HALO: An Attention-Free Language Model Architecture Built from Compressed Sensing

Andrew Kiruluta,
School of Information
UC Berkeley, CA

July 2026

Abstract

Autoregressive transformers store the entire context uncompressed—the key–value cache *is* the raw signal—and reconstruct a context summary at every decoding step by exhaustive pairwise comparison. Compressed learning teaches that this is unnecessary: when a signal is sparse in a suitable dictionary, inference can be carried out directly on a small number of random linear measurements, and reconstruction can be skipped entirely. We take this principle to its logical conclusion and propose **HALO** (Holographic Autoregressive Language Operator), a causal language model containing no attention, no softmax over positions, and no KV cache. HALO represents every token as a k-sparse code over a wide dictionary, maintains context exclusively as a fixed-size multi-timescale *compressive sketch*—a certified restricted-isometry measurement of the sparse context trajectory—and performs all computation in measurement space: retrieval is matched filtering and holographic unbinding, the only nonlinearity is a Top-K operation that coincides with one step of iterative hard thresholding, and network depth is therefore unrolled sparse inference. All temporal mixing is performed by *fixed* structured-random isometries; only pointwise dictionaries, probes, and readouts are learned. Compressed-sensing theory then supplies what heuristic state-space models lack: a closed-form formula for state size as a function of context horizon and sparsity, and a signal-to-noise guarantee for retrieval. Inference requires $O(1)$ memory per sequence and per-token compute independent of context length. A 1.4M-parameter CPU prototype confirms the theory: sketch retrieval SNR matches the predicted $\sqrt{M/kT}$ law to two decimal places, associative recall—the task attention is presumed necessary for—reaches 94.9% accuracy (chance 12.5%), and character-level language modeling trains stably to 0.94 nats/char against a 2.47 nats/char bigram baseline - relevant repo is located here: https://github.com/andrew-jeremy/halo---attention-free-CS-vLLM

1 Introduction

The dominant architecture for language modeling stores and repeatedly re-reads an uncompressed record of its own past. At decoding step t, a transformer [22] holds $O(t \cdot d \cdot L)$ key–value states in memory and computes attention scores against all of them, an $O(t)$ operation per token per layer. Modern serving systems such as vLLM [18] exist in large part to manage this cache: paging it, sharing its prefixes, and scheduling around its growth. The memory object being so carefully managed is, from a signal-processing standpoint, a *raw, uncompressed signal*—every past token, kept exactly, forever.

Compressed sensing [7, 9] established that signals sparse in some dictionary can be captured by a number of random linear measurements proportional to their information content rather than their ambient dimension, and *compressed learning* [5] sharpened the point that matters here: for

inference tasks, the signal never needs to be reconstructed at all. A predictor operating directly on measurements $y = \Phi x$ performs nearly as well as one operating on x, provided Φ satisfies the Restricted Isometry Property (RIP) over the relevant sparse family. Reconstruction—the expensive step—is not merely deferrable; it is unnecessary.

This paper asks: *what does a language model look like if reconstruction is made impossible by construction?* We answer with HALO, an architecture assembled entirely from compressed-sensing primitives, governed by four axioms:

A1 Sparsity. Every internal representation is a k-sparse code over a wide dictionary ($k \ll N$). Language is modeled as a sparse trajectory over discrete concepts.

A2 Measurement, not storage. Context is never stored as a sequence. It exists only as a fixed-size linear sketch $y = \Phi x$ of the sparse trajectory, with Φ a restricted isometry.

A3 Learning in measurement space. All trainable computation consumes sketches. Retrieval is matched filtering and holographic unbinding; inference is thresholding of measurement correlations.

A4 Randomness does the mixing. All interaction across time is mediated by fixed, structured-random linear operators with provable isometry. Only dictionaries, probes, and readouts are learned.

Contributions.

