A PROBABILISTIC ACCOUNT OF ASSOCIATIVE MEMORY OF A SINGLE ATTENTION HEAD

Anonymous authors
Paper under double-blind review

ABSTRACT

A single attention head that predicts the next token by attending to an earlier, related position in the sequence can also be seen as a form of associative memory. However, the reason and amount of such association storage are generally unclear. In the proposed work, we investigate a non-asymptotic explanation for a single-head, attention-only transformer. We introduce random spherical representations to construct an explicit query-key matrix and a value-output matrix that jointly bind a designated target position in each stored sequence to a prescribed output token. The proposed system allows arbitrary queries and gives each sequence-position pair its own representation, unlike constructions that fix the query to an end-of-sequence token or assign a single embedding per vocabulary item, thereby allowing a repeated token to carry different associations at different positions. Our analysis of the proposed system isolates three effects, namely, query-key retrieval interference, decoder interference, and the operator-norm normalisation of the value-output map, and controls them using sub-Gaussian and sub-exponential concentration, together with a matrix Bernstein bound. It has been proved that with high probability, the construction simultaneously stores associations whenever the sequence length and vocabulary size are polynomial in head dimension, a quadratic-over-logarithmic lower bound on the associative capacity attributable to attention alone. Our experiments confirm the predicted growth of the normalisation constant's scaling exponent and a sharp memorisation phase transition. Our experiments show capacity follows the ambient head dimension when the query/value maps are learned, and shifts toward a spectral dimension when the attention geometry is fixed and strongly anisotropic.

1 INTRODUCTION

Self attention lets every position in a sequence attend to every other position, giving Transformers a way to model long range dependencies without the sequential propagation required by recurrent networks Vaswani et al. (2017). Mechanistically, each attention head reads from and writes to a shared residual stream, the transformer's principal channel for moving and preserving contextual information between positions Elhage et al. (2021). Reading from one position and writing to another is, in effect, a memory operation, which raises a precise question: how much can a single attention head, viewed as such a read/write channel, actually store? Elhage et al. (2021) decompose a head into a query-key (QK) circuit that selects a position and an output-value (OV) circuit that writes its content a retrieve-by-key, write-a-value operation and training dynamics recover this structure from data. Bietti et al. (2023) show that gradient descent recovers this outer-product structure when embeddings are nearly orthogonal, and Cabannes et al. (2024) study the resulting object directly as a single matrix summing key-value outer products, read out by an argmax over inner products. Both, however, treat the memory as an abstract weight matrix with no query-key stage characterizing what can be stored once a position is already known, rather than the coupled retrieve-then-copy structure specific to attention, which we study directly.

Associative memory has a longer history in neural computation, from the classical Hopfield network Hopfield (1982), storing $O(d/\log d)$ random patterns with vanishing error McEliece et al. (1987), to modern/dense variants with exponential-in-dimension capacity. Ramsauer et al. (2021) show that such a network's update rule is identical to softmax attention, and Hu et al. (2024) tie the capacity of transformer-compatible Kernelized Hopfield Models to spherical codes, matching a known expo-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

nential lower bound with a tight upper bound at an optimally spread code. This is a different notion of capacity from ours, how many of M patterns are recoverable from a nearby or partial query, using an optimized arrangement, rather than how many arbitrary, exact associations an attention head implements from i.i.d. random points on the sphere the price being an exponential to polynomial drop. Li et al. (2025) instead let context reshape the Hopfield attractor geometry itself, addressing retrieval flexibility rather than capacity, and is largely orthogonal to our question.

Empirically, a single attention layer already memorizes a striking number of associations, Dana et al. (2024) report that memorization capacity grows quadratically in d_h, even though their own proven bound is only linear in d_h a gap left unexplained. A recent line of work turns the question into quantitative capacity bounds for Transformers and attention layers, giving parameter counts and, in several cases, matching lower bounds under varying assumptions on the inputs and architecture Kim et al. (2023); Mahdavi et al. (2024); Cabannes et al. (2024); Nichani et al. (2025); Kajitsuka & Sato (2025); Vural et al. (2026); Table 1 summarizes these results alongside our own.

However, these analyses, however, isolates the capacity of attention alone under conditions that reflect how it is used in practice, they restrict context length relative to the model dimension and target exact label fitting rather than cue-based recall, assign one embedding per vocabulary item rather than per position so a repeated token cannot carry different associations, fix the retrieval query across every stored sequence, or locate the memorization bottleneck in the feed-forward block rather than attention itself Mahdavi et al. (2024); Nichani et al. (2025); Dana et al. (2024); Kajitsuka & Sato (2025).

