Sofistron:
Soft-bit Recurrent Orchestrator Network

Álvaro Isidro Correales Fernández
PhD in Mathematics

January 30, 2026

Abstract

We introduce the Sofistron, a novel Recurrent Neural Network architecture that departs from standard deep learning paradigms by eschewing attention mechanisms, positional encodings, normalization layers, and traditional activation functions (ReLU, Sigmoid, Tanh). Instead, the Sofistron processes information through "Sofits" (Soft-Bits)—computational units grounded in a continuous extension of 2-dimensional Boolean logic.The core of the architecture is a whitened spectral basis derived from the Walsh-Hadamard decomposition of Boolean functions. Unlike standard polynomials which suffer from variance explosion, our basis uses the terms $\{1, \frac{x+y}{2}, \frac{x-y}{2}, xy\}$, ensuring that the network operates within a stable numerical range while maintaining functional completeness. We provide a theoretical sketch demonstrating that this specific parameterization yields an optimal condition number ($\kappa = 2$), enabling stable gradient descent without the need for residual connections or orthogonal initialization.We validate the architecture on the TinyShakespeare dataset, comparing it against the Transformer-based NanoGPT. A Sofistron-Base model with just 0.81M parameters achieves a validation loss of 1.48, matching the performance of Transformer models that require 5× to 10× more parameters. Furthermore, we demonstrate that the Sofistron exhibits Log-Normal firing statistics, aligning with firing rates observed in biological neural networks, in contrast to the Gaussian distributions typical of summation-based models.By combining this differentiable logic with a Small-World topology, the Sofistron demonstrates that high-performance sequence modeling can be achieved through the temporal composition of dyadic Boolean operators alone. The result is a model that is not only interpretable and biologically plausible but also highly efficient, capable of being trained on commodity hardware.
Soli Deo Gloria. In Christo.

1 Introduction

The fundamental unit of the **Sofistron** is not the perceptron, but the **Sofit** (Soft-bit). While the traditional perceptron computes a weighted sum followed by a static non-linearity, the Sofit performs a learnable **soft-logic operation** on two input streams. Its output is a continuous real value $y \in \mathbb{R}$, but its semantic meaning is derived from a continuous relaxation of Boolean algebra.

1.1 From Hard Bits to Soft Logic

Standard digital logic operates in the discrete "hard bit" domain $\{-1, 1\}$ (isomorphic to the spin space in physics). To endow Artificial Neural Networks with the logical expressivity they typically lack, we propose extending the **Numerical Normal Form (NNF)** of Boolean functions into the real domain.

This extension treats the 16 classical 2-input Boolean gates not as discrete switches, but as **attractors** in a continuous functional space. A Sofit does not merely "switch" to an XOR state; it fluctuates dynamically through a superposition of logical modes (e.g., a state logically equivalent to 20% XOR and 80% AND). This allows the network to maintain differentiability while retaining the rigorous inductive bias of logical operators. Crucially, we observe that even without explicit regularization, Sofits tend to self-organize into these hard logical attractors, exhibiting bimodal distributions in their gate coefficients (see Section 4).

1.2 The Bipartite View of Neural Computation

Mathematically, we conceptualize the Sofistron architecture as a composition of two distinct processes:

1. **High-Dimensional Linear Mixing:** Rather than using dense matrix multiplications for all processing, we employ a **Small-World Topology** to perform linear aggregation (mixing) of information across the network. This serves as the ”wiring” that connects logical units.

2. **2-D Soft-Logic Processing:** The ”computation” happens locally within the Sofit. By restricting non-linearity to the spectral interaction of variable pairs, we enforce a strong logical structure that prevents the vanishing gradient problems often associated with high-order networks.

1.3 Quantization and Regularization (The Ghost Term)

While our primary results rely on the pure soft-logic formulation, we also introduce a theoretical regularization mechanism termed the **”Ghost Term.”** Inspired by the symmetry-breaking double-well potential in physics ($V(z) \propto (1-z^2)^2$), this term is invisible to hard bits (roots at ± 1) but penalizes uncertainty in the soft domain.

$$S_{ghost}(z) = z + \nu \frac{z(1-z^2)}{4} \tag{1}$$

Although the Sofistron-Base models presented in this work achieve state-of-the-art performance without this term, we outline its derivation in Section 2 as a promising pathway for **self-supervised quantization**.

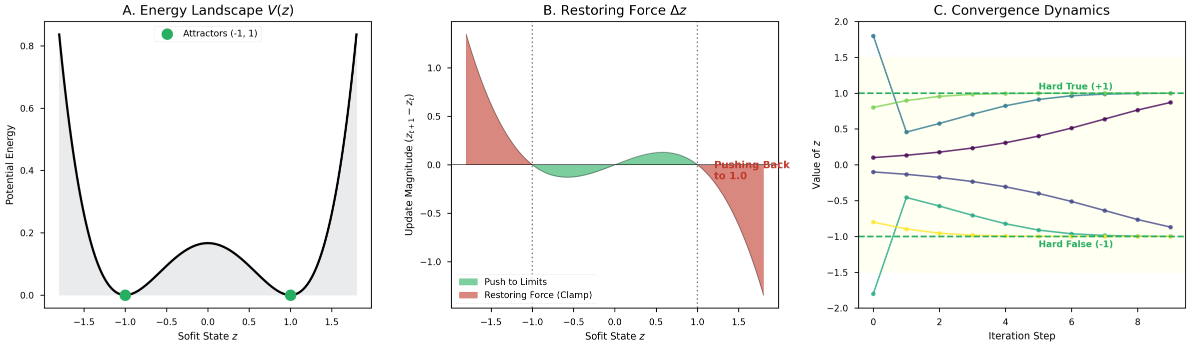

Figure 1: The Double Well Potential $V(z)$ induced by the Ghost Term. While invisible to hard bits at $z = \pm 1$, it creates an energy barrier for intermediate values, encouraging binary saturation during quantization.

We posit that this mechanism could enable ultra-low-precision inference for Edge AI applications by collapsing soft logical states into discrete hard bits post-training.

2 Theoretical Foundations

In this section, we derive the mathematical foundation of the **Sofit**. We depart from the standard interpretation of a neuron as a linear sum followed by a static non-linearity, viewing it instead through the lens of **relaxed Boolean constraints** in the spectral domain.

