TOWARDS CONTEXT-BASED RETRIEVAL IN ASSOCIATIVE MEMORIES

Moulik Choraria[1, *] UIUC
Argyrios Gerogiannis[1] UIUC
Vidhata Jayaraman UIUC
Ankur Mani University of Minnesota, Twin Cities
Lav R. Varshney Stony Brook University

ABSTRACT

Hopfield networks and their generalizations have established deep connections between biological associative memories, statistical physics, and transformers. Yet most models treat retrieval as a fixed query-to-memory mapping, ignoring the role of external context in recall. In this work, we propose a two-stage associative memory architecture within the modular energy framework of hierarchical associative memories. Herein, a context-gate sub-circuit reshapes the retrieval energy landscape before and during recall. We demonstrate how this structure can increase inter-memory separation and exponentially improve retrieval, while actively inducing sparsity over the space of memories available for recall. Our framework offers a promising step towards a principled account of how external context can reshape retrieval dynamics.

1 INTRODUCTION

The Hopfield network (Hopfield, 1982; 1984) established one of the earliest bridges between statistical physics and neural computation, casting memory as energy minimization whose attractors encode stored patterns. Recently, dense associative memories (Krotov & Hopfield, 2016; Demircigil et al., 2017) have allowed lifting the classical linear capacity to polynomial scaling via higher-order interactions, and modern Hopfield networks (MHNs) (Ramsauer et al., 2020), achieve exponential capacity while recovering softmax attention as the update rule—revealing an equivalence between associative recall and transformer attention (Vaswani et al., 2017). This connection has inspired architectures and new insights into in-context learning (ICL) (Burns et al., 2025; Wu et al., 2025), factual recall (Nichani et al., 2024), and energy-based generative modeling (Pham et al., 2025).

What remains less understood is how *external context* shapes retrieval. Biological memory is modulated by behavioral state and task demands (Smith & Vela, 2001), yet most associative memory models treat the query as the sole retrieval input. This context-dependence phenomenon also manifests in large language models (LLMs), for instance, when the *same* query elicits drastically different behaviors based on the system prompt (Nardo, 2023), or as ICL demonstrably improves with additional contextual examples (Brown et al., 2020). Interpretability adjacent work (Hendel et al., 2023; Merullo et al., 2024; Lv et al., 2024) further suggests a two-phase delineation in tasks like ICL or factual recall: (a) a *function determination* phase that only processes context to select the task (e.g., *get_capital()*), leaving the query untouched, and (b) an *execution* phase that retrieves the answer for the query (*get_capital(France) = Paris*), while evidence of similar redundancies in multimodal processing (Jiang et al., 2025; Choraria et al., 2026) hint at a similar phenomenon at play.

Recent work has begun addressing this gap: Betteti et al. (2025) proposed input-driven synaptic plasticity that reshapes Hopfield energy landscapes, while Podlaski et al. (2025) introduced context-modular networks that partition binary memories across contexts. However, a unified energy-based architecture modeling this interaction is still missing. To address this, we propose a **two-stage associative memory** within the framework of Krotov (2021), where context drives query-based retrieval.

[*]Correspondence: moulikc2@illinois.edu; [1]Equal contribution;

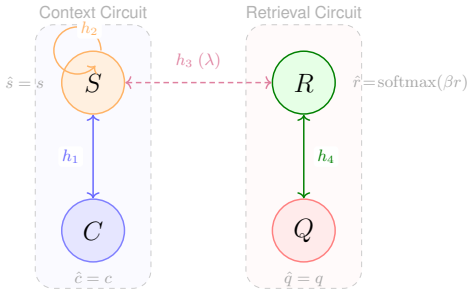

Figure 1: Two-stage associative memory architecture. The context circuit (left) establishes gate activations s from context c via alignment (h_1) and inter-memory competition (h_2). The retrieval circuit (right) recalls from stored memories via softmax attention (h_4). Cross-circuit coupling (h_3, dashed) enables bidirectional information flow controlled by λ.

Our contributions are as follows: **First**, we show theoretically that context gating increases the effective separation between memories, which translates into exponential improvements in retrieval via the connection to Modern Hopfield Networks. **Second**, we characterize a phase transition in the gate dynamics that induces winner-take-all selectivity, enabling sparse memory activation. We then validate both properties empirically on synthetic data. The organization of the paper is as follows: we outline our construction in Sec. 2, analyze its properties theoretically in Sec. 3, and validate it empirically in Sec. 4. We conclude by discussing open questions and future directions in Sec. 5.

2 Construction

We design a two-stage associative memory with the goal of exploiting the role of context in retrieval. For brevity, we defer a formal primer on the modular energy framework (Krotov et al., 2025) to Appendix A. Here, we describe its architecture and key components.

2.1 Neuron Layers

The architecture (Figure 1) comprises four layers. The **context** ($c \in \mathbb{R}^{d_c}$) receives external input specifying task or behavioral state, the **gates** ($s \in \mathbb{R}^N$) which contain one neuron per stored memory and selectively activate a sparse subset based on context, and the **query** ($q \in \mathbb{R}^{d_q}$) layer, which receives the noisy input for retrieval. Each uses a quadratic Lagrangian $L(x) = \frac{1}{2}\|x\|^2$, so the activation is the identity: $\hat{c} = c$, $\hat{s} = s$, $\hat{q} = q$. On the other hand, the **retrieval** layer ($r \in \mathbb{R}^N$) uses a LogSumExp Lagrangian, $L_r(r) = \frac{1}{\beta}\log\sum_{i=1}^{N}\exp(\beta r_i)$, yielding temperature-scaled softmax activations $\hat{r}_i = \exp(\beta r_i)/\sum_j \exp(\beta r_j)$. Since $\hat{r}_i \in [0,1]$ and $\sum_i \hat{r}_i = 1$, this layer naturally produces a distribution over the N stored memories.

2.2 Hypersynapses

We next describe the four hypersynapses encode the functional structure of the network.

(i) Context-gate alignment (h_1). A bilinear energy $E_{h_1} = -\hat{s}^\top W_{h_1}\hat{c}$ promotes alignment between the gate state and the context signal. The weight matrix $W_{h_1} \in \mathbb{R}^{N\times d_c}$ encodes how context activates gates, which in turn drive retrieval. This produces synaptic currents $I_s^{(h_1)} = W_{h_1}\hat{c}$ (context drives gates) and $I_c^{(h_1)} = W_{h_1}^\top \hat{s}$ (gates feed back to context).

(ii) Sparsity-inducing self-synapse (h_2). The crux of our construction, wherein a positive-energy self-connection $E_{\{h_2\}} = \frac{\alpha}{2}\hat{s}^\top W_{h_2}\hat{s}$ on the gate layer penalizes simultaneous activation of gates associated with similar memories. The weight matrix has entries $W_{h_2}^{ij} = \langle \zeta^i, \zeta^j\rangle$ for $i \neq j$ and zero diagonal, making the penalty proportional to inter-memory similarity and $\alpha \geq 0$ is the penalization strength. This produces a competitive signal $I_s^{(h_2)} = -2\alpha W_{h_2}\hat{s}$ that suppresses gates whose memories are close to already-active ones, enforcing sparse gate patterns without explicit regularization. The zero diagonal ensures neurons do not penalized beyond the intrinsic damping from E^{neuron}.

(iii) Cross-circuit coupling (h_3). An identity-weighted coupling $E_{h_3} = -\lambda\, \hat{s}^\top \hat{r}$ with scalar strength $\lambda > 0$ bridges the context subcircuit (S–C) and the retrieval subcircuit (R–Q). The bidirectional signals $I_s^{(h_3)} = \lambda \hat{r}$ and $I_r^{(h_3)} = \lambda \hat{s}$ allow context-derived gate activations to bias retrieval while retrieval outcomes refine the gates. Large λ forces tight coordination between subcircuits, while small λ permits near-independent operation. When the context signal has a larger magnitude, its influence on retrieval dominates the reverse direction.

