Gradual Forgetting: Logarithmic Compression for Extending Transformer Context Windows

Billy Dickson[1] and Zoran Tiganj[1,2]

[1]Department of Computer Science, Indiana University Bloomington
[2]Department of Psychological and Brain Sciences, Indiana University Bloomington
{dicksonb, ztiganj}@iu.edu

Abstract

Most approaches to long-context processing increase the complexity of the transformer's internal architecture by integrating mechanisms such as recurrence or auxiliary memory modules. In this work, we introduce an alternative approach that modifies the input representation itself, rather than the transformer architecture. Inspired by cognitive models of human memory, our method applies a scale-invariant logarithmic compression to the input tokens. The resulting compressed representation is processed by a standard, unmodified transformer, preserving architectural simplicity. We evaluate this approach on the WikiText-103 and PG-19 language modeling benchmarks, showing a reduction in perplexity compared to uncompressed baselines. Moreover, performance improves consistently with longer compressed temporal contexts, showing that input-level logarithmic compression is a simple and effective way to extend a transformer's long-range memory.

1 Introduction

Transformers have become a dominant architecture for sequence modeling across domains such as language modeling, time-series forecasting, and dialog systems (Vaswani et al., 2017; Lim et al., 2021; Devlin et al., 2019), yet their ability to process long sequences is constrained by the quadratic complexity of self-attention (Child et al., 2019; Beltagy et al., 2020). Existing solutions typically modify the model's architecture, employing segment-level recurrence (Dai et al., 2019; Rae et al., 2019; Bulatov et al., 2022), external memory modules (Graves et al., 2014; Weston et al., 2015; Ko et al., 2024; Kang et al., 2025), or sparse and approximate attention mechanisms (Wu et al., 2022; Child et al., 2019; Beltagy et al., 2020; Zaheer et al., 2020; Kitaev et al., 2020; Choromanski et al., 2021; Wang et al., 2020), which often increases complexity and introduces state dependencies. In contrast, our work proposes a scale-invariant input transformation that compresses the input history before it reaches a standard transformer, requiring no architectural changes. Inspired by cognitive models of human memory that posit a logarithmic encoding of temporal information (Shankar & Howard, 2012; Howard et al., 2014; Tano et al., 2020; Findling et al., 2025; Tiganj et al., 2019; De Vries & Principe, 1992; Grossberg & Schmajuk, 1989) and their applications in deep neural networks (Jacques et al., 2021, 2022; Mochizuki-Freeman et al., 2024), our method uses a bank of unimodal temporal filters to produce a log-compressed memory of the distant past. This compressed representation is concatenated with recent uncompressed tokens, preserving compatibility with existing attention mechanisms and supporting stateless batching. Evaluation on the WikiText-103 and PG-19 benchmarks shows that this approach improves perplexity over uncompressed baselines, with performance increasing as the compressed memory length grows.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: CogInterp: Interpreting Cognition in Deep Learning Models.

2 Model

We use a log-spaced bank of scale-invariant filters with impulse response $\Phi(t, \overset{*}{\tau}) = \frac{k^{k+1}}{k!}\left(\frac{t}{\overset{*}{\tau}}\right)^k e^{-kt/\overset{*}{\tau}}$, where k controls width and the peaks $\overset{*}{\tau}_i$ are geometrically spaced (Fig. 1), yielding rescaled copies that tile log-time. Let $f(t) \in \mathbb{R}^d$ denote the token embedding at discrete step t. The compressed representation at time t is obtained by a causal, depth-wise 1-D convolution:

$$\tilde{f}(t, \overset{*}{\tau}) = \sum_{t'=1}^{M} \Phi\big(t', \overset{*}{\tau}\big)\, f(t - t'). \tag{1}$$

For each scale $\overset{*}{\tau}$, $\tilde{f}(t, \overset{*}{\tau})$ is a smoothed, lagged estimate of $f(t - \overset{*}{\tau})$ (Fig. 1C) termed Scale-Invariant Temporal History (SITH) (Jacques et al., 2021, 2022; Shankar & Howard, 2012). We truncate the impulse responses at a finite horizon M; in our experiments, we set $M = \overset{*}{\tau}_{\max}$. All operations are vector-valued: the filter bank is applied independently to each embedding dimension (depth-wise 1-D convolution) to produce L d-dimensional "compressed slots" which are normalized via LayerNorm (Ba et al., 2016) and concatenated with the m most recent uncompressed tokens (Fig. 2) and subsequently processed by standard transformer layers (Fig. 3).