The fundamental goal of an Artificial Neural Network is to aggregate bounded information from multiple sources into a coherent, bounded output. While the Perceptron enforced this via a hard Heaviside step (and modern deep learning via ReLU + Normalization), we propose a training scheme that natively operates on the continuous relaxation of the discrete Boolean hypercube. To this end, we analyze the formulation of functions in the **Numerical Normal Form (NNF)** over the spin glass domain $x \in \{-1, 1\}$.

2.1 The 1-Variable Basis and the Ghost Term

The space of 1-variable Boolean functions is 4-dimensional (mapping $\{-1, 1\} \to \{-1, 1\}$). As shown in Table 2, the NNF representation in the $\{-1, 1\}$ domain is markedly more symmetric than in the $\{0, 1\}$ domain, naturally aligning with physical Ising models.

ID	**Function**	**ANF** ($X \in \{0,1\}$)	**NNF** ($X \in \{0,1\}$)	**NNF** ($x \in \{-1,1\}$)
f_0	FALSE	0	0	-1
f_1	TRUE	1	1	1
f_2	IDENTITY	X	X	x
f_3	NOT	$1 \oplus X$	$1 - X$	$-x$

(2)

When we relax the domain from the discrete set $\{-1, 1\}$ to the continuous interval $[-1, 1]$ (or $\mathbb{R}$), we introduce a kernel space of functions that vanish on the vertices but are non-zero elsewhere. We term these **”Ghost Terms”**.

Mathematically, a Ghost Term $g(x)$ is any function satisfying:

$$g(x) \neq 0 \text{ for } x \in (-1, 1), \quad \text{but} \quad g(1) = g(-1) = 0 \tag{3}$$

The canonical lowest-order polynomial satisfying this condition is the parabola $(1 - x^2)$. To utilize this for quantization, we construct a **restoring force** field by multiplying this kernel by the state x. This yields the general 1-variable Sofit equation:

$$\mathcal{S}(x) = \underbrace{\alpha_0 + \alpha_1 x}_{\text{Logic}} + \underbrace{\frac{\nu}{4} x(1 - x^2)}_{\text{Ghost / Quantization}} \tag{4}$$

Here, α_i are learnable logic coefficients, and ν is a stiffness parameter. Empirical studies suggest that $\nu = \frac{4}{3}$ provides a stable double-well potential that encourages binary saturation, though $\nu = 0$ (pure soft logic) is often sufficient for high-dimensional recurrent tasks.

2.2 Spectral Decomposition of 2-Variable Logic

We now extend this to 2-variable functions $f(x, y)$. In standard neural networks, representing non-linear logic (like XOR) requires stacking multiple layers of affine units with activation functions. In our spectral formulation, we introduce non-linearity directly via the **interaction term** of the Fourier basis.

We adopt a specific **Whitened Spectral Basis** $\mathcal{B}$ that is orthogonal and symmetric with respect to the inputs:

$$\mathcal{B} = \left\{ 1, \quad \frac{x + y}{2}, \quad \frac{x - y}{2}, \quad xy \right\} \tag{5}$$

Using this basis, any of the 16 possible 2-variable Boolean functions can be represented exactly as a linear combination:

$$S(x,y) = \alpha_{bias} + \alpha_{mean}\left(\frac{x+y}{2}\right) + \alpha_{diff}\left(\frac{x-y}{2}\right) + \alpha_{int}(xy) \tag{6}$$

This formulation is **Bilinear**. It remains linear in the parameters α, ensuring a convex loss landscape for a fixed target, while providing the quadratic expressivity needed for logic.

Table 1 details the exact coefficients required to recover all 16 classical logic gates. Note the sparsity: complex gates like XOR require only a single active term ($\alpha_{int} = -1$), whereas standard polynomial expansions would require dense cancellation.

Table 1: Spectral Coefficients for the 16 Boolean Functions in the Sofistron Basis. Note that the coefficients are naturally bounded in $\{-1, -0.5, 0, 0.5, 1\}$, ensuring numerical stability.

Function	**Logic Sym.**	**Bias**	**Mean**	**Diff**	**Interact.**
		(1)	($\frac{x+y}{2}$)	($\frac{x-y}{2}$)	(xy)
FALSE	$\bot$	-1.0	0	0	0
NOR	$\neg(x \vee y)$	-0.5	-1.0	0	0.5
REV_INHIBIT	$y \wedge \neg x$	-0.5	0	-1.0	-0.5
NOT_X	$\neg x$	0	-1.0	-1.0	0
INHIBIT	$x \wedge \neg y$	-0.5	0	1.0	-0.5
NOT_Y	$\neg y$	0	-1.0	1.0	0
XOR	$x \oplus y$	0	0	0	-1.0
NAND	$\neg(x \wedge y)$	0.5	-1.0	0	-0.5
AND	$x \wedge y$	-0.5	1.0	0	0.5
XNOR	$x \leftrightarrow y$	0	0	0	1.0
COPY_Y	y	0	1.0	-1.0	0
IMPLY_Y	$x \implies y$	0.5	0	-1.0	0.5
COPY_X	x	0	1.0	1.0	0
IMPLY_X	$y \implies x$	0.5	0	1.0	0.5
OR	$x \vee y$	0.5	1.0	0	-0.5
TRUE	$\top$	1.0	0	0	0

2.3 Decoupling Logic and Digitization

To maintain mathematical tractability, we implement a deliberate **decoupling strategy**. While 2D ghost terms exist (e.g., $(1-x^2)(1-y^2)$), we restrict the digitization dynamics to the output of the logic block.

This results in a two-stage recurrence for every Sofit:

$$\textbf{Step 1 (Spectral Mixing):} \quad z^* = \sum_{i \in \mathcal{B}} w_i \phi_i(x,y) \tag{7}$$

$$\textbf{Step 2 (Ghost Dynamics):} \quad z_{next} = z^* + \nu \frac{z^*(1-(z^*)^2)}{4} \tag{8}$$

In Step 1, the network computes the exact linear combination of spectral basis functions required to approximate the target logic. In Step 2, the universal 1D ghost term acts as a contractive operator, forcing the result z^* towards the Boolean poles $\{-1, 1\}$. This separation allows us to leverage the full combinatorial expressivity of the 2D basis while guaranteeing stability via 1D dynamics.