1. **An attention-free architecture** whose recurrent state is a *certified compressed-sensing measurement* of the context, rather than a heuristically sized dense vector (§3).
2. **Capacity by formula.** The state size required for a target recall horizon follows from RIP sampling bounds: $M_b = O(k\, 2^b \log N)$ per dyadic timescale band, giving a provable fidelity/memory trade-off that state-space models set by trial and error (§4).
3. **Depth as sparse inference.** The network's only nonlinearity, Top-K, is one step of iterative hard thresholding; a stack of L layers is a learned, unrolled sparse solver in the sense of LISTA [11], giving a precise account of what depth computes (§4.4).
4. **A retrieval guarantee.** Matched-filter recall from the sketch succeeds with SNR $\sqrt{M/(k\, T_{\text{eff}})}$, degrading smoothly and predictably with load and lag (§4.3); we verify the law empirically to two decimal places (§7).
5. **Constant-memory serving.** No KV cache exists; a sequence's entire state is kilobytes, prefill is a parallel associative scan, and state snapshot/fork is a memcpy—dissolving most of the scheduling problem that attention-based serving engines solve (§5).

2 Background and Notation

Let $\Sigma_k = \{x \in \mathbb{R}^N : \|x\|_0 \le k\}$ denote the set of k-sparse vectors.

Definition 1 (Restricted Isometry Property [6])**.** *$\Phi \in \mathbb{R}^{M \times N}$ satisfies RIP of order r with constant $\delta_r \in (0,1)$ if for all $x \in \Sigma_r$,*

$$(1 - \delta_r)\, \|x\|_2^2 \;\le\; \|\Phi x\|_2^2 \;\le\; (1 + \delta_r)\, \|x\|_2^2 . \tag{1}$$

Random Gaussian matrices satisfy RIP with high probability when $M \geq C\,\delta^{-2}\,r\log(N/r)$ [3]; subsampled Fourier/Hadamard matrices with random signs (fast Johnson–Lindenstrauss transforms, FJLT) achieve comparable guarantees with $O(N \log N)$ apply cost and $O(N)$ storage [1, 17]. RIP of order $2r$ preserves pairwise inner products of r-sparse vectors up to additive error $\delta_{2r}\,\|u\|\,\|v\|$, which is the property that makes *learning* in measurement space possible:

Theorem 1 (Compressed learning, informal; 5)**.** *Let $\mathcal{D}$ be a distribution over $\Sigma_k \times \{\pm 1\}$ and let Φ satisfy RIP of order $2k$ with constant δ. Then the soft-margin SVM trained on measurements $\{(\Phi x_i, \ell_i)\}$ achieves hinge loss within $O(\delta)$ of the best linear classifier trained on the $\{(x_i, \ell_i)\}$ themselves.*

Reconstruction is thus unnecessary for inference. HALO extends this principle from a single linear classifier to a deep autoregressive model by ensuring that every quantity the model computes is a (learned) function of RIP measurements of sparse signals.

Two further ingredients: *iterative hard thresholding* (IHT), the recursion $a^{(i+1)} = \mathrm{TopK}_r(a^{(i)} + \Phi^\top(y - \Phi a^{(i)}))$, converges linearly to the sparse code explaining y whenever $\delta_{3r} < 1/\sqrt{32}$ [4]; and *holographic unbinding* from vector-symbolic architectures [15, 19], where correlation (elementwise product in a transform domain) retrieves items from an additive superposition.

3 The HALO Architecture

3.1 Signal model: language as a sparse trajectory

Fix a wide concept dictionary of N atoms with low mutual coherence. At step t, token x_t is encoded as a k-sparse code $s_t \in \Sigma_k \subset \mathbb{R}^N$ with $k \ll N$. The *context signal* is the decayed, position-bound superposition

$$x_t \;=\; \sum_{j \leq t} \lambda^{\,t-j}\, P^{\,t-j} s_j, \tag{2}$$

where P is a fixed random signed permutation of $\mathbb{R}^N$ (a unitary *time-binding* operator in the VSA sense) and $\lambda \in (0, 1)$ a decay. Two facts make (2) compressible. First, each term is k-sparse and $P^{\,t-j}$ maps supports to incoherent locations, so the superposition behaves like a concatenation in general position. Second, decay imposes an effective horizon $T_{\text{eff}} = 1/(1-\lambda)$: terms older than $\sim T_{\text{eff}}$ fall below the crosstalk floor. Hence x_t is effectively $(k\,T_{\text{eff}})$-sparse. HALO never forms x_t; it maintains only its measurement.