We close this gap with a non-asymptotic, high-probability construction of a single attention head's query-key and output-value matrices from i.i.d. random spherical representations, in which every stored sequence draws its own query independently and every $(\text{sequence}, \text{position})$ pair not every vocabulary item receives its own representation. Decomposing the resulting error into query-key retrieval interference, decoder interference, and the operator-norm normalization of the value-output map, and bounding each via sub-Gaussian/sub-exponential concentration together with a matrix Bernstein inequality, we show:

- A single head stores $T = \Omega\Big(\frac{d_h^2}{\log d_h}\Big)$ associations whenever sequence length and vocabulary size are polynomial in d_h a quadratic over logarithmic capacity attributable to attention alone, isolated from any feed-forward memorization mechanism.
- To our knowledge, this is the first theoretical account consistent with the quadratic in d_h scaling that Dana et al. (2024) observe empirically but leave unexplained.
- When the attention geometry is instead fixed and strongly anisotropic, as in pretrained language models, capacity shifts from the ambient dimension d_h toward the smaller effective (spectral) dimension of that geometry, which we illustrate empirically.

2 Related Work

Symbols quoted from another paper's bound are that paper's own and do not always match our T, S, N, d_h (4).

Memorization capacity of attention and Transformers. The line of work closest to ours asks how many parameters suffice to exactly fit N arbitrary sequence-to-label associations, rather than to retrieve from a partial cue; Table 1 summarizes the assumptions and bounds. Kim et al. (2023) give the first such result for a full Transformer, in the tradition of sub-linear label-fitting bounds for feed-forward networks. Mahdavi et al. (2024) isolate multi-head attention under a context size $n < d$ and a linear-independence condition on the inputs, targeting exact label-fitting rather than cue-based recall. Cabannes et al. (2024) instead derive scaling laws for the abstract outer-product memory recall error against parameters and samples under Zipf-distributed labels a different quantity from the storage counts elsewhere in this line. Nichani et al. (2025) combine a self-attention layer with an MLP, but assign one random spherical embedding per vocabulary item rather than per position, and pin the retrieval query to a single embedding shared across every stored sequence. Kajitsuka & Sato (2025) obtain an analogous matching bound for a full transformer with a feed-forward block, showing that the feed-forward block, not attention, provably bottlenecks label association sharpen-

ing the question of what attention alone can store without answering it. Dana et al. (2024) remove the $n < d$ restriction for attention-only Transformers, improving on Mahdavi et al. (2024) with a bound linear in d_h, but, like Nichani et al. (2025), assign one embedding per vocabulary item; they also report empirical capacity growing quadratically in d_h, exceeding their own bound. Finally, Vural et al. (2026) move beyond hand-constructed weights, training a single-layer transformer by gradient descent with a matching finite-sample statistical lower bound, but, like the others, do not isolate attention's capacity from the MLP that follows it.

Relative to this literature, we isolate the capacity of a single attention head under assumptions closer to practice: an independently drawn query per stored sequence rather than one shared fixed embedding, one representation per (sequence, position) rather than per vocabulary item, and no restriction on context size relative to d_h.

High-dimensional probability toolkit. We control the three error sources in our analysis via sub-Gaussian/sub-exponential concentration for spherical inner products and a rectangular matrix Bernstein inequality, standard tools in high-dimensional probability Tropp (2012); Vershynin (2018).

3 Preliminaries

We describe and outline the notations employed throughout the paper. For norm and inner product, where we represent vectors as $x, y \in \mathbb{R}^{d_h}$, and $\langle x, y\rangle := x^\top y$ for their Euclidean inner product. The Euclidean norm is denoted by $\|x\|_2 = \sqrt{\langle x, x\rangle}$. For a matrix M, its operator norm is $\|M\|_{\mathrm{op}} := \sup_{\|x\|_2=1} \|Mx\|_2$. When there is no ambiguity, we use $\|M\|$ as shorthand for $\|M\|_{\mathrm{op}}$.

For a scalar random variable X, $\|X\|_{\psi_2}$ and $\|X\|_{\psi_1}$ denote its sub-Gaussian and sub-exponential Orlicz norms respectively.