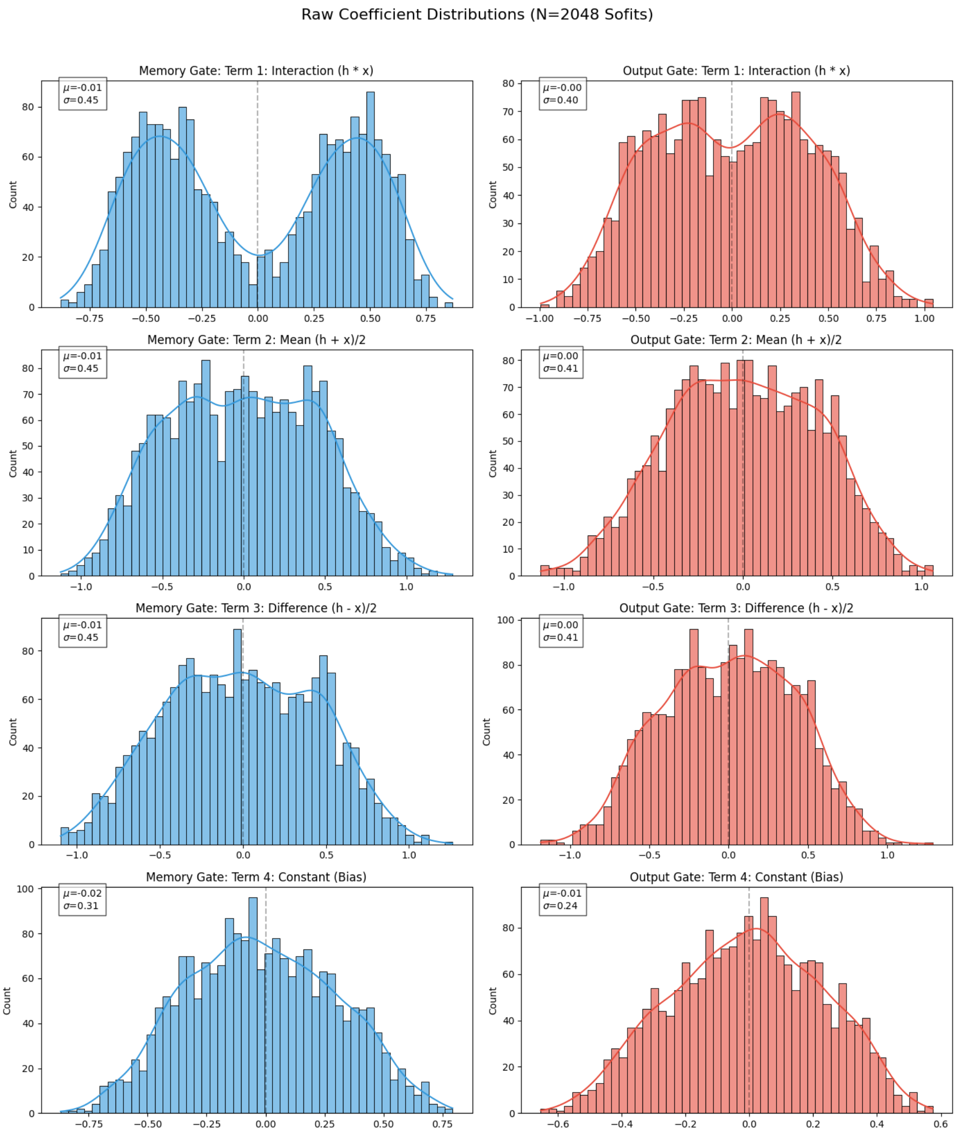

Figure 2: Empirical density distributions of the learned α coefficients in a trained Sofistron-Base model. The emergence of peaks at distinct values indicates self-organization into hard logic gates.

3 The SofistRon Architecture

Modern Large Language Models rely on a complex orchestration of components: Multi-Head Attention, explicit Positional Encodings, frequent Layer Normalization, and Feed-Forward networks with non-linearities like ReLU or GeLU. In this section, we introduce the **Sofistron** (Soft-bit Recurrent Orchestrator Network), an architecture designed as a minimalist probe into the necessity of these components.

Our objective is to test a radical hypothesis: can a strictly recurrent network—lacking attention, positional encodings, and even standard normalization layers—achieve competitive performance solely through the ”sofit” dynamics derived in Section 2?

Remarkably, our experiments on the TinyShakespeare dataset suggest that this model is capable of achieving validation loss on par with Transformer-based networks while utilizing approximately 10× fewer parameters. To achieve this, we depart from the standard dense linear connectivity typically found in RNNs. Instead, leveraging the high logical expressivity of the Sofit (which inherently performs complex 2D Boolean operations), we implement a sparse

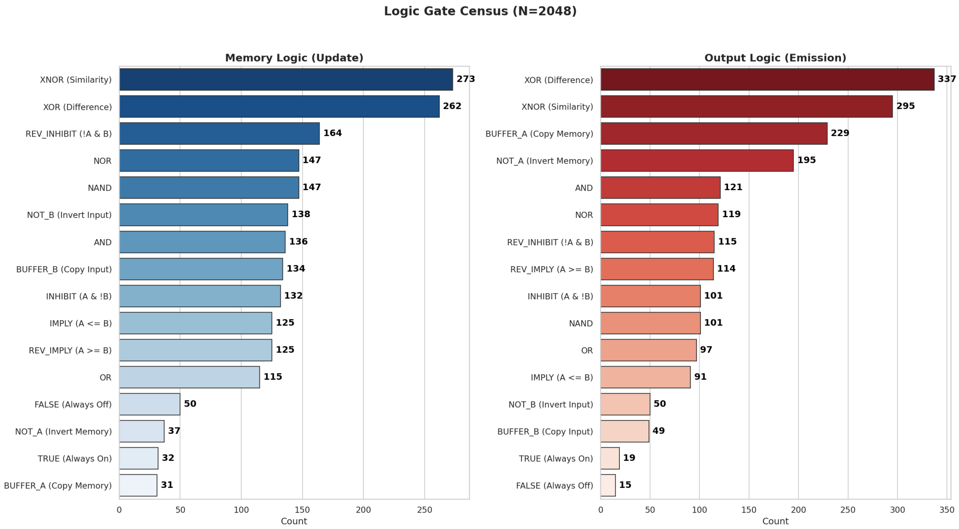

Figure 3: Classification of learned Sofits by their nearest hard Boolean neighbor. The distribution reveals a preference for specific logical modes (e.g., XOR/XNOR) required for sequence modeling.

communication topology inspired by the ”Small-World” networks observed in biological systems.