(iv) Memory retrieval (h_4). A bilinear energy $E_{h_4} = -\hat{q}^\top W_{h_4} \hat{r}$ with $W_{h_4} = [\zeta^1, \ldots, \zeta^N] \in \mathbb{R}^{d_q \times N}$ encodes the stored memory patterns. The softmax activations $\hat{r}$ act as attention weights over the columns of W_{h_4}, producing a signal $I_q^{(h_4)} = \sum_\mu \zeta^\mu \hat{r}_\mu$ that is a combination of memories. The reverse direction, $I_r^{(h_4)} = W_{h_4}^\top \hat{q}$ computes the similarity of the query and stored memories.

2.3 Evolution Equations

Combining all synaptic inputs with the self-damping from each neuron layer, we arrive at a succint description for the coupled dynamical system:

$$\tau_c\, \dot{c}_i = (W_{h_1}^\top \hat{s})_i - c_i \tag{1}$$

$$\tau_s\, \dot{s}_i = (W_{h_1} \hat{c})_i - \alpha \textstyle\sum_{j \neq i} \langle \zeta^i, \zeta^j \rangle\, \hat{s}_j + \lambda\, \hat{r}_i - s_i \tag{2}$$

$$\tau_r\, \dot{r}_\mu = \lambda\, \hat{s}_\mu + \langle \zeta^\mu, \hat{q} \rangle - r_\mu \tag{3}$$

$$\tau_q\, \dot{q}_i = \textstyle\sum_\mu \zeta_i^\mu\, \hat{r}_\mu - q_i \tag{4}$$

Equation 3 reveals the two-stage mechanism: the pre-softmax score for memory μ decomposes as a context-driven gate bias $\lambda \hat{s}_\mu$ and a query-memory similarity $\langle \zeta^\mu, \hat{q} \rangle$, with the gate bias able to reshape the retrieval landscape before the query has fully resolved. When the context subcircuit operates at a faster timescale ($\tau_c, \tau_s \ll \tau_r, \tau_q$), the gates effectively precondition retrieval dynamics.

3 Theory

Now, we highlight some theoretical consequences. First, we characterize the sufficient condition for reliable retrieval as a threshold on the separation gap between target and closest competitor. This gap decomposes into raw similarity (as in standard MHNs) plus a context-driven gate contrast, showing how favorable context eases via λ. The setup is as follows: we fix a target memory index μ and query q. Define the raw similarity gap $\Delta_{\text{raw}} := \langle \zeta^\mu, q \rangle - \max_{\nu \neq \mu} \langle \zeta^\nu, q \rangle$, the gate contrast $\Delta_{\text{gate}} := s_\mu - \max_{\nu \neq \mu} s_\nu$, and the effective separation gap $\Delta := \Delta_{\text{raw}} + \lambda\, \Delta_{\text{gate}}$.

Theorem 3.1. *The target pattern μ is a stable fixed point of the retrieval dynamics with retrieval probability $P(\mu) \geq 1 - \epsilon$ if*

$$\Delta = \Delta_{\text{raw}} + \lambda\, \Delta_{\text{gate}} \geq \frac{1}{\beta} \ln\left(\frac{(1-\epsilon)(N-1)}{\epsilon} \right).$$

In particular, when $\Delta_{\text{gate}} > 0$, increasing λ increases Δ and relaxes the required raw separation.

The proof is deferred to Appendix C.1. Importantly, since context can control Δ through Δ_{gate}, including the right context c can directly increase the separation gap between memories through its action on s. And because our retrieval component uses a LogSumExp Lagrangian, it takes the form of a standard MHN and inherits its exponentially large storage capacity and exponentially small retrieval error (Ramsauer et al., 2020) (refer to Appendix B for details on MHNs). Consequently, the increased gap translates into exponential improvements in retrieval, as formalized below.

Corollary 3.2 (Corollary of Theorems 4 & 5 in Ramsauer et al. (2020); Informal)**.** *An increase in the separation Δ_i between memories gives rise to an exponential improvement in the memory retrieval in one update. Similarly, an increase in Δ_i exponentially decreases the retrieval error.*

Next we characterize how our proposed construction intrinsically promotes sparsity among active memories, even without any explicit context-based biasing. To do this, we study the self-synapse h_2 in isolation by setting $\lambda = 0$ (to exclude influence from the query) and assuming a *uniform input*, i.e. $u := W_{h_1} \hat{c}_i = p \;\forall\, i$, so the context provides no distinguishing information between memories.

Notice then, that the fixed point equation for gate vector s (Eq. 2) is of the form $A(\alpha)s^* = u$, with the linear operator $A(\alpha) \triangleq I + \alpha W_{h2} = (1-\alpha)I + \alpha G$. Here $G \in \mathbb{R}^{N\times N}$ is the Gram matrix of N unit-norm memories $\{\zeta^\mu\}_{\mu=1}^N \subset \mathbb{R}^d$ with $G_{ij} = \langle \zeta^i, \zeta^j \rangle$ and $\mathrm{rank}(G) = \min(N, d) =: R$. Such an assumption is easily satisfied with high probability under random sampling of memories.

To analyze this, we consider the spectral decomposition of $G = \sum_{k=1}^R \mu_k\, v_k v_k^\top,\ \mu_1 \geq \cdots \geq \mu_R > 0$, with $\{v_k\}_{k=1}^R$ orthonormal. When $N > R$ (i.e., $N > d$), we can extend $\{v_k\}_{k=1}^R$ to an orthonormal basis $\{v_k\}_{k=1}^N$ of $\mathbb{R}^N$ by choosing $\{v_k\}_{k=R+1}^N$ to span $\ker(G)$, with eigenvalue $\mu_k = 0$ for $k > R$. In all cases, we define the smallest eigenvalue $\mu_{\min}$ of G and let $\mathcal{V}_{\min} := \mathrm{span}\{v_k : \mu_k = \mu_{\min}\}$ denote the corresponding eigenspace, with orthogonal projector $P_{\min}$. With this setup, we can now characterize our main result on sparsity due to $s(t)$ in softmax based retrieval.

Theorem 3.3 (Critical Regime; Informal)**.** *Let $\alpha_{\mathrm{crit}} \triangleq (1-\mu_{\min})^{-1}$. Let $s(t)$ be the state and u the input. Further, let $p(t) = \mathrm{softmax}(\beta s(t))$ be the probability distribution induced by $s(t)$.*

(i) Subcritical *($\alpha < \alpha_{\mathrm{crit}}$) Then, $s^* = A(\alpha)^{-1}u = \sum_{k=1}^N \frac{c_k}{\eta_k(\alpha)} v_k$, where c_k depends on v_k and the input u. Furthermore, the modes $v_k \in \mathcal{V}_{\min}$ are maximally amplified by a factor of $1/\eta_{\min}(\alpha)$ and $\eta_{\min}(\alpha) \to 0$ as $\alpha \to \alpha_{\mathrm{crit}}$.*

(ii) Supercritical *($\alpha > \alpha_{\mathrm{crit}}$) Then, $p(t) \to \delta_{i^*}(i)$, where $i \in \{1, 2, ..., N\}$ as $t \to \infty$, where selected memory i^* is some function of the state and input, along with the spectra of G.*

The proof relies on analyzing the spectra of $A(\alpha)$, which governs the simplified linear dynamics when $\lambda = 0$. We leave the details to Appendix C.2. The key insight is that the selected index i^* is governed by (a) the *geometry* of the memory interactions through the eigenspace $\mathcal{V}_{\min}$ of G corresponding to $\mu_{\min}$, and (b) the *initialization/input bias* w.r.t. the space of stored memories.

Remark 1: To probe the $\lambda > 0$ regime, we provide a preliminary analysis in Appendix C.3 deriving a modified critical threshold $\alpha_{\mathrm{crit}}^\lambda = \alpha_{\mathrm{crit}}\left(1 - \lambda^2\beta/N\right)$. Although derived under strong assumptions about the spectra of the Gram matrix G and the structure of the context and the query, this relationship suggests that context gating aids retrieval by lowering the threshold for mode collapse. We confirm this lowering of the threshold for $\lambda > 0$ empirically in Fig 3d.

4 Experiments

We validate the properties of our framework through synthetic experiments. First, we look at how context coupling affects retrieval, by measuring how varying λ impacts Δ, and track the resulting changes in retrieval accuracy and converged probabilities. Then, we study the sparsity-inducing mechanism in isolation by empirically verifying the phase transition characterized in Theorem 3.3. Due to space constraints, we defer the exact experimental setups to the Appendix.