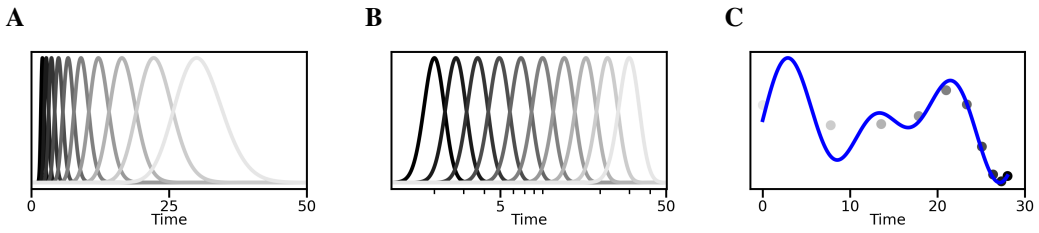

Figure 1: Log-compressed impulse responses from ten $\tilde{f}$ neurons with $k = 50$ show log-spaced peak times and a constant coefficient of variation (broader responses at later peaks). B. The same responses plotted on a log-time axis are uniformly spaced with equal widths. C. At time 30, $\tilde{f}$ neuron activations form a log-compressed memory of the input, approximating past values with finer resolution for more recent events. Grayscale dots correspond to impulse responses in panels A–B. Additional visualizations for $k = 10$ and $k = 100$ are shown in Fig. A1.

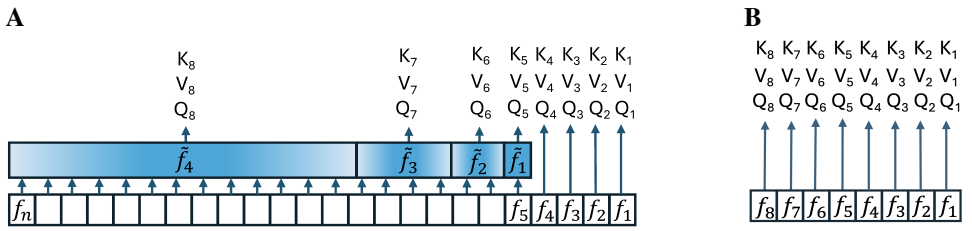

Figure 2: **A.** Illustration of compressed memory applied to the input sequence. A subset of the last m tokens, with $m = 4$ in this example, is used to compute four keys, queries, and values. The rest of the input sequence is used as an input to the compressed memory composed of L filters ($L = 4$ in this example), producing an additional four keys, queries, and values. **B.** Illustration of a standard transformer where each token in the sequence is used to generate keys, queries, and values.

Forming the compressed memory costs $O(MLd)$ once per input chunk (depth-wise 1-D convolution over embeddings of size d). Multi-head attention then runs on a fixed length $m + L$ sequence with cost $O\big((m + L)^2 d\big)$ per layer, instead of $O\big((m + M)^2 d\big)$ if the entire history were attended directly. Compression is applied only once as a preprocessing step before the first transformer block; subsequent layers operate on the concatenated sequence of fixed size. The training loss is computed only over the m uncompressed tokens. This approach allows the model to efficiently leverage information from a much longer context without incurring the full quadratic cost of attention over the entire sequence.