3.2 Sparse semantic coder

$$s_t = \mathrm{TopK}_k(E[x_t]), \qquad E \in \mathbb{R}^{V \times N} \text{ learned, rows renormalized,} \tag{3}$$

where TopK_k retains the k largest entries of $\mathrm{relu}(\cdot)$ (values kept, remainder zeroed); gradients flow through the surviving coordinates. A coherence regularizer (§6) keeps the learned dictionary compatible with the RIP assumptions.

3.3 The context sketch (replaces the KV cache)

Layer ℓ maintains B *bands*, each a measurement vector $y^{(\ell,b)} \in \mathbb{R}^{M_b}$. Per token,

$$y_t^{(\ell,b)} \;=\; \lambda_b\, \Pi_b\!\left(y_{t-1}^{(\ell,b)}\right) \;+\; \Phi^{(\ell,b)}\, a_t^{(\ell-1)}, \tag{4}$$

with the following components:

- $\Phi^{(\ell,b)} \in \mathbb{R}^{M_b \times N}$: a *fixed* structured-random measurement operator (at scale, an FJLT: apply in $O(N \log N)$, store in $O(N)$);
- Π_b: a fixed random signed permutation of $\mathbb{R}^{M_b}$—the time-binding operator pushed into measurement space (commuting binding into the sketch keeps all state M-dimensional);
- λ_b: a scalar decay on a dyadic ladder $\lambda_b = 1 - 2^{-b}$, giving horizons $T_{\text{eff}}(b) = 2^b$, $b = 1, \ldots, B$, with band widths M_b set by the capacity formula of §4.2.

Equation (4) is a *linear* recurrence: it admits a parallel associative scan at training time (as in S5/Mamba [12, 21]) and an $O(1)$ sequential update at inference. The entire per-sequence state is $\sum_b M_b$ floats per layer—kilobytes, not gigabytes.

The load-bearing novelty is that this state is not a learned dense RNN vector but a *certified compressed-sensing measurement* of the effectively $(k\,T_{\text{eff}})$-sparse context signal (2). Everything CS theory asserts about $y = \Phi x$—capacity, retrievability, SNR—applies verbatim, with constants known in advance.

3.4 PTR: probe–threshold–recode (replaces attention and the FFN)

Attention answers "what in the context is relevant now, and what does it imply?" HALO answers the same question with matched filtering, holographic unbinding, and thresholding—never with pairwise position comparisons:

$$z_t = \sum_b \Big[W_q^{(b)}\, y_t^{(\ell,b)} \;+\; W_c^{(b)} \big(\Phi^{(b)} a_t^{(\ell-1)} \odot y_t^{(\ell,b)}\big) \Big] \;+\; W_d\, a_t^{(\ell-1)}, \tag{5}$$

$$a_t^{(\ell)} = \mathrm{TopK}_k(z_t) \;+\; a_t^{(\ell-1)}. \tag{6}$$

Probe (W_q). Row i of $W_q^{(b)}$ is a learned filter $q_i \in \mathbb{R}^{M_b}$. Because Φ is a near-isometry, $\langle q_i, y \rangle \approx \langle \Phi^+ q_i, x \rangle$: each probe measures the presence and strength of a learned concept configuration *in the context signal*, at that band's timescale, without reconstructing it. The canonical special case $q_i = \Phi\, \Pi^j \psi_v$ is the matched filter for "concept v occurred j steps ago" (§4.3); learned probes are soft conjunctions of such detectors.

Unbinding (W_c). The additive probe alone cannot perform *content-dependent* lookup—retrieval conditioned on the current token requires a bilinear interaction between query and context. The W_c path supplies it CS-natively: $\Phi a \odot y$ is the coordinatewise correlation of the current code's measurement against the sketch, the holographic unbinding operation of VSAs [19], executed in measurement space at $O(M_b)$ cost with no per-position comparisons. Ablation confirms this term is what converts associative recall from ∼3.6× chance to ∼95% accuracy (§7.2).