We employ standard concentration bounds for isotropic spherical vectors. In particular, for $x \sim \mathrm{Unif}(\mathbb{S}^{d_h-1})$ and any fixed unit vector v,$\|\langle x, v\rangle\|_{\psi_2} \le \frac{C}{\sqrt{d_h}}$, $\Pr(|\langle x, v\rangle| \ge a) \le 2\exp(-cd_h a^2)$, and the product of two sub-Gaussian random variables is sub-exponential with $\|XY\|_{\psi_1} \le \|X\|_{\psi_2}\|Y\|_{\psi_2}$. We use these standard facts in the form given by Vershynin (2018). For independent, mean zero, sub-exponential $X_1, \ldots, X_m$, Bernstein's inequality results

$$\Pr\left(\left|\sum_{r=1}^{m} X_r\right| \ge a\right) \le 2\exp\left[-c\min\left\{\frac{a^2}{\sum_{r=1}^{m}\|X_r\|_{\psi_1}^2}, \frac{a}{\max_r \|X_r\|_{\psi_1}}\right\}\right],$$

following Vershynin (2018). Also, whenever $\|X_r\|_{\psi_1} \le C/d_h$,

$$\Pr\left(\left|\sum_{r=1}^{m} X_r\right| \ge a\right) \le 2\exp\left[-c\min\left\{\frac{d_h^2 a^2}{m}, d_h a\right\}\right].$$

In the Gaussian regime, this results the deviation $O(\sqrt{m\ell}/d_h)$ at failure probability $e^{-\Omega(\ell)}$. All simultaneous bounds below use the ordinary union bound.

4 Problem Formulation

We consider a single head attention only Transformer with head dimension d_h. The model is required to memorize T stored sequences, each of length S, together with a prescribed output token for each sequence. Throughout the paper, t indexes the stored sequence and s indexes the position within a sequence.

Stored Sequences: Let $t \in [T] := \{1, \ldots, T\}$ denote the sequence index, and let $s \in [S] := \{1, \ldots, S\}$ denote the position index within a sequence. For each sequence t and position s, we introduce a position-specific representation $x_s^t \in \mathbb{R}^{d_h}$. For each sequence t, let $j_t \in [S]$ denote the designated position that should be retrieved by the attention mechanism. We assume throughout that $\boxed{j_t \neq S \qquad \text{for every } t \in [T].}$ The corresponding target representation is $x_{j_t}^t$. Each sequence t is associated with an output token $y_t \in [N]$, where N is the vocabulary size. The corresponding unembedding vector is denoted by $u_{y_t} \in \mathbb{R}^{d_h}$.

Random Construction: All position specific representations are sampled independently and uniformly from the unit sphere $\mathbb{S}^{d_h-1} = \left\{x \in \mathbb{R}^{d_h} : \|x\|_2 = 1\right\}$. Thus, $x_s^t \overset{\text{i.i.d.}}{\sim} \text{Unif}\left(\mathbb{S}^{d_h-1}\right), \quad t \in [T],\ s \in [S]$. The index pair (t, s) identifies the position specific representation, so repeated token identities at different positions or in different stored sequences are not required to share the same vector.

The unembedding vectors $u_1, \ldots, u_N \in \mathbb{R}^{d_h}$ are also sampled independently and uniformly from $\mathbb{S}^{d_h-1}$. Here, all random vectors introduced above are mutually independent unless stated otherwise. Since the distributions are centrally symmetric,

$$\mathbb{E}[x_s^t] = \mathbb{E}[u_j] = 0.$$

Moreover, rotational invariance implies that for $x \sim \text{Unif}\left(\mathbb{S}^{d_h-1}\right)$, $\mathbb{E}[xx^\top] = \frac{1}{d_h} I_{d_h}$.

Distinct Target Tokens: Each stored sequence $t \in [T]$ is associated with a target output token $y_t \in [N]$. We assume throughout that the T target tokens are pairwise distinct $y_t \neq y_{t'}$ for all $t, t' \in [T]$ with $t \neq t'$. Under this assumption, the corresponding unembedding vectors $u_{y_1}, u_{y_2}, \ldots, u_{y_T}$ are mutually independent under the random construction above.