Context-Augmented Memory Separation. Figure 2(a) shows retrieval accuracy as a function of query noise. The accuracy is non-decreasing in λ, with the largest gains in the intermediate noise regime. Context does not eliminate error at extreme noise, but it substantially extends the operational noise range. Figure 2(b) plots retrieval probability against Δ for all trials. All empirical points lie on or above the exact logistic lower bound (red curve), confirming Theorem 3.1. The primary effect of increasing λ is to shift the point cloud rightward along the Δ-axis toward larger gaps. Figure 2(c) makes this effect explicit: the box plots show the distribution of Δ at fixed query noise $\sigma_q = 1.0$ for each λ. As λ increases, the entire gap distribution shifts upward, with higher λ values pushing the median gap well above the exact threshold required.

Phase Transition Experiments Figure 3(a) shows that the peak gate probability $\max_i p_i^*$ undergoes a sharp transition at α_{crit}: for $\alpha < \alpha_{\mathrm{crit}}$, it stays near $1/N$, consistent with the spectral filtering regime where $1/\eta_N(\alpha)$ remains finite; for $\alpha > \alpha_{\mathrm{crit}}$, it saturates near 1.0, matching the WTA limit of Theorem 3.3. Figure 3(b) resolves the transition region $[\alpha_{\mathrm{crit}},\ 1.18]$, revealing a smooth but steep curve. An example gate distribution at $\alpha = \alpha_{\mathrm{crit}}$ (Figure 3(c)) shows the onset of selectivity: multiple memories emerge above the uniform baseline, reflecting the divergent amplification of the least-correlated mode v_N as $\eta_N \to 0^+$. Figure 3(d) showcases the transition of the critical region as λ increases. As λ increases the empirical α_{crit} shifts to smaller values. While this agrees with our preliminary results in Appendix C.3, we emphasize that our theoretical threshold does not match the empirical threshold. Nevertheless, this suggests that context gating indeed aids retrieval.

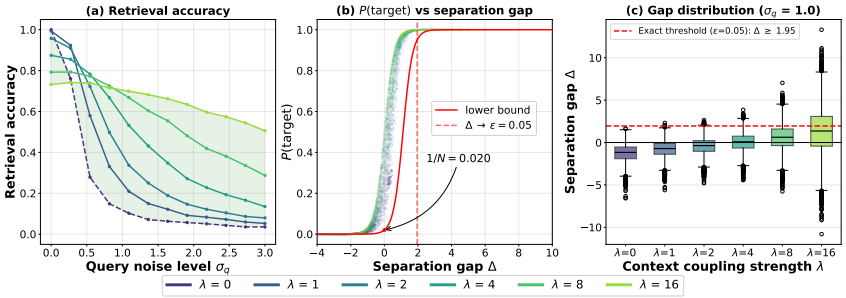

Figure 2: Context-augmented memory separation. **(a)** Retrieval accuracy vs. query noise. **(b)** Retrieval probability vs. effective separation gap Δ. **(c)** Distribution of Δ at $\sigma_q = 1.0$ for each λ. As λ increases the retrieval becomes more accurate under higher query noise and all points lie above the logistic lower bound. At the same time higher λ ensures larger effective separation gap Δ.

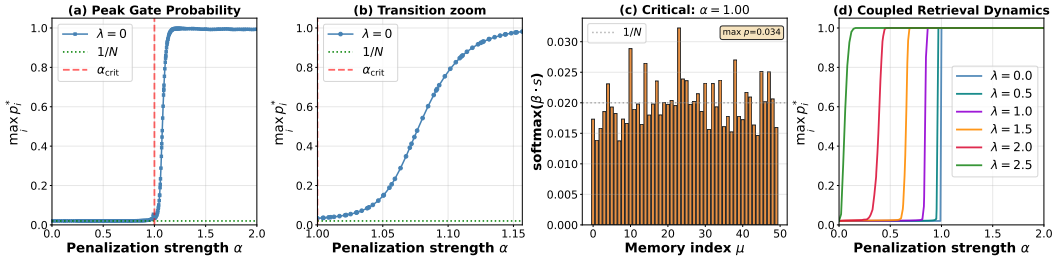

Figure 3: Phase transition in gate selectivity. **(a)** Peak gate probability vs. penalization strength ($\lambda = 0$). **(b)** Zoom on the transition region ($\lambda = 0$). **(c)** Example gate distribution at $\alpha = \alpha_{\text{crit}}$ ($\lambda = 0$). **(d)** Peak gate probability vs. penalization strength for varying λ. The retrieval dynamics demonstrate a sharp transition at α_{crit} towards a singe memory, breaking the symmetry of the input, while the increase in λ shifts the critical region to the left, meaning faster emergence of sparsity.

5 Discussion

Inspired by the role of context across domains, we propose a modular energy-based architecture to explain how context shapes associative retrieval. We establish desirable properties such as increased memory separation and sparsity, both of which facilitate improved retrieval. Nevertheless, our work opens up interesting theoretical and empirical directions, several of which we discuss below.

On the theoretical side, a formal characterization of the influence of gating on query-based retrieval is still needed. We provide preliminary analysis for this in Appendix C.3, though relaxing the assumptions therein requires careful thought. In particular, studying what happens when query and context compete (for instance, in LLMs, consider a context of "Be a liar" paired with the query "What is the capital of France?") could unlock key insights not only for associative memory design, but also for understanding transformer mechanisms more broadly. Relatedly, the context-query demarcation appears to emerge naturally in LLMs, and an open question is whether we can rigorously characterize the benefits it confers via our framework, such as sparsity enabling ease of representation learning and improved retrieval. One promising avenue is studying the abstraction and limits of top-k retrieval in the framework of Weller et al. (2025), and whether gating can improve on the constraints on minimum required embedding rank for successful retrieval.

Finally, our empirical evaluation is limited to synthetic data, and it remains unclear how one would extract context-query memory pairs suitable for showcasing the network's practical functioning on real data. A promising direction may be leveraging LLM embeddings in ICL or contextual question answering tasks; beyond validating our framework, such experiments could shed light on how context shapes retrieval in transformers more broadly.

REFERENCES

Simone Betteti, Giacomo Baggio, Francesco Bullo, and Sandro Zampieri. Input-driven dynamics for robust memory retrieval in Hopfield networks. *Science Advances*, 11(17):eadu6991, 2025. doi: 10.1126/sciadv.adu6991.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Thomas F Burns, Tomoki Fukai, and Christopher J Earls. Associative memory inspires improvements for in-context learning using a novel attention residual stream architecture, 2025. URL https://arxiv.org/abs/2412.15113.

Moulik Choraria, Xinbo Wu, Akhil Bhimaraju, Nitesh Sekhar, Yue Wu, Xu Zhang, Prateek Singhal, and Lav R. Varshney. Deepinsert: Early layer bypass for efficient and performant multimodal understanding, 2026. URL https://arxiv.org/abs/2504.19327.

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, May 2017. doi: 10.1007/s10955-017-1806-y.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors, 2023. URL https://arxiv.org/abs/2310.15916.

J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554.

J J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, 81(10):3088–3092, 1984. doi: 10.1073/pnas.81.10.3088.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proc. Comput. Vis. Pattern Recog. (CVPR)*, pp. 25004–25014, 2025.

Dmitry Krotov. Hierarchical associative memory, 2021. URL https://arxiv.org/abs/2107.06446.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/eaae339c4d89fc102edd9dbdb6a28915-Paper.pdf.

Dmitry Krotov, Benjamin Hoover, Parikshit Ram, and Bao Pham. Modern methods in associative memory, 2025. URL https://arxiv.org/abs/2507.06211.

Ang Lv, Yuhan Chen, Kaiyi Zhang, Yulong Wang, Lifeng Liu, Ji-Rong Wen, Jian Xie, and Rui Yan. Interpreting key mechanisms of factual recall in transformer-based language models. arXiv 2403.19521 [cs.CL], 2024.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. A mechanism for solving relational tasks in transformer language models, 2024. URL https://openreview.net/forum?id=ZmzLrl8nTa.

Cleo Nardo. The Waluigi effect (mega-post). AI Alignment Forum, March 2023. URL https://www.alignmentforum.org/posts/D7PumeYTDPfBTp3i7/the-waluigi-effect-mega-post. Accessed: 2026-02-07.