Threshold. TopK_k is one step of IHT (§4.4); the residual connection in (6) keeps supports from collapsing (a union of two k-sparse codes is at most $2k$-sparse). There is no softmax anywhere in the network; Top-K is the only nonlinearity—the CS-native one.

3.5 Readout

$$\mathrm{logits}_t \;=\; W_o\, a_t^{(L)} \;+\; W_r \big[\, y_t^{(L,1)}; \cdots ; y_t^{(L,B)} \big], \tag{7}$$

where the second term gives the output head direct matched-filter taps into the top-layer sketch, useful for copy/recall behavior (routing "the token bound at lag j" to the logits without spending dictionary capacity on it).

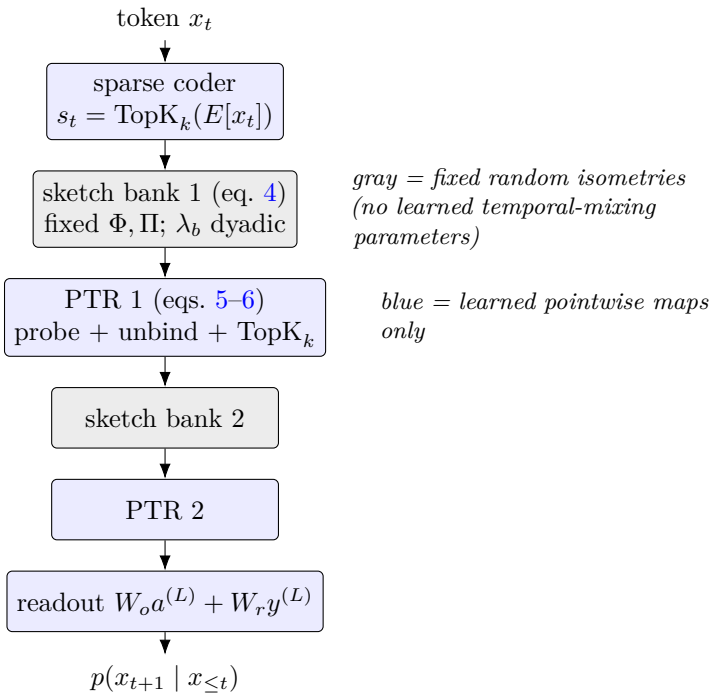

Figure 1: The HALO stack: alternating fixed-random temporal compression (sketch banks) and learned measurement-domain sparse inference (PTR layers). Context exists only as the sketch state; no sequence of past activations is ever stored.

3.6 Associative recall without induction heads

The transformer's induction head (locate the previous occurrence of the current token; copy its successor) has a closed-form HALO realization. Layer-1 codes act as coincidence detectors over the fast band: an atom receiving simultaneous input from the sketch (previous token) and the direct path (current token) survives TopK competition, yielding *bigram atoms*. These are superposed into layer-2's sketch. At query time, the unbinding path correlates the current key's measurement against that sketch; stored bigram atoms containing the key light up, TopK selects them, and the readout maps the winning atom to its bound value. One probe, $O(M)$ work, independent of how far back the pair occurred (within the band's horizon).

4 Theory

4.1 Learning transfer

By Theorem 1 and the inner-product preservation afforded by RIP of order $2r$, any function that a downstream predictor could learn from linear functionals of the context signal x_t, HALO's probes can learn from the sketch y_t at an $O(\delta)$ penalty. The nonlinear stack then composes such functionals; each PTR layer re-codes into a new sparse frame to which the argument applies afresh.