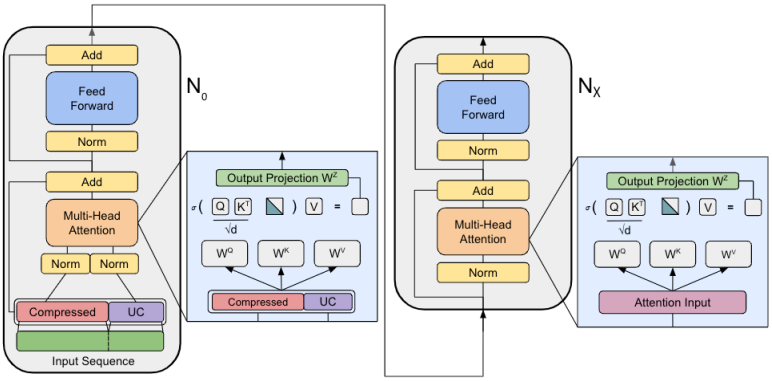

Figure 3: Architecture of the transformer based on scale-invariant compression. *Compressed* represents the portion of the input that is compressed with the scale-invariant filters, *UC* represents the uncompressed portion of the input. *Norm* depicts Layer Normalization, *Add* depicts residual connections, and *Feed-Forward* depicts a multi-layer perceptron at the end of each layer. At the *Multi-Head Attention* block in the first layer N_0, the output of the scale-invariant compression is concatenated with the uncompressed portion and projected to form queries, keys, and values. In subsequent layers N_X, where X is the layer number, the input is passed through without additional compression.

Fig. 4 compares strategies for modeling long-range dependencies. Transformer-XL, Compressive Transformer, and Recurrent Memory Transformer process input sequentially in segments, passing state between them via caching, compression, or recurrence (purple and red memory blocks and arrows). Our scale-invariant compression instead preprocesses the input: a subset of the full input (gray) is logarithmically compressed (red) into a fixed-size representation and combined with recent uncompressed tokens (green). This joint input is then processed by a standard transformer (blue layers) without cross-segment state management.

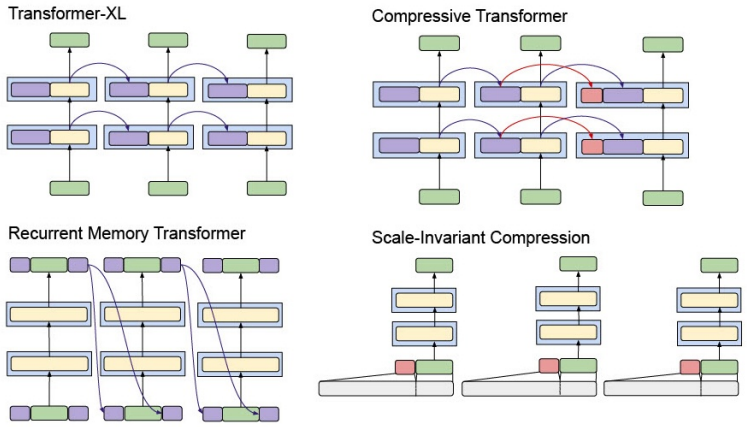

Figure 4: Comparison of long-context approaches. Transformer-XL, Compressive Transformer, and RMT handle long range by carrying state across segments (caching, compression, or recurrence). Our approach performs scale-invariant compression of the distant past to yield a fixed-size memory concatenated with recent tokens.

3 Training and Evaluation

Experiments were conducted on **WikiText-103** (Merity et al., 2016) and **PG-19** (Rae et al., 2019). **WikiText-103** contains over 100M words from Wikipedia, while **PG-19** comprises 28k Project Gutenberg books (1.9B words) exceeding 100k words each, designed for long-context modeling.

We trained transformer models and examined the effects of compression parameter k, and number of filters L, using an uncompressed sequence length of 256 tokens and $k \in \{100, 150, 200\}$. We fix the geometric spacing constant to $c = 0.19$. With $\overset{*}{\tau}_{\min} = 1$, the filter peaks are

$$\overset{*}{\tau}_i = \overset{*}{\tau}_{\min}(1+c)^{i-1}, \quad i = 1, \ldots, L,$$

which implies $\overset{*}{\tau}_{\max} = \overset{*}{\tau}_{\min}(1+c)^{L-1}$. Thus increasing L exponentially increases the temporal span (e.g., $L = 53 \Rightarrow \overset{*}{\tau}_{\max} \approx 8192$ tokens). In our sweeps we varied $L \in \{5, 9, \ldots, 53\}$, corresponding to filter windows from 2 to 8192 tokens.