Eshaan Nichani, Jason D. Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories, 2024. URL https://arxiv.org/abs/2412.06538.

Bao Pham, Gabriel Raya, Matteo Negri, Mohammed J. Zaki, Luca Ambrogioni, and Dmitry Krotov. Memorization to generalization: Emergence of diffusion models from associative memory, 2025. URL https://arxiv.org/abs/2505.21777.

William F. Podlaski, Everton J. Agnes, and Tim P. Vogels. High capacity and dynamic accessibility in associative memory networks with context-dependent neuronal and synaptic gating. *Phys. Rev. X*, 15:011057, Mar 2025. doi: 10.1103/PhysRevX.15.011057.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

S. M. Smith and E. Vela. Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review*, 8(2):203–220, 2001. doi: https://doi.org/10.3758/BF03196157.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval, 2025. URL https://arxiv.org/abs/2508.21038.

Weimin Wu, Teng-Yun Hsiao, Jerry Yao-Chieh Hu, Wenxin Zhang, and Han Liu. In-context learning as conditioned associative memory retrieval. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Zup6F3MwQO.

A PRIMER: GENERAL DENSE ASSOCIATIVE MEMORIES

A.1 MODULAR ENERGY FRAMEWORK

Here, we provide a brief introduction on the generalized abstraction of Energy-based AMs (HAMUX), introduced by Krotov et al. (2025). At a high level, Dense Associative Memories (DAMs) are composed of two major components: **neuron layers** (nodes) and **hypersynapses** (hyperedges). A neuron layer captures a non-linearity in the network (or activations such as ReLU), while a hypersynapse is a parameterized energy function that captures how similar or aligned the activations of its connected neuron layers are. The DAM is then described by summing the modular energies of all these components. The following exposition provides a gentle introduction. For a formal treatment, we point the reader towards Krotov (2021); Krotov et al. (2025).

A.1.1 TOTAL ENERGY

For a system with L neuron layers and S hypersynapses, the total energy is:

$$E_{\text{total}} = \sum_{\ell=1}^{L} E_{\ell}^{\text{neuron}} + \sum_{s=1}^{S} E_{s}^{\text{synapse}} \tag{5}$$

A.1.2 LOCAL UPDATE RULE

Let $\hat{\mathbf{x}}_\ell$ and $\mathbf{x}_\ell$ denote the activations and internal states of neuron layer ℓ. Let $\mathcal{N}(\ell)$ be the set of hypersynapses connected to layer ℓ. The internal states minimize total energy via:

$$\tau_\ell \frac{d\mathbf{x}_\ell}{dt} = -\frac{\partial E_{\text{total}}}{\partial \hat{\mathbf{x}}_\ell} = -\left(\sum_{s \in \mathcal{N}(\ell)} \frac{\partial E_s^{\text{synapse}}}{\partial \hat{\mathbf{x}}_\ell} \right) - \frac{\partial E_\ell^{\text{neuron}}}{\partial \hat{\mathbf{x}}_\ell} = \mathbf{I}_{\mathbf{x}_\ell} - \mathbf{x}_\ell \tag{6}$$

where $\mathbf{I}_{\mathbf{x}_\ell} := -\sum_{s \in \mathcal{N}(\ell)} \nabla_{\hat{\mathbf{x}}_\ell} E_s^{\text{synapse}}$ is the total synaptic input current and τ_ℓ is the time constant.

A.2 DYNAMICAL NEURONS AND LAGRANGIANS

Definition A.1 (Neuron Layer)**.** A neuron layer has internal state $\mathbf{x}$ and activation $\hat{\mathbf{x}}$, defined via a convex Lagrangian $\mathcal{L}_{\mathbf{x}}(\mathbf{x})$ and its Legendre transform $\mathcal{T}$:

$$\hat{\mathbf{x}} = \nabla \mathcal{L}_{\mathbf{x}}(\mathbf{x}) \quad \text{(activation function)} \tag{7}$$

$$E_{\mathbf{x}}(\hat{\mathbf{x}}) = \mathcal{T}[\mathcal{L}_{\mathbf{x}}] = \langle \mathbf{x}, \hat{\mathbf{x}} \rangle - \mathcal{L}_{\mathbf{x}}(\mathbf{x}) \quad \text{(dual energy)} \tag{8}$$

where $\langle \cdot, \cdot \rangle$ is the element-wise inner product.

Key Property: The energy gradient equals the hidden state:

$$\frac{\partial E_{\mathbf{x}}(\hat{\mathbf{x}})}{\partial \hat{\mathbf{x}}} = \mathbf{x} \tag{9}$$

Proof:

$$\begin{aligned} \frac{\partial E_{\mathbf{x}}(\hat{\mathbf{x}})}{\partial \hat{\mathbf{x}}} &= \frac{\partial}{\partial \hat{\mathbf{x}}} \left(\langle \mathbf{x}, \hat{\mathbf{x}} \rangle - \mathcal{L}_{\mathbf{x}}(\mathbf{x}) \right) \\ &= \mathbf{x} + \hat{\mathbf{x}} \frac{\partial \mathbf{x}}{\partial \hat{\mathbf{x}}} - \frac{\partial \mathcal{L}_{\mathbf{x}}(\mathbf{x})}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \hat{\mathbf{x}}} \\ &= \mathbf{x} + \hat{\mathbf{x}} \frac{\partial \mathbf{x}}{\partial \hat{\mathbf{x}}} - \hat{\mathbf{x}} \frac{\partial \mathbf{x}}{\partial \hat{\mathbf{x}}} = \mathbf{x} \end{aligned}$$

This implies exponential decay in isolation:

$$\frac{d\mathbf{x}}{dt} = -\nabla_{\hat{\mathbf{x}}} E_{\mathbf{x}}(\hat{\mathbf{x}}) = -\mathbf{x} \tag{10}$$

A.3 HYPERSYNAPSES

Definition A.2 (Hypersynapse)**.** A hypersynapse is a scalar-valued energy function defined on the activations of connected neuron layers. For layers X and Y with activations $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$:

$$E_{xy}(\hat{\mathbf{x}}, \hat{\mathbf{y}}; \Xi) \tag{11}$$

where Ξ represents learnable synaptic weights. Low energy indicates activations satisfy the relationship encoded by Ξ.

Example: Dense hypersynapse: $E_{\text{Dense}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}; \Xi) = -\hat{\mathbf{x}}^\top \Xi \hat{\mathbf{y}}$

Key Properties:

- Hypersynapses connect any number of layers (hyperedges)
- Undirected: all connected layers influence each other bidirectionally
- Signal to layer X: $\mathbf{I}_x = -\nabla_{\hat{\mathbf{x}}} E_{xy}(\hat{\mathbf{x}}, \hat{\mathbf{y}}; \Xi)$
- Signal to layer Y: $\mathbf{I}_y = -\nabla_{\hat{\mathbf{y}}} E_{xy}(\hat{\mathbf{x}}, \hat{\mathbf{y}}; \Xi)$

Hypersynapse Notation Conventions:

For synapses connecting multiple layers, we subscript with the identifiers of all connected layers:

- E_{xy} — synapse connecting layers X and Y
- E_{xyz} — synapse connecting layers X, Y, and Z
- $E_{xyz\ldots}$ — synapses connecting more than three layers (possible but rare)

Self-connections: To avoid confusion with neuron layer energy $E_{\mathbf{x}}$, we use curly brackets for synaptic self-connections:

- $E_{\{x\}}$ — interaction energy of a synapse connecting layer X to itself

Note: Since almost every interaction energy is parameterized, we generally omit Ξ from notation when not central to the discussion.