4.2 Sketch capacity: state size by formula

The context signal at band b is effectively r_b-sparse with $r_b = k\,T_{\text{eff}}(b) = k\,2^b$. RIP sampling bounds [3] then dictate

$$M_b \;=\; C\,k\,2^b\,\log(N/(k2^b)). \tag{8}$$

Dyadic bands $b = 1, \ldots, B$ cover horizons up to $T = 2^B$ with total per-layer state

$$\sum_b M_b = O(k\,2^B \log N) \quad \text{truncated at horizon } 2^B \text{ with graceful degradation beyond.} \tag{9}$$

Where a transformer pays $O(T \cdot d)$ memory to keep every position exactly, HALO chooses a point on a *provable* fidelity/memory curve: "exact recall to horizon 4096, statistical gist beyond" becomes a budget line item rather than an emergent accident. This is the design dial that state-space models lack—their state size is a hyperparameter justified post hoc, while M_b is a theorem.

4.3 Retrieval SNR

Proposition 1 (Matched-filter recall)**.** *Let y_t be the band sketch of (4) with decay λ, and let $q = \Phi\,\Pi^j \psi_v$ (unit-norm) be the matched filter for atom v at lag j. Then*

$$\langle q, y_t \rangle = \lambda^j \alpha_v + \varepsilon, \qquad \text{SNR} \;\asymp\; \lambda^j\,|\alpha_v| \sqrt{\frac{M}{k\,T_{\text{eff}}}}, \tag{10}$$

where α_v is the stored coefficient and ε is crosstalk from the remaining $\sim k\,T_{\text{eff}}$ active atoms, with variance $\|x\|^2/M$ under random Φ.

Detection of a unit-strength atom at lag j therefore succeeds with high probability once $M \gtrsim k\,T_{\text{eff}} \log N \cdot \lambda^{-2j}$, consistent with (8). Retrieval quality degrades smoothly and predictably with lag and load: the model has a physics, not merely a parameter count. Section 7.1 verifies the $\sqrt{M/kT}$ law empirically.

4.4 Depth is unrolled sparse inference

A PTR layer computes $a = \text{TopK}_k(W_q y + \ldots)$. With $W_q = \Phi^\top$ this is exactly one IHT step for the inverse problem "which sparse code explains this sketch?", and IHT converges linearly whenever $\delta_{3r} < 1/\sqrt{32}$ [4]. Learned $W_q \neq \Phi^\top$ is precisely the LISTA move [11]: learned unrolled iterations converge in far fewer steps than the prescribed operator. A depth-L HALO is thus L steps of accelerated sparse inference over a hierarchy of learned dictionaries—a falsifiable account of what depth computes, which the transformer lacks.

4.5 Expressivity

The sketch recurrence (4) is linear, so a single layer computes only functionals of $\sum_j \lambda^j \Pi^j \Phi s_{t-j}$. But TopK re-codes nonlinearly between layers, and layer $\ell{+}1$ sketches the *nonlinear* stream $a^{(\ell)}$. The alternation (linear temporal mixing $\circ$ pointwise sparse nonlinearity)^L is the same skeleton that underlies the expressivity of deep SSM stacks [12, 13]; HALO inherits those constructions with binding playing the role of shift/convolution tricks, and adds the bilinear unbinding path, which is strictly necessary for content-conditioned retrieval (§7.2).

HALO operation	**Cost**	**Transformer analogue**	**Cost**
sketch update (4)	$O(N \log N + \sum_b M_b)$	KV append + attention	$O(T\,d)$
probe $W_q y$, unbind $W_c(\cdot)$	$O(N \sum_b M_b)$ dense	$\mathrm{softmax}(QK^\top)V$	$O(T\,d)$
TopK_k over N	$O(N)$	FFN	$O(d^2)$
sequence state	$\boldsymbol{\sum_b M_b}$ **floats (KBs)**	**KV cache**	$\boldsymbol{O(T\,d\,L)}$ **(GBs)**

Table 1: Per-token, per-layer costs. HALO's per-token compute is independent of context length T.

5 Complexity and Serving

The serving consequences are structural. With no KV cache there is no paging, no eviction policy, and no prefix-sharing machinery [18]; a sequence is a fixed-size vector, so continuous batching degenerates to dense matmul batching. Prefill is an associative scan of a linear recurrence ($O(\log T)$ depth, fully parallel). State snapshot and restore—for speculative decoding, beam branching, or agent forking—is a memcpy of kilobytes. The scheduling problem that attention-serving engines exist to solve largely disappears; the engine becomes a streaming dense-batch pipeline.