For each k, we trained 13 models with L log-spaced filters (5–53), spanning window sizes from 2–8192 tokens. Each setting included a compression model (Fig. 2A) and a delta-pulse control (Fig. 2B) that bypassed compression via shift-register buffers, equivalent to retaining embeddings of the L preceding tokens. All models followed the GPT-2 Small architecture (Radford et al., 2019) (12 layers, 12 heads, 768-d embeddings, 3,072-d MLP, 50,304 vocab (50,257 rounded up to the nearest multiple of 64 for training efficiency); ∼124M parameters) with minor variation in number of parameters due to size of the learned positional encoding. Training used AdamW (β=0.9,0.95), weight decay 0.1, linear warmup for 700 steps to a peak learning rate of 6×10^{-4} with cosine decay to 6×10^{-5}, batch size ∼16,384 tokens/step, gradient clipping 1.0, and no dropout. Each model trained ∼48h on a 40 GB A100 GPU (90 epochs on WikiText-103; 4 on PG-19). We used a sliding window with stride m when forming batches, computing loss only over the m uncompressed tokens at each step. We report *raw perplexity* as the perplexity computed using the GPT-2 BPE tokenizer, given by e^{Loss_μ}, where $Loss_\mu$ denotes the average cross-entropy over the tokenized train or test subset. Following Rae et al. (2019) we report *per-word perplexity* as $e^{\text{Loss}_{tot}/n_{\text{words}}}$, where $Loss_{tot}$ denotes the total cross-entropy over the tokenized subset and n_{words} is the total number of words in the given subset. This facilitates direct comparisons between our GPT-2 BPE tokenizer-based models and other models using whitespace tokenization.

4 Results

The proposed scale-invariant compression filters consistently improve model performance by efficiently extending the temporal context on both the WikiText-103 and PG-19 datasets. This approach outperforms a control model using simple delta pulse filters, particularly as the number of memory filters L increases. As shown in Fig. 5, test perplexity on both datasets generally decreases as the number of filters L increases, with small non-monotonicities. Because the filter peaks are geometrically spaced, this corresponds to an approximately log-scale dependence on the maximum filter peak time (see Appendix A for numerical values of perplexity in each tested condition). This improvement becomes most apparent when the temporal context window significantly exceeds the number of filters (e.g., for $L > 17$ on WikiText-103 and $L > 21$ on PG-19), underscoring the benefit of capturing long-range dependencies.

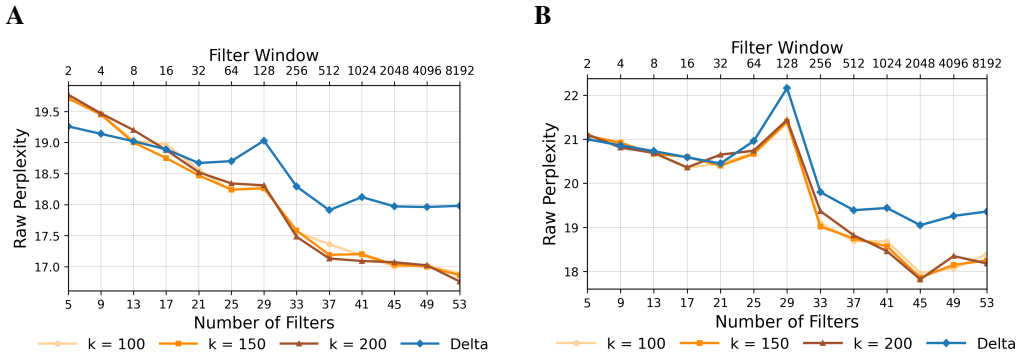

Figure 5: Test set perplexity decreases as the number of filters L increases. WikiText-103 (A) shows an approximately linear decline; PG-19 (B) shows a downward, step-like trend with a knee around mid-range L. The top axis indicates the corresponding filter window $\tau_{\max}$ (geometric spacing). The results are consistent for different values of k.