A.4 ENERGY DESCENT DYNAMICS

Theorem A.3 (Guaranteed Energy Descent)**.** *The dynamics in Eq. (2) decrease the global energy:*

$$\frac{dE_{total}}{dt} = \sum_{\ell=1}^{L} \frac{\partial E_{total}}{\partial \hat{\mathbf{x}}_\ell} \frac{\partial \hat{\mathbf{x}}_\ell}{\partial \mathbf{x}_\ell} \frac{d\mathbf{x}_\ell}{dt} = -\sum_{\ell=1}^{L} \tau_\ell \frac{d\mathbf{x}_\ell}{dt} \frac{\partial^2 \mathcal{L}_{\mathbf{x}}}{\partial \mathbf{x}_\ell \partial \mathbf{x}_\ell} \frac{d\mathbf{x}_\ell}{dt} \leq 0 \tag{12}$$

The Hessian $\frac{\partial^2 \mathcal{L}_{\mathbf{x}}}{\partial \mathbf{x}_\ell \partial \mathbf{x}_\ell}$ is positive semi-definite due to convexity of $\mathcal{L}_{\mathbf{x}}$, ensuring energy never increases.

Convergence: If energy is bounded below:

- Strictly positive definite Hessian $\Rightarrow$ converges to fixed point
- Zero eigenvalues $\Rightarrow$ may converge to fixed manifold with non-zero velocity

A.5 SUMMARY

To design an Associative Memory:

1. Choose convex Lagrangian $\mathcal{L}_{\mathbf{x}}(\mathbf{x})$ for each neuron layer (defines activation function)
2. Design hypersynapse energies E_s^{synapse} encoding desired relationships
3. Total energy: sum of all neuron and hypersynapse energies
4. Dynamics: minimize energy via local gradient descent (Eq. 2)
5. Guaranteed convergence with bounded activations

B MODERN HOPFIELD NETWORKS

To keep this work self-contained, we provide a brief introduction to Modern Hopfield Networks (MHN) (Ramsauer et al., 2020). MHNs are a form of DAM with an energy function of the form

$$E = -\text{lse}(\beta, X^T\xi) + \frac{1}{2}\xi^T\xi + \beta^{-1}\log N + \frac{1}{2}M^2 \tag{13}$$

where $\text{lse}(\beta, \cdot)$ is the LogSumExp function with temperature parameter β, N is the number of stored patterns, M is the norm of the largest pattern (i.e. $M = \max_i \|x_i\|$), and ξ is the query. The update rule for the MHN is then given by

$$\xi^{new} = X\text{softmax}(\beta X^T\xi). \tag{14}$$

Ramsauer et al. (2020) provide theorems to guarantee convergence for this form of Associative Memory and further establishes an exponential storage capacity (see Theorems 1, 2, and 3 in their paper). Of particular interest in this paper are Theorems 4 and 5.

Theorem B.1 (Theorem 4 in Ramsauer et al. (2020))**.** *With query ξ, pattern x_i, fixed point x_i^*, and separation of x_i to other memories Δ_i, after one update, the distance between $f(\xi)$ and x_i^* is exponentially small. Specifically,*

$$\|f(\xi) - x_i^*\| \leq 2\beta N M^2 (N-1)\exp(-\beta(\Delta_i - 2\max\{\|\xi - x_i\|, \|x_i^* - x_i\|M))\|\xi - x_i^*\|.$$

Theorem B.2 (Theorem 5 in Ramsauer et al. (2020))**.** *The retrieval error $\|f(\xi) - x_i\|$ is bounded by*

$$\|f(\xi) - x_i\| \leq 2(N-1)\exp(-\beta(\Delta_i - 2\max\{\|\xi - x_i\|, \|x_i^* - x_i\|M))M$$

and for $\|x_i - x_i^*\| \leq \frac{1}{2\beta M}$ together with $\|x_i - \xi\| \leq \frac{1}{2\beta M}$ we have

$$\|x_i - x_i^*\| \leq 2e(N-1)M\exp(-\beta\Delta_i).$$

We see from these theorems that increasing the separation between memories yields an exponential improvement in both retrieval after one update and retrieval error.

B.1 MHN'S CONNECTION TO TRANSFORMERS

Notably, MHNs bear a strong mathematical resemblance to the attention mechanism in Transformers (Vaswani et al., 2017). This connection can be seen below.

Suppose we have N stored patterns y_i and S state query patterns r_i that are mapped to a space of dimension d_k. Set $x_i = W_K^T y_i$ and $\xi_i = W_Q^t r_i$ and then multiply the result of the update rule (Eq. 14) by W_V where $W_K \in \mathbb{R}^{d_y \times d_k}$, $W_Q \in \mathbb{R}^{d_r \times d_k}$, $W_V \in \mathbb{R}^{d_k \times d_v}$.

If we combine everything into matrix operations as is commonly done for attention, we arrive at the following. Let $Y = (y_1, \ldots, y_N)^T$, $R = (r_1, \ldots, r_N)^T$. Define $X^T = K = YW_K$, $\Xi^T = Q = RW_Q$, and $V = YW_KW_V = X^TW_V$. Let the temperature parameter in MHN $\beta = \frac{1}{\sqrt{d_k}}$ and let the output of softmax be a row-vector. Then for the update rule in matrix form multiplied by W_V we get

$$Z = \text{softmax}(\frac{1}{\sqrt{d_k}}QK^T)V = \text{softmax}(\beta RW_QW_K^TY^T)YW_KW_V. \tag{15}$$

We can recognize the left part of the equation as being the attention mechanism, while the right side is the update rule for MHNs, followed by a matrix multiplication with W_V.

C THEORETICAL PROOFS

C.1 PROOF OF THEOREM 3.1

Proof. Let,

$$\mathcal{S}(\mu, q) := \langle \zeta^\mu, q\rangle + \lambda\, s_\mu \tag{16}$$

At the retrieval fixed point, $r_k^* = \mathcal{S}(k, q)$ for each memory k. The retrieval probability for the target μ is the softmax output:

$$P(\mu) = \hat{r}_\mu = \frac{e^{\beta\,\mathcal{S}(\mu,q)}}{\sum_{k=1}^{M} e^{\beta\,\mathcal{S}(k,q)}}. \tag{17}$$

Dividing numerator and denominator by $e^{\beta\,\mathcal{S}(\mu,q)}$:

$$P(\mu) = \frac{1}{1 + \sum_{\nu\neq\mu} e^{-\beta\,\Delta_\nu}}, \tag{18}$$

where $\Delta_\nu := \mathcal{S}(\mu, q) - \mathcal{S}(\nu, q) \geq \Delta$ for all $\nu \neq \mu$, by definition of Δ as the minimum gap. Since the exponential is monotonically decreasing, $e^{-\beta\,\Delta_\nu} \leq e^{-\beta\,\Delta}$ for each $\nu \neq \mu$. Summing over the $M - 1$ distractor terms:

$$P(\mu) \geq \frac{1}{1 + (M-1)\,e^{-\beta\,\Delta}}, \tag{19}$$

To derive the threshold condition, we require $P(\mu) \geq 1 - \epsilon$, i.e.:

$$\frac{1}{1 + (M-1)\,e^{-\beta\,\Delta}} \geq 1 - \epsilon. \tag{20}$$

Inverting and rearranging:

$$1 + (M-1)\,e^{-\beta\,\Delta} \leq \frac{1}{1-\epsilon}, \tag{21}$$

$$(M-1)\,e^{-\beta\,\Delta} \leq \frac{1}{1-\epsilon} - 1 = \frac{\epsilon}{1-\epsilon}. \tag{22}$$

Taking logarithms:

$$-\beta\,\Delta \leq \ln\left(\frac{\epsilon}{(1-\epsilon)(M-1)}\right), \tag{23}$$

which leads to the final result. □

C.2 Formal version of Theorem 3.3 and Proof

Theorem C.1. *Let $\alpha_{\text{crit}} \triangleq (1 - \mu_{\min})^{-1}$, and for state $s(t)$ and input u, define projections on the eigenvectors as $a_k(t) = v_k^\top s(t)$ and $c_k = v_k^\top u$ respectively. Finally, define $p(t) = \text{softmax}(\beta s(t))$ as the distribution induced by $s(t)$ post softmax.*

(i) Subcritical *($\alpha < \alpha_{\text{crit}}$). Then, $A(\alpha) \succ 0$ and the fixed point solution is unique:*

$$s^* = A(\alpha)^{-1} u = \sum_{k=1}^{N} \frac{c_k}{\eta_k(\alpha)}\, v_k, \qquad p^* = \text{softmax}(\beta s^*).$$

The modes $v_k \in \mathcal{V}_{\min}$ have the maximum possible amplification which is given by $\frac{1}{\eta_{\min}(\alpha)}$. Furthermore, $\eta_{\min}(\alpha) \to 0$ as $\alpha \to \alpha_{\text{crit}}$ leading to larger amplifications.