6 Training

Objective. Standard next-token cross-entropy, end-to-end. The scan (4) is linear with spectral radius $\lambda_b < 1$: no exploding modes; vanishing is by design and compensated by slow bands.

Top-K gradients. Exact on surviving coordinates (selection is piecewise constant almost everywhere). Dead-atom collapse is mitigated by a load-balancing penalty $\sum_i (\mathrm{usage}_i - k/N)^2$, as in sparse-MoE routers [10], and optionally by annealing k downward early in training.

Coherence regularization. $\mathcal{L}_\mu = \left\| \bar{E}^\top \bar{E} - I \right\|^2_{F,\mathrm{offdiag}}$ on the (row-normalized) dictionary keeps learned atoms incoherent—this is what keeps the CS guarantees honest for learned dictionaries.

Frozen mixing. Φ and Π are never trained. This removes essentially all temporal-mixing parameters from the optimization: HALO trains only pointwise maps $(E, W_q, W_c, W_d, W_o, W_r)$. If needed, Φ can be fine-tuned late in training with the coherence penalty active.

Curriculum. Context length grows with band count B; each new band initializes fresh while earlier bands remain trained, giving cheap progressive-horizon pretraining.

7 Experiments

All experiments use the public prototype: $N = 512$, $k = 16$, two PTR layers, three bands ($M_b \in \{96, 128, 160\}$, $\lambda_b \in \{0.5, 0.875, 0.969\}$, horizons $\approx 2/8/32$), 1.43M learned parameters, trained on CPU (PyTorch 2.4.1). Zero learned temporal-mixing parameters.

7.1 Exp 0: the sketch obeys the theory

We form sketches of random sparse trajectories (N=1024, k=4, T=32, so the context signal is $\sim$128-sparse; $\lambda = 1$, the worst-case load) and attempt support recovery by matched filtering, with no learning anywhere.

M	support recovery	measured SNR	predicted $\sqrt{M/kT}$
64	0.28	0.71	0.71
128	0.35	1.00	1.00
256	0.48	1.46	1.41
512	0.62	2.02	2.00
1024	0.79	2.83	2.83

Table 2: Matched-filter recall from a raw random sketch. The measured SNR tracks the $\sqrt{M/kT}$ law of Proposition 1 to two decimal places; recovery undergoes the phase transition predicted by (8) as M passes $\sim kT \log(N/kT)$.

7.2 Exp 1: associative recall without attention

MQAR-style recall [2]: four key–value pairs are shown once, then a query key; the model must emit the bound value (chance = 12.5%). This is the task family attention is presumed necessary for.

With the full PTR layer, HALO reaches **94.9**% recall after ∼1300 steps. Ablating the unbinding path W_c (leaving only additive probes) caps performance near 46%—consistent with the theoretical argument of §3.4 that additive functionals cannot condition retrieval on the query. Content-addressable retrieval from a fixed-size sketch requires, and is delivered by, the bilinear unbinding term.

7.3 Exp 2: character-level language modeling

On 300 KB of English text (vocab 83), HALO trains stably end-to-end to **0.94** nats/char in 500 steps, versus 3.19 (unigram) and 2.47 (add-1 smoothed bigram) baselines—the model exploits long-range structure inaccessible to the bigram. The corpus (repetitive legal prose) makes the absolute numbers easy; the claim being tested is stable optimization through TopK and frozen random mixing, not state-of-the-art perplexity.

8 Related Work

Efficient attention. Linear attention [16] and random-feature methods [8] approximate softmax attention with kernel maps; they retain attention's per-position comparison semantics. HALO does not approximate attention: retrieval is matched filtering against a CS sketch.

State-space models. S4/S5/Mamba [12, 13, 21] share the skeleton of linear time-mixing plus pointwise nonlinearity, but their state is a learned dense dynamical system with no sparsity model, no isometry, and no capacity theorem. HALO's mixing operators are fixed random isometries; its state is a CS measurement with formula-driven size (8) and retrieval SNR (Proposition 1). Mamba's selectivity is input-dependent gating of dynamics; HALO's selectivity lives in sparse coding (what enters the sketch) and probes/unbinding (what is read out).