On WikiText-103, our best model achieved a per-word perplexity of **23.56** (with $L = 53$ filters covering an 8192-token window). As detailed in Table 1, this result is competitive with other transformer architectures utilizing long-context memory approaches with similar parameter counts. While these results are promising, we acknowledge that models with significantly more parameters (over 200M) achieve lower perplexity scores. The performance trend was replicated on the PG-19 dataset, though a direct comparison with other models was not possible due to a lack of published results at a similar scale or with per-word normalization of perplexity.

Model	**# param**	**Attn. Length**	**Per-word PPL**
Delta Pulse Filters (Our control model)	124M	309	25.48
Transformer-XL Standard (Dai et al., 2019)	151M	300	24.00
RMT (Bulatov et al., 2022)	151M	175	24.85
Transformer-XL + RMT (Bulatov et al., 2022)	151M	310	23.99
Scale-Invariant Compression Filters (Ours)	124M	309	**23.56**

Table 1: Comparison of transformer models of similar size (*# param*) on the WikiText-103 test set. Our model achieves the lowest per-word perplexity. *Attn. Length* is the input size considered by attention ($m + L$ for our model, with $m = 256$ uncompressed tokens plus $L = 53$ compressed inputs giving the total of 309 tokens).

5 Discussion

Unlike transformers, the human brain does not store a verbatim record of linguistic input. Instead, it learns statistical regularities in language by dynamically maintaining a memory representation of the recent past (Saffran et al., 1996; Kuhl, 2004). Cognitive scientists have argued that this representation forms a *mental timeline* of the past (Brown et al., 2007; Howard et al., 2015). Consistent with scale-invariant power-law decay of memory (Ebbinghaus, 2013; Wixted, 2004) and the Weber–Fechner law (Fechner, 1860), this timeline is thought to have a temporal resolution that gradually decreases from recent to distant events, yielding a logarithmically compressed representation. Neuroscience studies support this view, reporting neurons that activate sequentially with logarithmically compressed temporal receptive fields (Cao et al., 2022; Tiganj et al., 2018; Eichenbaum, 2014). Our approach augments transformer input representations with such a memory timeline, enabling the model to capture temporal dependencies across a wide input range. Logarithmic compression implies that the effective temporal range expands exponentially with the number of attention scores, offering a resource-efficient representation. Moreover, it entails that the density of neurons decreases as a power-law function of the peak times of their receptive fields. Power-law decay of long-range correlations is a pervasive phenomenon observed in DNA sequences (Mantegna et al., 1994), musical rhythm spectra (Levitin et al., 2012), earthquake statistics (Abe & Suzuki, 2004), and natural language (Ebeling & Neiman, 1995; Altmann et al., 2012). Lin & Tegmark (2017) similarly showed that mutual information between symbols can decay as a power law with increasing separation in context-free grammars. Integrating transformer architectures with power-law–decaying memory thus provides a principled framework for modeling domains where such statistical regularities naturally arise. A complementary, neurally motivated *time-local transformer* builds a log-compressed timeline via fixed recurrent dynamics and attends to it at each step, trading higher per-step compute for strict time-locality and greater biological plausibility (Dickson et al., 2025).

6 Conclusion

Our approach augments transformers with a scale-invariant memory inspired by a cognitive model of human memory. By incorporating temporal logarithmic compression as an input preprocessing step, our model efficiently encodes long-range dependencies into a fixed-size representation, allowing a standard transformer architecture to capture these dependencies effectively while maintaining computational tractability. Through experiments on the WikiText-103 and PG-19 datasets, we demonstrated that our model outperforms other transformer models of similar size. The observed gradual decrease in perplexity with increased temporal context highlights the efficacy of the scale-invariant memory representation in capturing long-range correlations. These findings illustrate that integrating cognitive principles into neural architectures can lead to more efficient language models.