5 Our Approach

The query key interaction matrix is constructed as

$$W_{QK} = \beta \sum_{r=1}^{T} x_S^r (x_{j_r}^r)^\top,$$

where $\beta > 0$ is a scalar amplification factor. For sequence t, using the final position representation x_S^t, the attention score assigned to position $\ell \in [S]$ is $\gamma_\ell^t = (x_S^t)^\top W_{QK} x_\ell^t$. The corresponding attention coefficient is

$$\alpha_\ell^t = \frac{\exp(\gamma_\ell^t)}{\sum_{r=1}^{S} \exp(\gamma_r^t)}. \tag{1}$$

The retrieval objective is to ensure $\alpha_{j_t}^t \geq 1 - \epsilon_1$ simultaneously for all $t \in [T]$.

Retrieval Score Decomposition: For a target position j_t and a distractor position $\ell \neq j_t$, consider $(x_S^t)^\top W_{QK} x_{j_t}^t - (x_S^t)^\top W_{QK} x_\ell^t$. Substituting the definition of W_{QK} results

$$(x_S^t)^\top W_{QK} x_{j_t}^t - (x_S^t)^\top W_{QK} x_\ell^t = \beta \sum_{r=1}^{T} \langle x_S^t, x_S^r \rangle \left(\langle x_{j_r}^r, x_{j_t}^t \rangle - \langle x_{j_r}^r, x_\ell^t \rangle \right).$$

The contribution corresponding to $r = t$ contains the unit signal $\langle x_S^t, x_S^t \rangle \langle x_{j_t}^t, x_{j_t}^t \rangle = 1$. The remaining terms constitute the random retrieval interference, which are detailed in the subsequent sections.

Decoder Construction: We define the unnormalized decoder matrix $W'_{OV} = \sum_{t=1}^{T} u_{y_t} (x_{j_t}^t)^\top$. Applying W'_{OV} to the target representation $x_{j_t}^t$ gives

$$W'_{OV} x_{j_t}^t = \sum_{r=1}^{T} u_{y_r} \left\langle x_{j_r}^r, x_{j_t}^t \right\rangle.$$

The term corresponding to $r = t$ equals $u_{y_t} \left\langle x_{j_t}^t, x_{j_t}^t \right\rangle = u_{y_t}$. Hence $W'_{OV} x_{j_t}^t = u_{y_t} + \sum_{r \neq t} \left\langle x_{j_r}^r, x_{j_t}^t \right\rangle u_{y_r}$. The first term is the desired decoding signal, whereas the remaining terms constitute decoder interference.

Normalized Decoder: Let $\kappa > 0$ be a high probability upper bound on the operator norm of W'_{OV}, so that $\|W'_{OV}\|_{\text{op}} \leq \kappa$. We define the normalized decoder by $W_{OV} = \frac{W'_{OV}}{\kappa}$. On the event $\|W'_{OV}\|_{\text{op}} \leq \kappa$, we have $\|W_{OV}\|_{\text{op}} \leq 1$. Also, for every unit vector $e \in \mathbb{R}^{d_h}$, $\|W_{OV} e\|_2 \leq 1$. In particular, the ideal value associated with the target position is $v_{j_t}^t := W_{OV} x_{j_t}^t$, and therefore $\|v_{j_t}^t\|_2 \leq 1$. A high probability upper bound on κ is established in Section C.

Attention Output: Let v_s^t denote the value vector associated with position s of sequence t. At the target position, $v_{j_t}^t = W_{OV} x_{j_t}^t$. The actual attention output for sequence t is $z^t = \sum_{s=1}^{S} \alpha_s^t v_s^t$. Once the retrieval analysis establishes $\alpha_{j_t}^t \geq 1 - \epsilon_1$ and all value vectors satisfy $\|v_s^t\|_2 \leq 1$, we will establish $\|z^t - v_{j_t}^t\|_2 \leq 2\epsilon_1$.

Decoding Criterion: The output token for sequence t is decoded according to $y_t = \arg\max_{j \in [N]} \langle u_j, z^t \rangle$. It is therefore sufficient to establish $\langle u_{y_t}, z^t \rangle > \langle u_j, z^t \rangle \quad \forall j \neq y_t$. The proof first establishes a positive decoding margin for the ideal target value $v_{j_t}^t$, and then show that the perturbation $z^t - v_{j_t}^t$ is sufficiently small to preserve the ordering of the output logits.

Failure Probabilities: Let $\delta_1, \delta_2, \delta_3 > 0$ denote the failure-probability budgets assigned to the principal probabilistic events in the analysis. We assume $\delta_1 + \delta_2 + \delta_3 \leq \delta_{\text{fail}}$.

A final union bound will therefore yield an overall success probability of at least $1 - \delta_{\text{fail}}$.