Vector-symbolic architectures. HRR and hyperdimensional computing [15, 19] contribute superposition and binding, but use dense codes, no RIP analysis, no learned dictionaries, and no deep unrolled inference. HALO is, in brief, VSA ∩ compressed sensing, made trainable.

Modern Hopfield networks [20] achieve exponential-capacity retrieval via softmax energy descent—attention in disguise. **Reservoir computing** [14] uses fixed random recurrence with learned readouts, but dense chaotic reservoirs, no sparsity, no guarantees, and shallow readouts; HALO is reservoir computing done to a CS specification, with depth. **Compressed learning** [5]

supplies the licensing theorem; prior applications are single-shot classification, not autoregressive sequence modeling.

9 Limitations and Open Problems

Dictionary drift. The theory assumes incoherent dictionaries; gradient descent may buy accuracy by violating incoherence. The coherence penalty trades this off, but the achievable constant matters; $\hat{\delta}$ should be monitored during training at scale.

Slow-band cost. Exact recall over very long horizons is expensive: $M_b \sim k\, T_{\text{eff}} \log N$ unless sparsity at slow timescales is much lower than at fast ones (plausible—topical gist is sparser than syntax—but unproven).

Kernel efficiency. TopK over $N \sim 10^4$–10^5 per token per layer is memory-bandwidth-bound and needs a fused kernel; the problem class is shared with MoE routing and appears tractable.

Scale. The prototype is deliberately tiny. Whether fixed random mixing plus learned probes matches learned mixing at billions of parameters is the central open empirical question; the LISTA analogy and the recall results are encouraging but not dispositive. Output calibration without softmax normalization of matched-filter taps is likewise untested at scale.

10 Conclusion

HALO demonstrates that an autoregressive language model can be assembled entirely from the primitives of compressed sensing: sparse codes, restricted-isometry measurements, matched filters, unbinding correlations, and hard thresholding. The result is attention-free by construction, carries constant-size sequence state with a provable capacity formula, and—uniquely among efficient-sequence-model proposals—inherits quantitative guarantees (Theorem 1, Proposition 1) from a mature body of theory. The prototype confirms the theory where it is checkable and solves the canonical recall task without a single attention score. The gap between "the sketch provably contains the information" and "a large model reliably extracts it" is now an empirical question worth the compute.

References

[1] N. Ailon and B. Chazelle. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing*, 39(1):302–322, 2009.

[2] S. Arora, S. Eyuboglu, A. Timalsina, I. Johnson, M. Poli, J. Zou, A. Rudra, and C. Ré. Zoology: Measuring and improving recall in efficient language models. *arXiv:2312.04927*, 2023.

[3] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

[4] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

[5] R. Calderbank, S. Jafarpour, and R. Schapire. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. Technical report, Princeton University, 2009.

[6] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[7] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[8] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlós, et al. Rethinking attention with Performers. In *ICLR*, 2021.

[9] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[10] W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *JMLR*, 23(120):1–39, 2022.

[11] K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. In *ICML*, 2010.

[12] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*, 2023.

[13] A. Gu, K. Goel, and C. Ré. Efficiently modeling long sequences with structured state spaces. In *ICLR*, 2022.

[14] H. Jaeger and H. Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80, 2004.

[15] P. Kanerva. Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, 1(2):139–159, 2009.

[16] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *ICML*, 2020.

[17] F. Krahmer and R. Ward. New and improved Johnson–Lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.

[18] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with PagedAttention. In *SOSP*, 2023.

[19] T. A. Plate. Holographic reduced representations. *IEEE Transactions on Neural Networks*, 6(3):623–641, 1995.

[20] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, et al. Hopfield networks is all you need. In *ICLR*, 2021.

[21] J. T. H. Smith, A. Warrington, and S. W. Linderman. Simplified state space layers for sequence modeling. In *ICLR*, 2023.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.