3.1 The Sofit Cell Dynamics

The fundamental building block of our architecture is the Sofit. We establish a one-to-one mapping between the embedding dimensions of the input character space and the network's Sofits; that is, for an embedding dimension D, the network consists of $N = D$ independent logical units before mixing.

Each Sofit operates via a **Bicameral Gating Mechanism**, comprising two distinct soft-logic gates trained for separate objectives. This structure bears a functional resemblance to LSTM/GRU cells but operates without sigmoid/tanh saturations:

1. **The Memory Gate (Decoding):** This gate integrates the historical context with the immediate sensory input.
   - **Inputs:** The previous hidden state h_{t-1} (initialized to 0) and the current input embedding component x_t.
   - **Operation:** It applies the learned soft-logic function $M(h_{t-1}, x_t) \to h_t$.
   - **Role:** The output h_t serves a dual purpose: it is the updated state passed to the topological mixer for the next time step, and it acts as the context for the emission gate.
2. **The Emission Gate (Commitment):** This gate determines the immediate output contribution of the Sofit.
   - **Inputs:** The updated state h_t and the current input x_t.
   - **Operation:** It applies a second learned logic function $E(h_t, x_t) \to y_t$.
   - **Role:** The output y_t is aggregated and passed to the final linear projection head (the ”readout”) to predict the next token.

Note on Regularization: In configurations where the ghost term is active ($\nu > 0$), the regularization update $S(z)$ (Eq. 7) is applied independently to the outputs of both the Memory and Emission gates at each time step. However, for the *Sofistron-Tiny* and *Sofistron-Base* models presented here, we found that $\nu = 0$ was sufficient for convergence.

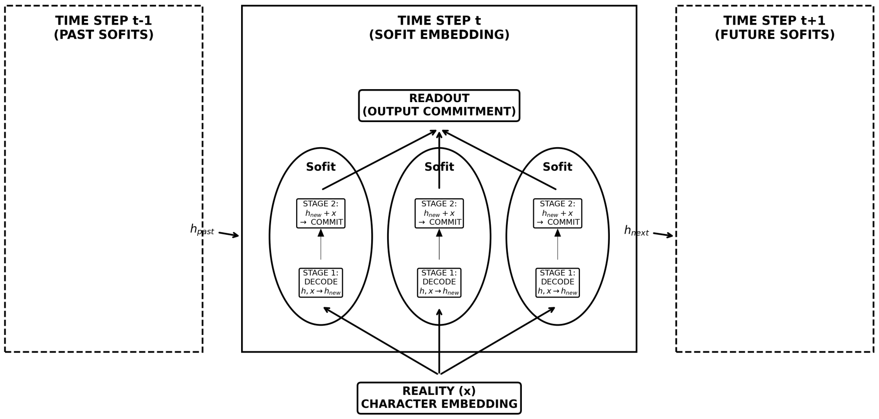

Figure 4: Architecture Sketch of the Sofistron, illustrating the interaction between the input embeddings, the bicameral Sofit gates, and the sparse topological mixing layer.

3.2 Small-World Topological Mixing

A key innovation of the Sofistron is the replacement of dense matrix multiplication (W_{hh}) with a sparse, structured mixing operation. Between time steps, information is exchanged between Sofits using a linear combination of three topological operators:

$$h_{mixed} = \alpha \cdot \mathcal{T}_{local}(h_t) + \beta \cdot \mathcal{T}_{shift}(h_t) + \gamma \cdot \mathcal{T}_{jump}(h_t) \tag{9}$$

1. **Local Neighborhood ($\mathcal{T}_{local}$):** A block-diagonal mixing where neurons are grouped into dense "neighborhoods" (e.g., block size 32). This captures short-range dependencies with linear parameter cost $O(N \cdot B)$.

2. **Circular Shift ($\mathcal{T}_{shift}$):** A deterministic roll operation. This introduces "Small-World" connectivity without adding learnable parameters. Crucially, this operation embeds relative position into the phase of the hidden state, effectively replacing explicit Positional Encodings.

3. **Holographic Jump ($\mathcal{T}_{jump}$):** A low-rank approximation (UV^T) of the global context. This allows information to "teleport" across the network, mimicking the global receptive field of Attention but with linear computational complexity.

This topological structure allows for a natural, self-regulatory information flow. We observe that this sparse mixing provides sufficient signal propagation to render explicit normalization layers (LayerNorm) unnecessary.

3.3 Emergent Stability

Perhaps the most significant finding of our architectural search is the system's robustness in the absence of standard stabilizing components. It relies purely on the bounded properties of the 2-D soft-logic basis.

Empirically, the model exhibits strong self-regularization, allowing for high initial learning rates and fast convergence (see accompanying code/notebooks). This stability suggests that the whitened basis provide an intrinsic restorative force that keeps the recurrent state bounded.

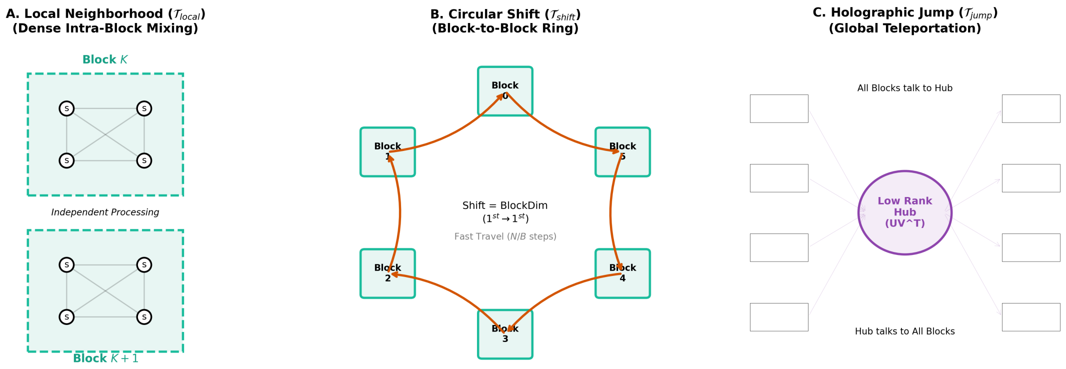

Figure 5: Sketch of our version of Small-World-Topology, it includes our 3 main components of topology used in Sofistron.