Interference Thresholds: Let τ_1 denote the threshold used to control retrieval interference, and let τ_2 denote the corresponding threshold used in the decoder analysis.

The construction requires $\tau_1 < \frac{1}{2}, \tau_2 < \frac{1}{2}$. These inequalities ensure that the signal contribution dominates the corresponding random interference and that the relevant margins remain strictly positive. The allowable scaling of T are derived from the corresponding concentration inequalities.

5.1 Objective

The objective is to determine the largest order of T for which the construction simultaneously satisfies $\alpha_{j_t}^t \geq 1 - \epsilon_1$ for every $t \in [T]$, and $y_t = \arg\max_{j \in [N]} \langle u_j, z^t \rangle$ for every $t \in [T]$ with probability at least $1 - \delta_{\text{fail}}$. We establish this result via separately controlling retrieval interference, decoder interference, and the operator norm of W'_{OV}, and then combining the resulting probabilistic guarantees through a final union bound.

5.2 Attention Retrieval

We first analyze the ability of the query key attention mechanism to retrieve the target at the designated position j_t in each stored sequence $t \in [T]$. For every distractor position $\ell \neq j_t$, we compare the attention score assigned to the target representation $x_{j_t}^t$ with that assigned to the distractor representation x_ℓ^t.

Sufficient Score Gap for Retrieval: If for a fixed sequence t, $\gamma_{j_t}^t - \gamma_\ell^t \geq \log\left(\frac{S-1}{\epsilon_1}\right) \quad \forall \ell \neq j_t$, then from the softmax definition, $\alpha_{j_t}^t \geq \frac{1}{1+\epsilon_1} \geq 1 - \epsilon_1$. Thus it suffices to establish a uniform lower bound on $\gamma_{j_t}^t - \gamma_\ell^t$ for every sequence t and every distractor position $\ell \neq j_t$.

We now choose a common deviation level that controls all required retrieval events. We define $\ell_1 := \log\left(\frac{c_0 TS}{\delta_1}\right)$, where $c_0 > 0$ is a sufficiently large absolute constant, and define $\tau_1 := C_1 \left(\frac{\sqrt{T\ell_1}}{d_h} + \sqrt{\frac{\ell_1}{d_h}}\right)$, where $C_1 > 0$ is an absolute constant.

The first term controls the cross sequence sub-exponential interference, while the second term controls the individual distractor inner product.

Proposition 1 (Simultaneous retrieval score gap)**.** *With probability at least* $1 - \delta_1$*, simultaneously for every sequence* $t \in [T]$ *and every distractor position* $\ell \neq j_t$*,* $D_{t,\ell} \geq 1 - 2\tau_1$*. Equivalently,* $\gamma_{j_t}^t - \gamma_\ell^t \geq \beta(1 - 2\tau_1)$.

5.3 Choice of the Attention Scale

The retrieval guarantee requires $\tau_1 < \frac{1}{2}$. Under this condition, the uniform score gap bound implies $\alpha_{j_t}^t \geq 1 - \epsilon_1$ provided that $\beta \geq \frac{\log\left(\frac{S-1}{\epsilon_1}\right)}{1-2\tau_1}$.

5.4 Retrieval based Constraint on T

The condition $\tau_1 < \frac{1}{2}$ requires $C_1\left(\frac{\sqrt{T\ell_1}}{d_h} + \sqrt{\frac{\ell_1}{d_h}}\right) < \frac{1}{2}$, where $\ell_1 = \log\left(\frac{c_0 TS}{\delta_1}\right)$. A sufficient condition is $\frac{T\ell_1}{d_h^2} \le c$ for a sufficiently small absolute constant $c > 0$. Equivalently, $T \log\left(\frac{c_0 TS}{\delta_1}\right) \le d_h^2$. This is the retrieval based constraint on the number of stored sequences. The inversion of this constraint will be combined with the corresponding decoder based constraint in the final capacity analysis.

6 Decoder based Concentration

We next analyze the decoder constructed from the target representations and their corresponding target unembedding vectors. The notion is to show that, for every stored sequence t, the target unembedding vector u_{y_t} has a larger inner product with the ideal decoded value than any competing unembedding vector. Utilising the unnormalized decoder $W'_{OV} = \sum_{t=1}^{T} u_{y_t}(x_{j_t}^t)^\top$. For brevity, we define the target representation $x_t := x_{j_t}^t$. Then, $W'_{OV}x_t = \sum_{r=1}^{T} u_{y_r}\langle x_r, x_t\rangle$. Since $\langle x_t, x_t\rangle = 1$, we obtain $W'_{OV}x_t = u_{y_t} + \sum_{r\neq t}\langle x_r, x_t\rangle u_{y_r}$. The first term is the desired decoding signal, while the summation is the decoder interference.