(ii) Supercritical *($\alpha > \alpha_{\text{crit}}$). Define $\gamma_k(\alpha) \triangleq \alpha(1 - \mu_k) - 1$, and $\xi \triangleq P_{\min} \sum_{\{k:\mu_k=\mu_{\min}\}} w_k\, v_k$, with $w_k \triangleq a_k(0) + c_k/\gamma_k(\alpha)$.*

$$\xi \triangleq P_{\min}\, w, \qquad w \triangleq \sum_{\mu_k=\mu_{\min}} w_k\, v_k, \qquad w_k \triangleq a_k(0) + \frac{c_k}{\gamma_{\min}(\alpha)}.$$

Assume $\xi \neq 0$ and $i^* := \arg\max_{i\in[N]} \xi_i$ is unique. Then

$$p(t) \to \delta_{i,i^*} \quad \text{as } t \to \infty.$$

In particular, for any $\delta \in (0, 1)$, $p_{i^*}(t^*(\alpha)) \geq 1 - \delta$ with

$$t^*(\alpha) = \frac{\tau_s}{\gamma_{\min}(\alpha)}\left[\ln\left(\frac{\kappa}{\beta\,\Delta_\xi}\right) + O(1)\right], \quad \Delta_\xi := \xi_{i^*} - \max_{j\neq i^*} \xi_j > 0,$$

where $\kappa > 0$ is independent of α.

Proof. Let N unit-norm memories $\{\zeta^\mu\}_{\mu=1}^N \subset \mathbb{R}^d$ with Gram matrix $G \in \mathbb{R}^{N\times N}$, $G_{ij} = \langle \zeta^i, \zeta^j \rangle$, so $G_{ii} = 1$. Define $W_{h_2} := G - I$ and

$$A(\alpha) := I + \alpha W_{h_2} = (1-\alpha)I + \alpha G. \tag{24}$$

Let $R := \operatorname{rank}(G) = \min(N, d)$ and write the spectral decomposition

$$G = \sum_{k=1}^{R} \mu_k \, v_k v_k^\top, \qquad \mu_1 \geq \cdots \geq \mu_R > 0, \tag{25}$$

with orthonormal $\{v_k\}_{k=1}^R$. When $N > R$, extend $\{v_k\}_{k=1}^R$ to an orthonormal basis $\{v_k\}_{k=1}^N$ of $\mathbb{R}^N$ by choosing $\{v_k\}_{k=R+1}^N$ to span $\ker(G)$, and set $\mu_k := 0$ for $k > R$. Define the smallest eigenvalue

$$\mu_{\min} := \mu_N = \begin{cases} \mu_N > 0 & \text{if } N \leq d, \\ 0 & \text{if } N > d, \end{cases} \tag{26}$$

and let $\mathcal{V}_{\min} := \operatorname{span}\{v_k : \mu_k = \mu_{\min}\}$ denote its eigenspace with orthogonal projector

$$P_{\min} := \sum_{\mu_k = \mu_{\min}} v_k v_k^\top. \tag{27}$$

Then

$$A(\alpha)v_k = \eta_k(\alpha) v_k, \qquad \eta_k(\alpha) = 1 - \alpha(1-\mu_k), \qquad k = 1, \ldots, N, \tag{28}$$

so the smallest eigenvalue is

$$\eta_{\min}(\alpha) := \min_k \eta_k(\alpha) = 1 - \alpha(1 - \mu_{\min}), \tag{29}$$

attained for all k with $\mu_k = \mu_{\min}$. Define

$$\alpha_{\text{crit}} := \frac{1}{1 - \mu_{\min}}. \tag{30}$$

Write the input as $u = \sum_{k=1}^N c_k v_k$ with $c_k := v_k^\top u$ and recall $p(t) = \operatorname{softmax}(\beta s(t))$.

Subcritical case ($\alpha < \alpha_{\text{crit}}$). Since $\alpha < \alpha_{\text{crit}}$ iff $\eta_{\min}(\alpha) > 0$, we have $\eta_k(\alpha) \geq \eta_{\min}(\alpha) > 0$ for all k, hence $A(\alpha) \succ 0$ and is invertible. The fixed point is unique and satisfies

$$s^* = A(\alpha)^{-1} u = \sum_{k=1}^{N} \frac{c_k}{\eta_k(\alpha)} \, v_k, \qquad p^* = \operatorname{softmax}(\beta s^*). \tag{31}$$

Thus mode k is amplified by $1/\eta_k(\alpha)$, and the modes in $\mathcal{V}_{\min}$ are maximally amplified by $1/\eta_{\min}(\alpha)$ with $\eta_{\min}(\alpha) \to 0$ as $\alpha \to \alpha_{\text{crit}}$.

Supercritical case ($\alpha > \alpha_{\text{crit}}$). Consider the gate dynamics

$$\tau_s \, \dot{s}(t) = u - A(\alpha) s(t), \qquad p(t) = \operatorname{softmax}(\beta s(t)), \tag{32}$$

and define modal coordinates $a_k(t) := v_k^\top s(t)$. Projecting equation 32 onto v_k gives

$$\tau_s \, \dot{a}_k(t) = c_k - \eta_k(\alpha) a_k(t), \tag{33}$$

so

$$a_k(t) = \frac{c_k}{\eta_k(\alpha)} + \Big(a_k(0) - \frac{c_k}{\eta_k(\alpha)}\Big) e^{-\eta_k(\alpha) t/\tau_s}. \tag{34}$$

Since $\alpha > \alpha_{\text{crit}}$, $\eta_{\min}(\alpha) < 0$ on $\mathcal{V}_{\min}$. Set

$$\gamma_{\min}(\alpha) := -\eta_{\min}(\alpha) = \alpha(1 - \mu_{\min}) - 1 > 0. \tag{35}$$

For each k with $\mu_k = \mu_{\min}$ (hence $\eta_k(\alpha) = \eta_{\min}(\alpha)$), rewrite equation 34 as

$$a_k(t) = -\frac{c_k}{\gamma_{\min}(\alpha)} + w_k \, e^{\gamma_{\min}(\alpha) t/\tau_s}, \qquad w_k := a_k(0) + \frac{c_k}{\gamma_{\min}(\alpha)}. \tag{36}$$

All modes with $\mu_k > \mu_{\min}$ satisfy $\eta_k(\alpha) > \eta_{\min}(\alpha)$ and, in particular, remain stable in a neighborhood of α_{crit}, contributing bounded terms to $s(t)$.

Expanding $s(t) = \sum_{k=1}^{N} a_k(t)v_k$ and collecting all stable-mode contributions and constant offsets into $r(t)$ yields, for each coordinate i,

$$s_i(t) = \xi_i\, e^{\gamma_{\min}(\alpha)t/\tau_s} + r_i(t), \quad \xi_i := \sum_{\mu_k=\mu_{\min}} w_k\,(v_k)_i = \big(P_{\min}w\big)_i, \tag{37}$$

where $w := \sum_{\mu_k=\mu_{\min}} w_k v_k \in \mathcal{V}_{\min}$ and $|r_i(t)| \le C$ for all t, for a constant C independent of t. Assume $\xi := (\xi_i)_{i=1}^N \neq 0$ and choose

$$i^* := \arg\max_{i\in[N]} \xi_i, \tag{38}$$

with the maximizer unique. For any $j \neq i^*$, subtracting equation 37 gives

$$s_{i^*}(t) - s_j(t) = (\xi_{i^*} - \xi_j)e^{\gamma_{\min}(\alpha)t/\tau_s} + O(1) \xrightarrow[t\to\infty]{} +\infty, \tag{39}$$

since $\xi_{i^*} > \xi_j$. Then

$$p_j(t) = \frac{e^{\beta s_j(t)}}{\sum_\ell e^{\beta s_\ell(t)}} \le \frac{e^{\beta s_j(t)}}{e^{\beta s_{i^*}(t)}} = e^{-\beta(s_{i^*}(t)-s_j(t))} \xrightarrow[t\to\infty]{} 0, \tag{40}$$

so $p(t) \to \delta_{i,i^*}$, proving the WTA claim.