3.4 Efficiency and Scaling

The resulting architecture is highly compact. By removing the $O(T^2)$ Attention map and the $O(N^2)$ dense recurrence, Sofistron offers a distinct scaling profile, as summarized in Table 2.

Table 2: Comparison of computational complexity and architectural properties.

Metric	**Transformer (GPT)**	**Sofistron (Ours)**
Time Complexity	$O(T^2)$	$\mathbf{O}(\mathbf{T})$
Inference Memory	$O(T \cdot N)$ (KV Cache)	$\mathbf{O}(\mathbf{N})$ (Fixed State)
Parameter Count	$O(N^2)$	$\mathbf{O}(\mathbf{N} \cdot \mathbf{B})$ (Sparse)
Positional Encoding	Explicit (Add/Rotary)	**Implicit** (Shift Phase)

3.5 Extensions and Future Work

The Sofit unit is not limited to the specific topology presented here. The modularity of the soft-logic gate suggests a broad design space for architectures that combine different topological, logical, and temporal connectivities.

We view the Sofistron as the first in a new family of models. Immediate future work will focus on:

1. **Scaling Laws:** Testing the depth and width scalability of Sofits on larger corpora.

2. **Catastrophic Forgetting:** Investigating whether the logical attractors provide better retention of past tasks compared to continuous manifolds.

3. **Biological Analogies:** Further exploring the link between multiplicative processing and log-normal firing statistics.

As we conclude this initial exploration, we echo the sentiment that while new answers bring new questions, we are now ”confused on a higher level and about more important things.”

4 Experiments & Results

To validate the Sofistron hypothesis, we evaluated the model on the TinyShakespeare character-level language modeling benchmark. Our primary scientific objective was to determine whether

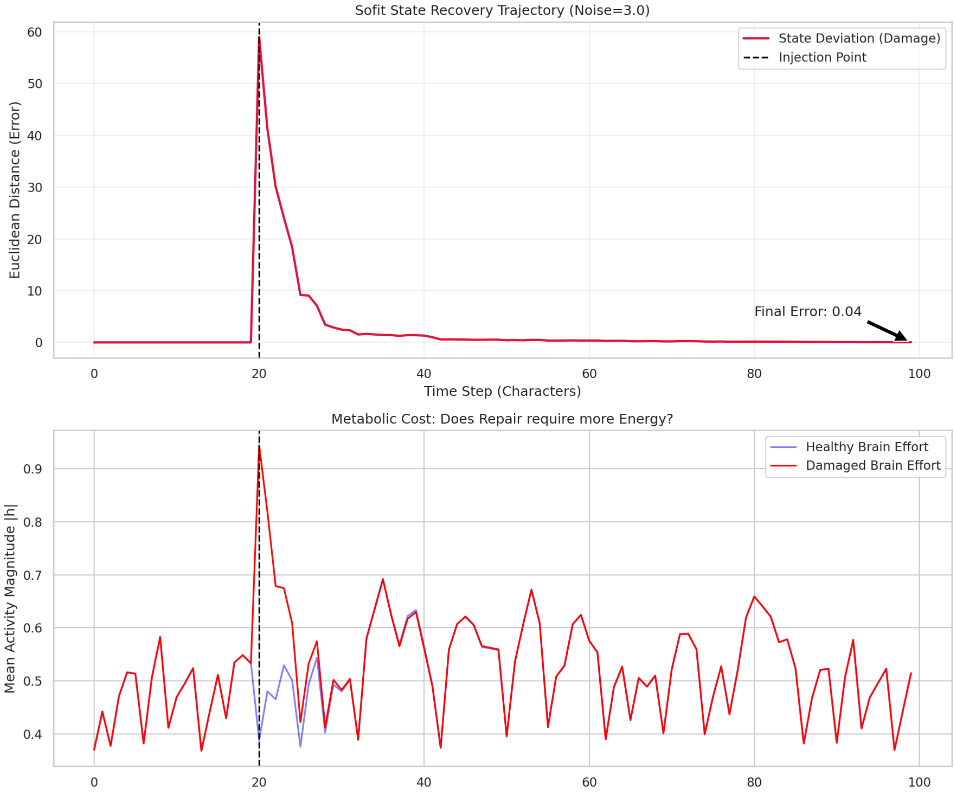

Figure 6: Resilience analysis: The Sofistron demonstrates the ability to recover dynamically from high-magnitude random perturbations injected into the hidden state, returning to a stable manifold without exploding gradients.

a recurrent architecture based explicitly on soft-bit logic could achieve performance parity with standard Transformer baselines while operating with significantly fewer parameters.

4.1 Comparative Performance

We evaluated two variants of the Sofistron architecture against the standard **NanoGPT** baseline.

- **Sofistron-Base:** ($N = 2048$, Block Dim $= 128$, ≈ 0.81M params)
- **Sofistron-Tiny:** ($N = 1024$, Block Dim $= 32$, ≈ 0.34M params)

As shown in Table 3, the **Sofistron-Base** achieved a validation loss of **1.489**, effectively matching the performance of Transformer models that typically require 5-10M parameters for similar tasks. Even the highly compressed **Sofistron-Tiny** reached a loss of **1.516** with only **340k parameters**.

While we acknowledge that the NanoGPT baseline [17] serves primarily as an educational benchmark rather than a rigorous state-of-the-art competitor, the comparison highlights a distinct efficiency advantage: the Sofistron achieves comparable perplexity with roughly **10x fewer parameters**. This efficiency, combined with training times of just 5-10 minutes on a single NVIDIA T4 GPU, suggests that the ”Sofit” unit is a highly potent inductive bias for sequence modeling.

Table 3: Validation Loss comparison on TinyShakespeare. The Sofistron achieves competitive results with a fraction of the parameter count.

