wide margin, and the supervised bound shows the same gap (CV R^2 0.325 versus 0.032). The strong form of Osgood's prediction, however, is not supported: valence is not concentrated in PC1 ($|\rho|$=0.20) but spread across components (max $|\rho|$=0.32 at PC5). Hebbian co-occurrence statistics therefore encode human valence as a major, linearly decodable axis aligned with the model's polarity direction, rather than as the single principal factor of the space.