6.1 Decoder Score Difference

We fix sequence $t \in [T]$ and an incorrect token $j \in [N] \setminus \{y_t\}$. We define the unnormalized decoder score difference $D_{t,j} := \langle u_{y_t}, W'_{OV}x_t\rangle - \langle u_j, W'_{OV}x_t\rangle$. Using the expansion of $W'_{OV}x_t$,

$$\begin{aligned} D_{t,j} &= \left\langle u_{y_t}, u_{y_t} + \sum_{r\neq t}\langle x_r, x_t\rangle u_{y_r}\right\rangle - \left\langle u_j, u_{y_t} + \sum_{r\neq t}\langle x_r, x_t\rangle u_{y_r}\right\rangle \\ &= 1 - \langle u_j, u_{y_t}\rangle + \sum_{r\neq t}\langle x_r, x_t\rangle\left(\langle u_{y_t}, u_{y_r}\rangle - \langle u_j, u_{y_r}\rangle\right). \end{aligned}$$

We define the direct unembedding correlation $R_{t,j} := \langle u_j, u_{y_t}\rangle$ and, for $r \neq t$, define the decoder interference terms $I_r^{t,j} := \langle x_r, x_t\rangle\left(\langle u_{y_t}, u_{y_r}\rangle - \langle u_j, u_{y_r}\rangle\right)$. Then, $D_{t,j} = 1 - R_{t,j} + \sum_{r\neq t} I_r^{t,j}$. The first term is the unit decoder signal. The quantities $R_{t,j}$ and $I_r^{t,j}$ constitute the decoder based interference.

We detail the constructs in Appendix A. We now combine the retrieval, decoder, operator-norm, and robustness results into a single high probability memorization guarantee. We recall that the retrieval analysis is controlled by $\tau_1 = C_1\left(\frac{\sqrt{T\ell_1}}{d_h} + \sqrt{\frac{\ell_1}{d_h}}\right), \qquad \ell_1 = \log\left(\frac{c_1 TS}{\delta_1}\right)$, while the decoder analysis is controlled by $\tau_2 = C_2\left(\frac{\sqrt{T\ell_2}}{d_h} + \sqrt{\frac{\ell_2}{d_h}}\right), \qquad \ell_2 = \log\left(\frac{c_2 TN}{\delta_2}\right)$.

The operator norm of the decoder is controlled with failure probability at most δ_3 using the value of κ established in Section C.

Theorem 2 (High probability memorization guarantee)**.** *Consider the random construction described in Section 4, with T stored sequences of length S and vocabulary size N. Assume that the target tokens $y_1, \ldots, y_T$ are pairwise distinct. Let $\delta_1, \delta_2, \delta_3 > 0$ satisfy $\delta_1 + \delta_2 + \delta_3 \le \delta_{\text{fail}}$. Suppose that $\tau_1 < \frac{1}{2}, \tau_2 < \frac{1}{2}$, and choose $\beta \ge \frac{\log\left(\frac{S-1}{\epsilon_1}\right)}{1-2\tau_1}$. Let κ satisfy $\Pr\left(\|W'_{OV}\|_{\text{op}} \le \kappa\right) \ge 1-\delta_3$, and suppose that $\epsilon_1 < \frac{1-2\tau_2}{4\kappa}$. Then, with probability at least $1 - \delta_{\text{fail}}$,the constructed attention only Transformer simultaneously satisfies $\alpha_{j_t}^t \ge 1 - \epsilon_1 \qquad$ for every $t \in [T]$, and $y_t = \arg\max_{j\in[N]}\langle u_j, z^t\rangle \qquad$ for every $t \in [T]$.*

A detail proof of which is deferred to Appendix.