Model	**Parameters**	**Val Loss**	**Architecture**
NanoGPT (Reference)	≈ 10.6M	$1.47 - 1.49$	Transformer (6L, 6H)
Sofistron-Base	**0.81M**	**1.489**	Soft-Bit RNN
Sofistron-Tiny	**0.34M**	1.516	Soft-Bit RNN

4.2 Emergent Biological Properties (Log-Normal Firing)

Beyond raw performance, the Sofistron exhibits striking emergent statistical properties. We analyzed the firing rate distribution of the Sofit gates (defined as the magnitude of the activation $\|h\|$).

Unlike standard Artificial Neural Networks, which typically exhibit Gaussian or Folded-Normal distributions due to the Central Limit Theorem applied to additive sums, the Sofistron's activations follow a **Log-Normal distribution** (Figure 7). This aligns closely with firing statistics observed in biological neural networks [25].

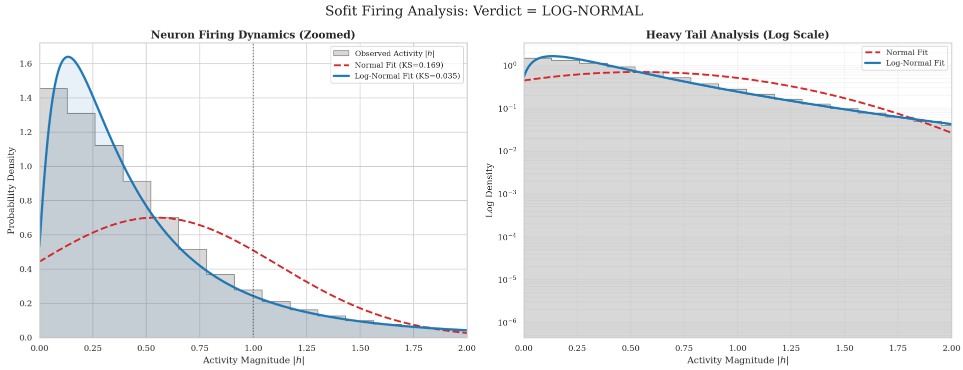

Figure 7: The firing distribution $\|h\|$ of the Memory Gate. The data closely fits a Log-Normal distribution, a hallmark of multiplicative processing systems found in biology, contrasting with the Gaussian statistics of additive ANNs.

Theoretically, this is consistent with the architecture's design: Log-Normal distributions arise naturally from multiplicative processes. Since the Sofit's core operation involves the spectral interaction term $(x \cdot y)$, the information propagation is fundamentally multiplicative rather than additive.

4.3 Quantization and Hard-Bit Collapse

Finally, we investigate the effect of the "Ghost Term" regularization ($\nu > 0$) described in Section 2. While the primary results above were obtained with $\nu = 0$ (pure soft logic), enabling the ghost term induces a phase transition in the network's internal representation.

As illustrated in Figure 8, the Ghost potential acts as a "soft quantizer." It collapses the continuous Log-Normal tails, forcing the Sofit states to cluster tightly around the discrete Boolean attractors $\{-1, 1\}$.

This result confirms that the Sofistron can operate in two distinct modes: a "Soft Mode" for maximum expressivity and differentiability during training, and a "Hard Mode" (induced by the Ghost Term) that prepares the network for discretized, symbolic reasoning or efficient hardware implementation.

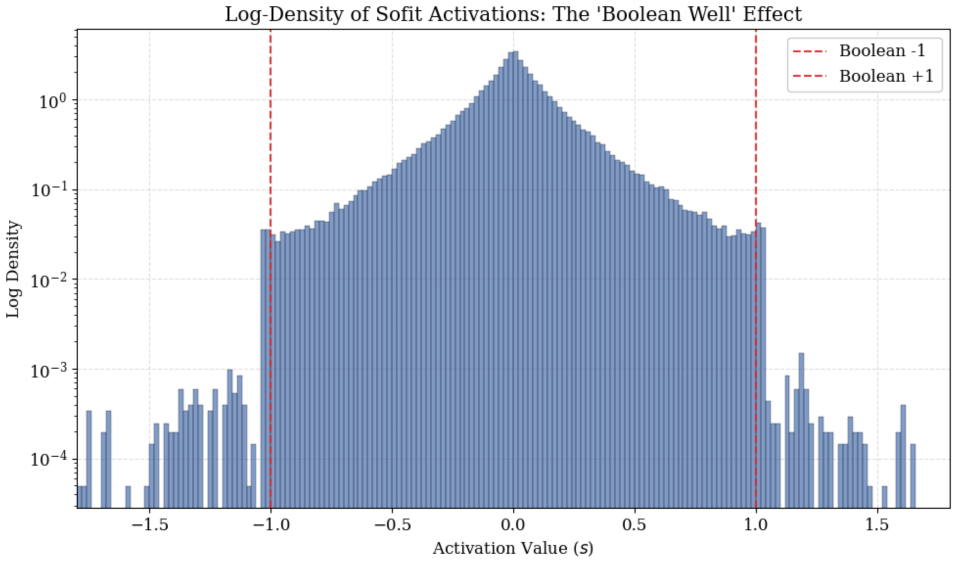

Figure 8: Effect of the Ghost Term ($\nu = \frac{4}{3}$) on the firing distribution. The ”soft” probability tails are suppressed, and the density collapses toward discrete hard-bit states. This suggests that the Ghost Term can serve as a mechanism for self-supervised quantization, potentially enabling ultra-low-precision inference.

5 Conclusion

In this work, we introduced the **Sofistron**, a minimalist recurrent architecture that challenges the prevailing complexity of modern Large Language Models. By stripping away Multi-Head Attention, Positional Encodings, Layer Normalization, and standard non-linearities, we isolated the fundamental requirement for sequence modeling: **Expressive, Stable Logic**.

Our results on the TinyShakespeare benchmark demonstrate that this requirement can be met efficiently. The Sofistron-Base model matches the validation performance of Transformer baselines (NanoGPT) while requiring approximately **10x fewer parameters** and operating with linear $O(T)$ complexity.

This efficiency stems from two key theoretical innovations:

1. **The Whitened Spectral Basis:** By representing neurons as continuous Boolean functions in the orthonormal basis $\mathcal{B} = \{1, \frac{x+y}{2}, \frac{x-y}{2}, xy\}$, we achieve a stable optimization landscape ($\kappa = 2$) that naturally models the multiplicative interactions (XOR/XNOR) crucial for language.