Let $j^* \in \arg\max_{j\neq i^*} \xi_j$ be a strongest competitor and define $\Delta_\xi := \xi_{i^*} - \xi_{j^*} > 0$. The softmax identity gives

$$p_{i^*}(t) = \frac{1}{1 + \sum_{j\neq i^*} e^{-\beta(s_{i^*}(t)-s_j(t))}} \ge \frac{1}{1 + (N-1)e^{-\beta(s_{i^*}(t)-s_{j^*}(t))}}. \tag{41}$$

Thus $p_{i^*}(t) \ge 1-\delta$ is ensured by

$$\beta\big(s_{i^*}(t) - s_{j^*}(t)\big) \ge \ln\Big(\frac{(1-\delta)(N-1)}{\delta}\Big) =: L_\delta. \tag{42}$$

By equation 37, the leading gap satisfies

$$s_{i^*}(t) - s_{j^*}(t) = \Delta_\xi\, e^{\gamma_{\min}(\alpha)t/\tau_s} + O(1). \tag{43}$$

Solving equation 42 using equation 43 and absorbing additive constants into an $O(1)$ bracket yields, for some $\kappa > 0$ independent of α,

$$t^*(\alpha) = \frac{\tau_s}{\gamma_{\min}(\alpha)}\Big[\ln\Big(\frac{\kappa}{\beta\,\Delta_\xi}\Big) + O(1)\Big], \tag{44}$$

which is exactly the claimed form.

Finally, when $N \le d$ and μ_N is simple, $\mathcal{V}_{\min} = \mathrm{span}\{v_N\}$ so $\xi_i = w_N(v_N)_i$ and $i^* = \arg\max_i w_N(v_N)_i$. When $N > d$, $\mu_{\min} = 0$ and $\gamma_{\min}(\alpha) = \alpha - 1$, while $P_{\min}$ is the projector onto $\ker(G)$, so the winner is determined by the projection of w onto $\ker(G)$. □

C.3 Analysis for $\lambda > 0$

Theorem C.2 (Critical Regime with $\lambda > 0$)**.** *Let $\lambda > 0$ and $\tau_s \ll \tau_r$ so that the gates settle to $s^* = A(\alpha)^{-1}(u + \lambda\hat{r})$ on the fast timescale. Further assume that the context input is uniform (i.e. $u = u_0\mathbf{1}$), the query provides no bias (i.e. $\langle\zeta^\mu, \hat{q}\rangle = q_0$ for all $\mu \in [N]$), and $G\mathbf{1} = \mu_G\mathbf{1}$ for some $\mu_G > 0$.*

Let $\mu_{\min}^\perp$ denote the smallest eigenvalue of G restricted to the subspace $\mathbf{1}^\perp \triangleq \{v \in \mathbb{R}^N : \langle v, \mathbf{1}\rangle = 0\}$, and define $\alpha_{\mathrm{crit}} \triangleq (1 - \mu_{\min}^\perp)^{-1}$. Then, provided $\lambda^2\beta < N$, the uniform retrieval state *$\hat{r} = \frac{1}{N}\mathbf{1}$ is a fixed point of the slow retrieval dynamics whose stability is governed by the threshold*

$$\alpha_{\mathrm{crit}}^\lambda \;=\; \alpha_{\mathrm{crit}}\left(1 - \frac{\lambda^2\beta}{N}\right).$$

The uniform fixed point is stable for $\alpha < \alpha_{\mathrm{crit}}^\lambda$ and unstable for $\alpha > \alpha_{\mathrm{crit}}^\lambda$.

In particular, for $\alpha_{\mathrm{crit}}^\lambda < \alpha < \alpha_{\mathrm{crit}}$, the gate subsystem alone is subcritical *(its uniform state is stable), and it is the retrieval feedback loop through the softmax nonlinearity that destabilizes the uniform state, inducing winner-take-all retrieval.*

Proof. Recall $A(\alpha) = (1-\alpha)I + \alpha G$. By assumption, $\mathbf{1}$ is an eigenvector of G with eigenvalue μ_G, hence of $A(\alpha)$ with eigenvalue

$$\eta_G(\alpha) = 1 - \alpha(1-\mu_G). \tag{45}$$

Observe that G is symmetric, so all eigenvectors other than $\mathbf{1}$ lie in $\mathbf{1}^\perp$. Thus, for $\alpha < \alpha_{\text{crit}}$ all eigenvalues $\eta_k(\alpha) = 1 - \alpha(1-\mu_k)$ are positive (since $\mu_k \geq \mu_{\min}^\perp$), so $A(\alpha)$ is invertible with $A(\alpha)^{-1}\mathbf{1} = \eta_G^{-1}\,\mathbf{1}$.

At the uniform retrieval state $\hat{r} = \frac{1}{N}\mathbf{1}$, the gate steady-state is

$$s^* = A(\alpha)^{-1}\big(u_0\mathbf{1} + \tfrac{\lambda}{N}\mathbf{1}\big) = \frac{u_0 + \lambda/N}{\eta_G(\alpha)}\,\mathbf{1}.$$

Substituting into the retrieval dynamics equation 3 and absorbing the constant query bias $q_0\mathbf{1}$ into a scalar, the effective slow dynamics can be written as the vector field

$$\tau_r \dot{r} \;=\; \underbrace{\lambda\, A(\alpha)^{-1} u}_{\text{constant bias}} \;+\; \lambda^2\, A(\alpha)^{-1}\,\hat{r} \;+\; q_0\,\mathbf{1} \;-\; r, \tag{46}$$

where $\hat{r} = \mathrm{softmax}(\beta r)$. Observe that $r^* = c\mathbf{1}$ where

$$c = \frac{\lambda u_0}{\eta_G(\alpha)} + \frac{\lambda^2}{N\eta_G(\alpha)} + q_0,$$

is indeed a fixed point of the dynamics.

The stability of this system is determined by the Jacobian of the vector field $F(r) = \lambda^2 A(\alpha)\hat{r} - r$ evaluated at the uniform fixed point. The Jacobian of the softmax at $\hat{r} = \frac{1}{N}\mathbf{1}$ is

$$\mathcal{J} \;\triangleq\; \frac{\partial\,\hat{r}}{\partial\, r}\bigg|_{r=c\mathbf{1}} = \beta\big(\mathrm{diag}(\hat{r}) - \hat{r}\hat{r}^\top\big) = \frac{\beta}{N}\left(I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\right) \;=\; \frac{\beta}{N}\,P_\perp,$$

where $P_\perp = I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$ is the orthogonal projector onto $\mathbf{1}^\perp$. The Jacobian of the vector field equation 46 at $r = c\mathbf{1}$ is therefore

$$J_{\text{sys}} \;=\; \frac{\lambda^2\beta}{N}\,A(\alpha)^{-1}\,P_\perp \;-\; I. \tag{47}$$

Let $\{v_k\}_{k=1}^N$ be an orthonormal eigenbasis of G with $Gv_k = \mu_k v_k$. By assumption, one eigenvector is $v_1 = \mathbf{1}/\sqrt{N}$ with eigenvalue μ_G, and the remaining $\{v_k\}_{k=2}^N$ span $\mathbf{1}^\perp$.