References

Sumiyoshi Abe and Norikazu Suzuki. Scale-free network of earthquakes. *Europhysics Letters*, 65(4): 581–586, 2004.

Eduardo G Altmann, Giampaolo Cristadoro, and Mirko Degli Esposti. On the origin of long-range correlations in texts. *Proceedings of the National Academy of Sciences*, 109(29):11582–11587, 2012.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Gordon DA Brown, Ian Neath, and Nick Chater. A temporal ratio model of memory. *Psychological review*, 114(3):539, 2007.

Aydar Bulatov, Yury Kuratov, and Mikhail Burtsev. Recurrent memory transformer. *Advances in Neural Information Processing Systems*, 35:11079–11091, 2022.

Rui Cao, John H Bladon, Stephen J Charczynski, Michael E Hasselmo, and Marc W Howard. Internally generated time in the rodent hippocampus is logarithmically compressed. *Elife*, 11: e75353, 2022.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *Advances in Neural Information Processing Systems*, pp. 11969–11979, 2019.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, 2019.

Bert De Vries and Jose C Principe. The gamma model—a new neural model for temporal processing. *Neural networks*, 5(4):565–576, 1992.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1*, pp. 4171–4186, 2019.

Billy Dickson, James Mochizuki-Freeman, Md Rysul Kabir, and Zoran Tiganj. Time-local transformer. *Computational Brain & Behavior*, pp. 1–13, 2025.

Hermann Ebbinghaus. Memory: A contribution to experimental psychology. *Annals of neurosciences*, 20(4):155, 2013.

Werner Ebeling and Alexander Neiman. Long-range correlations between letters and sentences in texts. *Physica A: Statistical Mechanics and its Applications*, 215(3):233–241, 1995.

Howard Eichenbaum. Time cells in the hippocampus: a new dimension for mapping memories. *Nature Reviews Neuroscience*, 15(11):732–744, 2014.

Gustav Theodor Fechner. *Elemente der psychophysik*, volume 2. Breitkopf u. Härtel, 1860.

Charles Findling, Felix Hubert, International Brain Laboratory, Luigi Acerbi, Brandon Benson, Julius Benson, Daniel Birman, Niccolò Bonacchi, E Kelly Buchanan, Sebastian Bruijns, et al. Brain-wide representations of prior information in mouse decision-making. *Nature*, 645(8079):192–200, 2025.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Stephen Grossberg and Nestor A Schmajuk. Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural networks*, 2(2):79–102, 1989.

Marc W Howard, Christopher J MacDonald, Zoran Tiganj, Karthik H Shankar, Qian Du, Michael E Hasselmo, and Howard Eichenbaum. A unified mathematical framework for coding time, space, and sequences in the hippocampal region. *Journal of Neuroscience*, 34(13):4692–4707, 2014.

Marc W. Howard, Karthik H. Shankar, William R. Aue, and Amy H. Criss. A distributed representation of internal time. *Psychological Review*, 122(1):24–53, 2015.

Brandon Jacques, Zoran Tiganj, Marc Howard, and Per B Sederberg. Deepsith: Efficient learning via decomposition of what and when across time scales. *Advances in Neural Information Processing Systems*, 34:27530–27541, 2021.

Brandon G Jacques, Zoran Tiganj, Aakash Sarkar, Marc Howard, and Per Sederberg. A deep convolutional neural network that is invariant to time rescaling. In *International conference on machine learning*, pp. 9729–9738. PMLR, 2022.

Jikun Kang, Wenqi Wu, Filippos Christianos, Alex J Chan, Fraser Greenlee, George Thomas, Marvin Purtorab, and Andy Toulis. Lm2: Large memory models. *arXiv preprint arXiv:2502.06049*, 2025.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.

Ching-Yun Ko, Sihui Dai, Payel Das, Georgios Kollias, Subhajit Chaudhury, and Aurelie Lozano. Memreasoner: A memory-augmented LLM architecture for multi-hop reasoning. In *The First Workshop on System-2 Reasoning at Scale, NeurIPS'24*, 2024.

Patricia K Kuhl. Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11):831–843, 2004.