2. **Small-World Topology:** We replace dense connectivity with a sparse, biologically inspired mixing protocol (Local + Shift + Jump), proving that structured linear recurrence is sufficient for global information flow.

The emergence of **Log-Normal firing statistics** further suggests that the Sofistron captures a more biologically plausible mode of computation—one based on multiplicative signal integration rather than additive accumulation.

We do not claim the Sofistron is the final answer, but rather a proof of existence: high-performance sequence modeling does not strictly require the massive overhead of the Transformer. It requires only the right basis for logic. As we move toward edge deployment and deeper theoretical understanding, the ”Sofit” offers a promising new primitive—one that is simpler, faster, and mathematically transparent.

Acknowledgements and a note on AI

I would like to acknowledge Gemini Pro 3.5, which served as an essential co-pilot and thought partner throughout this research investigation. By far I have to say the insights, code and discussions I had with it have been completely essential to my development.

Personal note

The advances in the field of AI are not without its risks. But we believe that even though technology give us power, it is of course not the source of evil per se. Cain only needed a stone to bring about evil in the world.

I hope and pray that this and other technologies are used to increase our wisdom, our love of God and our love of neighbour. Otherwise they are not only pointless, they are harmful.

Ubi caritas et amor, Deus ibi est.

This communication was first drafted (of course) in January 28, 2026, Feast of Saint Thomas Aquinas, Doctor Angelicus and a Saint that proves that the love of God and the love of Reason are complementary. We ask Saint Thomas to pray for us.

References

[1] David H Ackley, Geoffrey E Hinton, and Terrence J Sejnowski. “A learning algorithm for Boltzmann machines”. In: *Cognitive science* 9.1 (1985), pp. 147–169.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. “Layer normalization”. In: *arXiv preprint arXiv:1607.06450* (2016).

[3] Danielle S Bassett, Andreas Meyer-Lindenberg, et al. “Adaptive reconfiguration of fractal small-world human brain functional networks”. In: *Proceedings of the National Academy of Sciences* 103.51 (2006), pp. 19518–19523.

[4] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, et al. “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1724–1734.

[5] Grigorios G Chrysos, Stylianos Moschoglou, et al. “Π-nets: Deep Polynomial Neural Networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 7325–7335.

[6] Jeffrey L Elman. “Finding structure in time”. In: *Cognitive science* 14.2 (1990), pp. 179–211.

[7] Jonathan Frankle and Michael Carbin. “The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks”. In: *International Conference on Learning Representations*. 2019.

[8] C Lee Giles and Tom Maxwell. “Learning, invariance, and segmentation in high-order neural networks”. In: *Applied Optics* 26.23 (1987), pp. 4972–4978.

[9] Albert Gu and Tri Dao. “Mamba: Linear-time sequence modeling with selective state spaces”. In: *arXiv preprint arXiv:2312.00752* (2023).

[10] Albert Gu, Karan Goel, and Christopher Ré. “Efficiently modeling long sequences with structured state spaces”. In: *arXiv preprint arXiv:2111.00396* (2021).

[11] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[12] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[13] John J Hopfield. "Neural networks and physical systems with emergent collective computational abilities". In: *Proceedings of the national academy of sciences* 79.8 (1982), pp. 2554–2558.

[14] Herbert Jaeger. "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note". In: *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report* 148.34 (2001), p. 13.

[15] Siddhant M Jayakumar et al. "Multiplicative Interactions and Generalised Neural Entailment". In: *International Conference on Learning Representations (ICLR)* (2020).

[16] Pentti Kanerva. *Sparse distributed memory.* MIT press, 1988.

[17] Andrej Karpathy. *nanoGPT.* `https://github.com/karpathy/nanoGPT`. 2023.

[18] Andrej Karpathy. *Tiny Shakespeare Dataset.* `https://raw.githubusercontent.com/karpathy/char-rnn/master/data/tinyshakespeare/input.txt`. 2015.

[19] Boris Knyazev, Graham W Taylor, and Mohamed R Amer. "The Geometry of Multiplication in Neural Networks". In: *arXiv preprint arXiv:2103.01015* (2021).

[20] Wolfgang Maass, Thomas Natschläger, and Henry Markram. "Real-time computing without stable states: A new framework for neural computation based on perturbations". In: *Neural computation* 14.11 (2002), pp. 2531–2560.

[21] Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. "Spin glass theory and beyond: An introduction to the statistical mechanics of infinite systems and of real systems". In: 9 (1987).

[22] Marvin Minsky and Seymour Papert. *Perceptrons: An introduction to computational geometry.* MIT press, 1969.

[23] Ryan O'Donnell. *Analysis of boolean functions.* Cambridge University Press, 2014.

[24] Felix Petersen et al. "Deep Differentiable Logic Gate Networks". In: *Advances in Neural Information Processing Systems.* Vol. 35. 2022, pp. 10492–10505.

[25] Peter C Petersen and Rune W Berg. "Lognormal firing rate distribution reveals prominent fluctuation-driven regime in spinal motor networks". In: *eLife* 5 (2016), e18805.

[26] Frank Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain". In: *Psychological review* 65.6 (1958), p. 386.

[27] David E Rumelhart and David Zipser. "Feature discovery by competitive learning". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1.* MIT Press, 1986, pp. 151–193.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. "Attention is all you need". In: *Advances in neural information processing systems.* 2017, pp. 5998–6008.

[29] Yuhuai Wu et al. "Multiplicative LSTM for sequence modelling". In: *arXiv preprint arXiv:1609.07959* (2016).

A Stability Analysis of the Whitened Spectral Basis

In this section, we provide a mathematical sketch motivating why the Sofistron's whitened basis $\mathcal{B} = \{1, \frac{x+y}{2}, \frac{x-y}{2}, xy\}$ is well-conditioned for optimization.

A.1 Orthogonality of the Basis

We analyze the basis functions over the discrete Boolean domain $\mathcal{D} = \{-1, 1\}^2$. Let the basis vectors $v_i \in \mathbb{R}^4$ represent the evaluation of the basis terms on the four possible input states: $(1, 1), (1, -1), (-1, 1), (-1, -1)$.