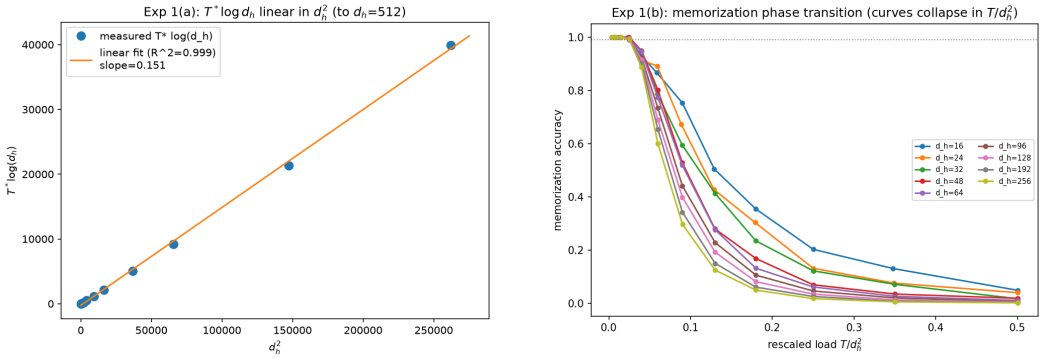

Figure 1: **Capacity law:** Left: $T^\star \log d_h$ is linear in d_h^2 up to $d_h = 512$, where $R^2 = 0.9992$). Right: memorisation accuracy versus the rescaled load T/d_h^2, the per-d_h curves collapse onto a single sharp phase transition.

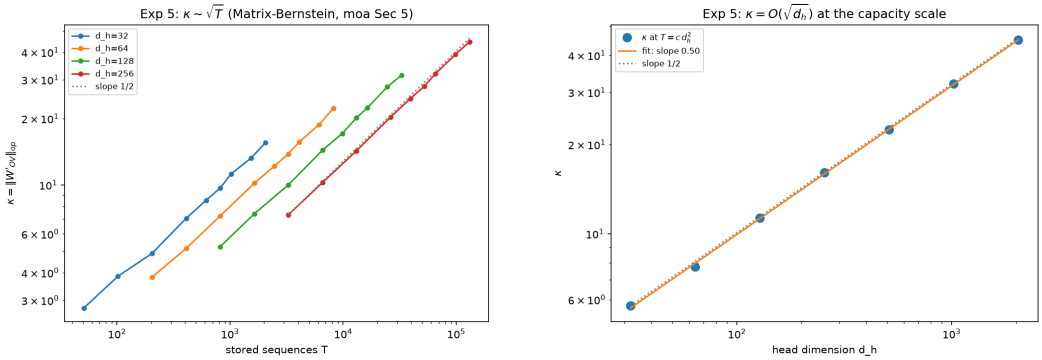

Figure 2: **Operator-norm law.** Left: $\kappa \sim \sqrt{T}$ at fixed d_h. Right: at $T = \frac{1}{4}d_h^2$, $\kappa = O(\sqrt{d_h})$ over $d_h \in \{32, \ldots, 2048\}$ (fitted exponent 0.501, $\kappa/\sqrt{d_h} \approx 1$).

7 Experiments

We evaluate our construction, which follows gradient-free methods in d_h dimensions. A memory efficient streaming evaluator comprises chunked construction of the $d_h \times d_h$, power iteration κ, batched decoding, and single precision arithmetic is employed to reach large d_h. The representations and unembeddings are randomly distributed on Unif($\mathbb{S}^{d_h-1}$), the vocabulary is minimal $N = T$, the target position is isotropic in $[S-1]$, and T^* depicts the largest load whose mean memorisation accuracy over random seeds stays at least 0.99.

We sweep d_h and locate $T^\star(d_h)$ by bisection, and test the prediction that $T^\star \log d_h$ is linear in d_h^2, the inverse of the feasibility constraint $T \log(d_h \max\{S, N\}) \le d_h^2$ reported in the left of Figure 1. Over $d_h \in \{16, \ldots, 512\}$ the linear fit gives $T^\star \log d_h = 0.151\, d_h^2 - 261$ with coefficient of determination $R^2 = 0.9992$, and the log-log power-law exponent is 1.96 on the asymptotic tail, where 2.00 over the full range, matching the predicted exponent of less than 2. The largest point reaches $T^\star = 6400$ stored associations at $d_h = 512$. We plot accuracy against the rescaled load T/d_h^2 collapses the per-d_h curves onto a single sharp phase transition reported in the right of Figure 1, which confirms that d_h^2 is the governing scale.