Daniel J Levitin, Parag Chordia, and Vinod Menon. Musical rhythm spectra from bach to joplin obey a 1/f power law. *Proceedings of the National Academy of Sciences*, 109(10):3716–3720, 2012.

Bryan Lim, Sercan O Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 2021.

Henry W Lin and Max Tegmark. Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7):299, 2017.

Rosario N Mantegna, Sergey V Buldyrev, Ary L Goldberger, Shlomo Havlin, Chung-Kang Peng, M Simons, and H Eugene Stanley. Linguistic features of noncoding dna sequences. *Physical review letters*, 73(23):3169–3172, 1994.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

James Mochizuki-Freeman, Md Rysul Kabir, and Zoran Tiganj. Incorporating a cognitive model for evidence accumulation into deep reinforcement learning agents. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*, 2019.

Jenny R Saffran, Richard N Aslin, and Elissa L Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, 1996.

Karthik H Shankar and Marc W Howard. A scale-invariant internal representation of time. *Neural Computation*, 24(1):134–193, 2012.

Pablo Tano, Peter Dayan, and Alexandre Pouget. A local temporal difference code for distributional reinforcement learning. *Advances in neural information processing systems*, 33:13662–13673, 2020.

Zoran Tiganj, Jason A Cromer, Jefferson E Roy, Earl K Miller, and Marc W Howard. Compressed timeline of recent experience in monkey lateral prefrontal cortex. *Journal of cognitive neuroscience*, 30(7):935–950, 2018.

Zoran Tiganj, Samuel J Gershman, Per B Sederberg, and Marc W Howard. Estimating scale-invariant future in continuous time. *Neural Computation*, 31(4):681–709, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. In *Advances in Neural Information Processing Systems*, 2020.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

John T Wixted. The psychology and neuroscience of forgetting. *Annu. Rev. Psychol.*, 55(1):235–269, 2004.

Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. In *International Conference on Learning Representations*, 2022.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

A Additional visualizations and results

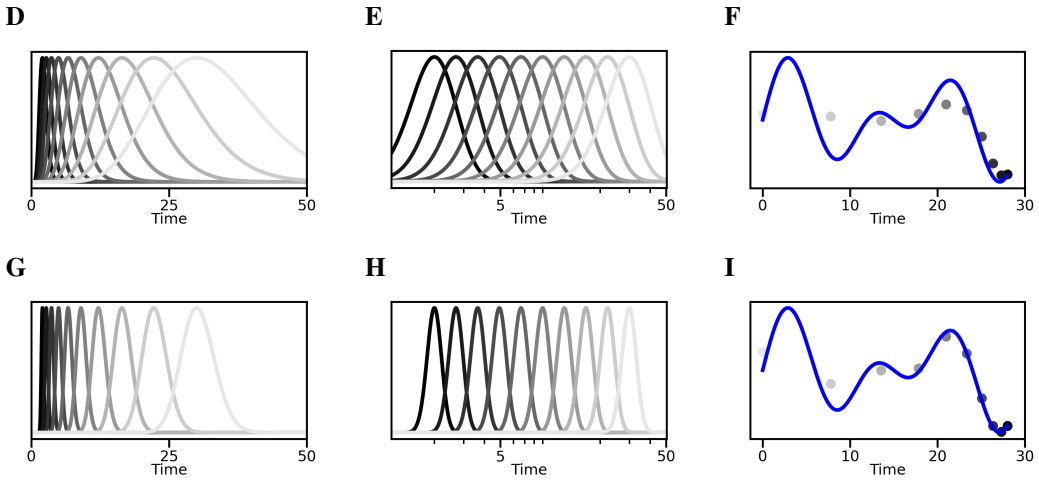

Figure A1: Continuation of Fig. 1. **D, E, F.** Same as panels A, B, C in the main text, but for $k = 10$. Note that a lower value of k results in wider filters but does not break the scale-invariance. Wider filters smooth the signal more, making the memory representation less able to follow changes in the signal (panel *F*). **G, H, I.** Same as panels A, B, C, but for $k = 100$. For very large values of k, the filters become narrower, making the space between them larger. Tokens that fall between peaks of the filters get less well represented than tokens that fall closer to the peaks.