For any $v_k \in \mathbf{1}^\perp$:

$$P_\perp\, v_k = v_k, \tag{48}$$

$$A(\alpha)\, v_k = \eta_k(\alpha)\, v_k, \quad \text{where } \eta_k(\alpha) = 1 - \alpha(1-\mu_k). \tag{49}$$

By multiplying both sides of equation 47 by v_k we get:

$$J_{\text{sys}}\, v_k \;=\; \left(\frac{\lambda^2\beta}{N\,\eta_k(\alpha)} - 1\right) v_k, \qquad k = 2, \ldots, N. \tag{50}$$

The uniform fixed point is unstable when J_{sys} has a positive eigenvalue. From equation 50, the eigenvalue is maximized for the mode with the *smallest* $\eta_k(\alpha)$, which corresponds to $\mu_{\min}^\perp$:

$$\frac{\lambda^2\beta}{N\,\eta_{\min}^\perp(\alpha)} - 1 > 0 \quad\Longleftrightarrow\quad \eta_{\min}^\perp(\alpha) < \frac{\lambda^2\beta}{N}. \tag{51}$$

Substituting $\eta_{\min}^\perp(\alpha) = 1 - \alpha(1-\mu_{\min}^\perp)$ and solving for α:

$$1 - \alpha(1-\mu_{\min}^\perp) < \frac{\lambda^2\beta}{N} \quad\Longrightarrow\quad \alpha > \frac{1 - \lambda^2\beta/N}{1-\mu_{\min}^\perp} = \alpha_{\text{crit}}\left(1 - \frac{\lambda^2\beta}{N}\right).$$

The requirement $\lambda^2\beta < N$ ensures $\alpha_{\text{crit}}^\lambda > 0$: when $\lambda^2\beta \geq N$, the feedback gain is strong enough that the uniform state is unstable for *all* $\alpha > 0$. □

D EXPERIMENTAL DETAILS

D.1 CONTEXT-AUGMENTED MEMORY SEPARATION (FIGURE 2)

Stored memories. We fix $N = 50$ unit-norm memories in $\mathbb{R}^d$ with $d = 10$ by sampling i.i.d. Gaussian vectors and normalizing:

$$\zeta^\mu \sim \mathcal{N}(0, I_d), \qquad \zeta^\mu \leftarrow \zeta^\mu / \|\zeta^\mu\|_2, \qquad \mu \in [N].$$

The inverse temperature is $\beta = 3.5$.

Query and context noise model. Each trial selects a target index μ uniformly from $[N]$ and draws

$$q = \zeta^\mu + \sigma_q \, \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_d),$$

$$c = \zeta^\mu + \sigma_c \, \eta, \qquad \eta \sim \mathcal{N}(0, I_d), \qquad \sigma_c = 0.3.$$

We sweep σ_q over 12 evenly spaced noise levels in $[0, 3]$.

Metrics and plotting quantities. Retrieval accuracy is the fraction of trials for which $\arg\max_{\nu \in [N]} \hat{r}_\nu = \mu$. For each trial we also compute the effective separation gap

$$\Delta = S_\mu(q; \lambda) - \max_{\nu \neq \mu} S_\nu(q; \lambda)$$

and the target probability $\hat{r}_\mu$. We plot: (i) accuracy vs. σ_q for each $\lambda \in \{0, 1, 2, 4, 8, 16\}$, (ii) a scatter of $(\Delta, \hat{r}_\mu)$ over all trials, overlaid with the exact logistic lower bound

$$\hat{r}_\mu \geq \frac{1}{1 + (N-1)e^{-\beta\Delta}},$$

and (iii) box plots of Δ at fixed noise $\sigma_q \approx 1.0$ (we collect trials with $|\sigma_q - 1.0| < 0.15$). Each (σ_q, λ) pair uses 6000 trials (so $12 \times 6000 \times 5$ total softmax evaluations). Finally, we set $\alpha = 0.1$.

D.2 PHASE TRANSITION IN GATE SELECTIVITY (FIGURE 3)

Single-cluster memory model. Each trial generates a single cluster of $N = 50$ unit-norm memories in $\mathbb{R}^d$ with $d = 10$ as follows. Sample a random centroid $g \sim \mathcal{N}(0, I_d)$, normalize, and scale to norm 2:

$$g \leftarrow 2\, g / \|g\|_2.$$

Then for each $\mu \in [N]$ draw

$$\zeta^\mu = g + \sigma \, \xi^\mu, \qquad \xi^\mu \sim \mathcal{N}(0, I_d), \qquad \sigma = 0.3,$$

and normalize $\zeta^\mu \leftarrow \zeta^\mu / \|\zeta^\mu\|_2$.

Competition matrix, input, and output. Let $G = \zeta^\top \zeta$ and define $W_{h_2} = G - I$ (equivalently, set the diagonal of G to zero). For each penalization strength α we form

$$A(\alpha) = I + \alpha W_{h_2}, \qquad u = \frac{1}{N}\mathbf{1}, \qquad p^*(\alpha) = \mathrm{softmax}\big(\beta s(\alpha)\big),$$

with $\beta = 3.5$. When $A(\alpha)$ is numerically stable (see below), we use the exact fixed point $s(\alpha) = A(\alpha)^{-1}u$.

Estimating the critical point. To estimate $\mu_N = \lambda_{\min}(G)$, we sample 1000 independent clusters, compute the smallest eigenvalue of each Gram matrix G, and report the mean μ_N. We then set

$$\alpha_{\mathrm{crit}} = \frac{1}{1 - \mu_N}.$$

Sweep over α and number of trials. We run 1000 independent trials per α (seeds $0, \ldots, 999$). The sweep uses: (i) a coarse grid of 17 points on $[0, 2]$ and (ii) a dense grid of 60 points between $\min\{\alpha_{\mathrm{crit}}, 1.25\}$ and $\max\{\alpha_{\mathrm{crit}}, 1.25\}$, plus the value α_{crit} itself; duplicates are removed after rounding to 4 decimals.

Stable vs. unstable handling (implementation detail). Let $\lambda_{\min}(A(\alpha))$ be the smallest eigenvalue of $A(\alpha)$. If $\lambda_{\min}(A(\alpha)) > 0.01$, we compute $s(\alpha) = A(\alpha)^{-1}u$. Otherwise, we evaluate the closed-form transient solution of the linear ODE $\dot{s} = u - A(\alpha)s$ at a finite horizon t_{end} (chosen adaptively as per Theorem 3.3), and set $s(\alpha) = s(t_{\text{end}})$; this produces a numerically well-defined proxy that approaches the WTA behavior in the unstable regime.

Reported quantity. For each α, we report the mean peak probability

$$\mathbb{E}\Big[\max_{i\in[N]} p_i^*(\alpha)\Big]$$

over the 1000 trials, and visualize the sharp transition around α_{crit} (with a zoomed-in panel near the transition).

D.3 Coupled Gate-Retrieval Dynamics (Figure 3, Panel d)

Model and coupling mechanism. We extend the gate-only model (Section D.2) by coupling retrieval activations back through the gates via a cross-circuit parameter $\lambda \geq 0$. The gate dynamics settle fast ($\tau_s \ll \tau_r$), yielding the quasi-static fixed point

$$s^*(\alpha, \lambda) = A(\alpha)^{-1}\big(u + \lambda\,\hat{r}\big),$$

where $\hat{r} = \text{softmax}(\beta r)$ are the retrieval probabilities. The retrieval activations r then satisfy the self-consistent equation

$$r^* = \lambda\, A(\alpha)^{-1}u + \lambda^2\, A(\alpha)^{-1}\hat{r}^*.$$

Fixed-point iteration. For each (α, λ) pair, we solve for r^* by iterating

$$r^{(k+1)} = \lambda\, A(\alpha)^{-1}u + \lambda^2\, A(\alpha)^{-1}\text{softmax}(\beta r^{(k)}),$$

initialized with small random noise $r^{(0)} \sim \mathcal{N}(0, \sigma_{\text{init}}^2 I)$ where $\sigma_{\text{init}} = 10^{-4}$. We run up to 500 iterations or until $\|r^{(k+1)} - r^{(k)}\|_\infty < 10^{-8}$. The retrieval probabilities are then $\hat{r}^* = \text{softmax}(\beta r^*)$.

Alpha sweep for panel (d). To capture the smooth transitions across all λ values uniformly, we augment the sweep from Section D.2 with 100 uniformly spaced points on $[0, 2]$. Additionally, for each $\lambda > 0$, we add a dense grid of 60 points near the theoretical $\alpha_{\text{crit}}^{\lambda}$. After removing duplicates (rounded to 4 decimals), this yields approximately 150–180 α values.

Memory generation and trials. Each trial uses the same single-cluster generation protocol (Section D.2), with $N = 50$, $d = 10$, $\sigma = 0.3$, and $\beta = 5.0$. We run 1000 independent trials per α (seeds $0, \ldots, 999$), computing retrieval probabilities for all λ values simultaneously by reusing $A(\alpha)^{-1}$ for efficiency.

Reported quantity. For each (α, λ) pair, we report the mean peak retrieval probability

$$\mathbb{E}\Big[\max_{i\in[N]} p_i^*(\alpha, \lambda)\Big]$$

over the 1000 trials. Panel (d) visualizes how the phase transition shifts leftward (to smaller α) as λ increases.