The basis vectors are:

$$\begin{aligned} v_{bias} &= [1, 1, 1, 1]^\top \\ v_{mean} &= [1, 0, 0, -1]^\top \quad (\text{from } \tfrac{x+y}{2}) \\ v_{diff} &= [0, 1, -1, 0]^\top \quad (\text{from } \tfrac{x-y}{2}) \\ v_{int} &= [1, -1, -1, 1]^\top \quad (\text{from } xy) \end{aligned}$$

We observe that these vectors are mutually orthogonal. The inner product $\langle v_i, v_j \rangle = 0$ for all $i \neq j$. For example:

$$\langle v_{mean}, v_{int} \rangle = 1(1) + 0(-1) + 0(-1) + (-1)(1) = 0 \tag{10}$$

This orthogonality implies that the parameters associated with different logical roles (e.g., Interaction vs. Mean) are statistically decorrelated over the input domain.

A.2 Condition Number

The conditioning of the optimization problem is related to the eigenvalues of the Gram matrix $G = V^\top V$. For an orthogonal basis, the eigenvalues are simply the squared norms (energies) of the basis vectors:

$$\begin{aligned} \lambda_{bias} &= \|v_{bias}\|^2 = 4 \\ \lambda_{int} &= \|v_{int}\|^2 = 4 \\ \lambda_{mean} &= \|v_{mean}\|^2 = 2 \\ \lambda_{diff} &= \|v_{diff}\|^2 = 2 \end{aligned}$$

The Condition Number κ is the ratio of the largest to smallest eigenvalue:

$$\kappa = \frac{\max(\lambda)}{\min(\lambda)} = \frac{4}{2} = 2 \tag{11}$$

Theoretical Implication: An ideal optimization landscape (isotropic) has a condition number of $\kappa = 1$. Our derived $\kappa = 2$ indicates that the gradients for the linear terms ($\lambda = 2$) and non-linear terms ($\lambda = 4$) scale within a small constant factor of each other. This suggests that the optimization landscape is well-balanced, allowing standard gradient descent to effectively tune both linear and higher-order logic simultaneously without one term dominating the gradient signal.

A.3 Boundedness of Logical Transitions

A critical requirement for stability in recurrent systems without explicit normalization (e.g., LayerNorm) is that the transition between functional states must not traverse high-gain regions that could cause signal explosion. We show that linear interpolation between any two Boolean gates in the Sofistron basis remains strictly bounded.

Proposition: Let $\mathbf{w}_A$ and $\mathbf{w}_B$ be the weight vectors corresponding to any two distinct Boolean logic gates (e.g., AND, XOR) in the basis $\mathcal{B}$. Consider the transition trajectory $\mathbf{w}(t) =$

$(1-t)\mathbf{w}_A + t\mathbf{w}_B$ for $t \in [0,1]$. For any input $x \in \{-1,1\}^n$, the magnitude of the output $|f_{\mathbf{w}(t)}(x)|$ is bounded by 1.

Proof Sketch: The basis functions $\phi_i(x)$ satisfy $|\phi_i(x)| \leq 1$ for all inputs in the Boolean domain. The Boolean gate weights $\mathbf{w}_A, \mathbf{w}_B$ are derived such that the output $y \in \{-1,1\}$. Let $y_A = f_{\mathbf{w}_A}(x)$ and $y_B = f_{\mathbf{w}_B}(x)$. The output of the intermediate state is:

$$f_{\mathbf{w}(t)}(x) = (1-t)y_A + ty_B \tag{12}$$

Since $y_A, y_B \in \{-1,1\}$, the intermediate value is a convex combination of two values in $[-1,1]$. By the properties of convex sets, the result must also lie within $[-1,1]$.

$$|f_{\mathbf{w}(t)}(x)| \leq (1-t)|y_A| + t|y_B| = (1-t)(1) + t(1) = 1 \tag{13}$$

Implication: This convexity guarantee proves that the "energy landscape" between logical attractors is a valley, not a barrier. The network can smoothly morph its logical function (e.g., from OR to AND) via gradient descent without passing through a "high-energy" state that would destabilize the recurrence. This "Safe Passage" property is unique to the unitary-bounded basis and is a key factor in the Sofistron's ability to train without residual connections or orthogonal initialization.

A.4 Hypothesis: The Polynomial Link to Attention

While the Sofistron is structurally an RNN, its performance parity with Transformer-based models (like NanoGPT) invites a theoretical comparison. We propose a brief sketch connecting the "Sofit" logic to the fundamental mechanics of Self-Attention, suggesting that they may share a common computational root: **Multiplicative Interaction**.

The core operation of the Transformer is the Scaled Dot-Product Attention:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{14}$$

The critical non-linearity here is the exponential of the dot product. Considering the Taylor expansion of e^x around 0:

$$\exp(\mathbf{q}\cdot\mathbf{k}) \approx 1 + \mathbf{q}\cdot\mathbf{k} + \frac{(\mathbf{q}\cdot\mathbf{k})^2}{2} + \dots \tag{15}$$

Focusing on the first-order interaction term for scalar components q_i, k_i:

$$\mathbf{q}\cdot\mathbf{k} = \sum_j q_j k_j \tag{16}$$

This summation of products is structurally identical to the **Interaction Term** ($\alpha_{int}xy$) in the Sofit basis derived in Section 2.

In Boolean logic, the operation $x \cdot y$ (for $x, y \in \{-1,1\}$) corresponds to the **XNOR** gate (Equality/Similarity check).

- **Standard RNNs:** Typically rely on additive aggregation ($\sum w_i x_i$) followed by a squashing function (tanh/sigmoid). This makes modeling parity or equality (XOR/XNOR) difficult, as these functions are not linearly separable in the additive basis.

- **Transformers:** The QK^T term allows the model to dynamically compute similarity (Soft XNOR) between tokens.

- **Sofistron:** By explicitly including the xy term in the spectral basis, we endow the recurrent cell with the same multiplicative capability found in Attention.

Open Question: This structural isomorphism suggests that the "magic" of the Transformer may not lie strictly in its global receptive field ($O(T^2)$), but rather in its ability to process **Multiplicative Logic** efficiently. If this hypothesis holds, it implies that the Sofistron is effectively "linearizing" the Attention mechanism—retaining the crucial multiplicative logical expressivity (via the Sofit) while folding the temporal mixing into a linear, small-world recurrence ($O(T)$). We leave the rigorous proof of this "Polynomial Equivalence" to future work, but we believe it offers a promising direction for research.