# Filters	**Delta Filters**	**Filter Window**	**Scale-Invariant Compression Filters**		
	PPL **Raw / Per-Word**	$(\overset{*}{\tau}_{max})$	$k = 100$ **Raw / Per-word**	$k = 150$ **Raw / Per-word**	$k = 200$ **Raw / Per-word**
5	19.26/30.78	2	19.74/31.31	19.71/31.26	19.77/31.37
9	19.14/30.44	4	19.46/30.80	19.45/30.80	19.47/30.82
13	19.02/30.10	8	19.03/28.46	19.00/28.41	19.20/28.75
17	18.89/29.76	16	18.97/30.87	18.75/30.45	18.88/30.69
21	18.67/28.92	32	18.56/28.72	18.47/28.56	18.52/28.64
25	18.70/29.79	64	18.23/28.91	18.24/28.93	18.34/29.13
29	19.03/28.82	128	18.29/27.54	18.26/27.49	18.31/27.58
33	18.29/28.67	256	17.56/27.35	17.58/27.39	17.48/27.21
37	17.91/25.37	512	17.36/24.51	17.19/24.24	17.13/24.14
41	18.12/29.89	1024	17.18/28.07	17.20/28.12	17.09/27.90
45	17.97/27.41	2048	17.00/25.72	17.03/25.78	17.07/25.85
49	17.96/28.08	4096	17.02/26.38	17.00/26.35	17.02/26.38
53	17.98/25.48	8192	16.89/23.76	16.86/23.71	**16.76/23.56**

Table A1: WikiText-103 perplexity on the test set for an uncompressed size of 256. *Baseline* represents delta pulse filters where no compression occurs (control model). Our model, based on scale-invariant compression filters, was evaluated with $k = 100$, $k = 150$, and $k = 200$. Note that $\overset{*}{\tau}_{max}$ value is relevant only for our scale-invariant compression model. Each column reports two metrics in the format PPL / Per-Word PPL: PPL is the raw perplexity computed using the GPT-2 tokenizer, and Per-Word PPL is computed as $e^{\text{Loss}_{tot}/n_{\text{words}}}$ where $Loss_{tot}$ is the total cross-entropy loss over the tokenized test set and n_{words} is the number of words in the test set.

# Filters	Delta Filters	Filter Window	Scale-Invariant Compression Filters		
	PPL **Raw / Per-Word**	$(\overset{*}{\tau}_{max})$	$k = 100$ **Raw / Per-word**	$k = 150$ **Raw / Per-word**	$k = 200$ **Raw / Per-word**
5	21.00/102.11	2	21.05/101.66	21.08/101.90	21.11/102.14
9	20.86/101.04	4	20.86/100.29	20.92/100.72	20.81/99.91
13	20.73/99.97	8	20.69/99.62	20.67/99.42	20.69/99.61
17	20.59/98.90	16	20.35/96.77	20.60/98.60	20.36/96.83
21	20.45/97.83	32	20.46/96.87	20.40/97.43	20.65/99.28
25	20.96/101.75	64	20.68/99.65	20.67/99.56	20.74/100.11
29	22.16/110.45	128	21.36/104.48	21.40/104.74	21.44/105.09
33	19.80/92.66	256	19.10/87.72	19.02/87.21	19.37/89.67
37	19.39/90.12	512	18.69/85.26	18.75/85.64	18.82/86.14
41	19.44/90.11	1024	18.68/84.81	18.57/84.02	18.46/83.23
45	19.05/87.83	2048	17.97/80.41	17.86/79.64	**17.82/79.40**
49	19.26/88.86	4096	18.07/80.66	18.15/81.18	18.35/82.56
53	19.36/89.29	8192	18.39/82.60	18.25/81.62	18.18/81.16

Table A2: PG-19 perplexity on test set after 4 epochs. The models have the same configuration as those used on WikiText-103.