We measure $\kappa = \|W'_{OV}\|_{\mathrm{op}}$. At fixed d_h it grows as $\kappa \sim T^b$ with $b \in [0.47, 0.49]$ and predicted $1/2$). At the capacity scale $T = \frac{1}{4}d_h^2$ the ratio $\kappa/\sqrt{d_h}$ stays within $[0.97, 1.01]$ across the $64\times$ range $d_h \in \{32, \ldots, 2048\}$, with fitted exponent 0.501 reported in Figure 2. This confirms the Matrix Bernstein prediction of Section C and is the prediction that reaches largest d_h, because κ requires only the $d_h \times d_h$ matrix and the T target pairs. We also inspect how the learned heads

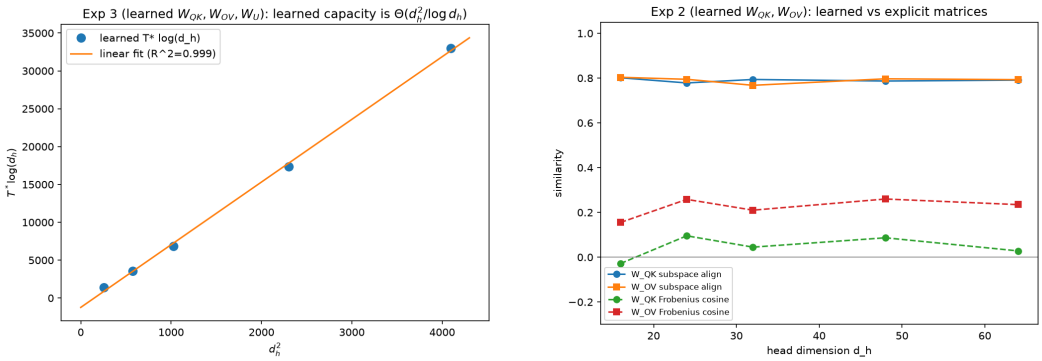

Figure 3: **Learned heads.** Left: learning W_{QK}, W_{OV}, and the unembedding recovers the $\Theta(d_h^2/\log d_h)$ law (exponent 2.00). Right: learned versus explicit matrix similarity near zero Frobenius cosine but ≈ 0.8 leading-subspace overlap.

realise the bound. Theorem 2 exhibits a feasible head. We ask whether gradient descent finds one with the same scaling. Freezing the representations and unembedding and learning W_{QK}, W_{OV} by cross entropy reproduces a linear capacity line ($R^2 = 0.999$) with fitted exponent 1.59. The frozen readout caps the growth relative to the theory. Additionally learning the unembedding recovers the full law exponent 2.00, $R^2 = 0.998$ reported in Figure 3) with a constant roughly an order of magnitude larger than the explicit construction, exactly as expected for a lower bound. The learned matrices are not entrywise close to the explicit ones (their Frobenius cosine is near zero) because the construction is identified only up to gauge i.e. an overall scale for W_{QK}. A scale and orthogonal head-basis change for the W_{OV} / unembedding pair, but their leading singular subspaces overlap by ≈ 0.8, gradient descent recovers the same low dimensional structure.

We also investigate whether the sequence length enters only through $\log(\max\{S, N\})$. The bound predicts that S affects capacity only through the logarithmic denominator. Sweeping S from 2 to 8192 at fixed d_h, $T^\star$ changes by only tens of percent and then saturates, for instance, $116 \to 96$ as S:2 $\to$ 8192 at $d_h = 64$) never polynomially. Reintroducing the $\max\{S, N\}$ term into the capacity invariant $T^\star \log(c_0 T^\star \max\{S, T^\star\}/\delta)$ makes it flatter in S than dropping the term i.e. coefficient of variation 0.058 versus 0.067 at $d_h = 64$, 0.279 versus 0.315 at $d_h = 32$ confirms the $\log(\max\{S, N\})$ form.

For the anisotropic geometry, we analyse the ambient versus effective dimension. We probe what governs capacity when the geometry is anisotropic, as real embeddings are measuring the effective dimension by the participation ratio $d_{\text{eff}} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ of the representation second-moment spectrum i.e. $d_{\text{eff}} = d_h$ when isotropic. Two regimes emerge reported in Figure 4. *(a) Learned head, real embeddings.* Drawing token embeddings from a pretrained language model, where PCA-projected to d_h, mild anisotropy, $d_h/d_{\text{eff}} \in [1.16, 1.26]$) and learning W_{QK}, W_{OV} and the unembedding, capacity still follows the ambient d_h. The capacity line fits d_h^2 with $R^2 = 0.9989$ slightly better than d_{eff}^2 ($R^2 = 0.9976$), and $T^\star/d_h^2 \approx 1.94$ is constant. Learnable maps can re-expand a low-d_{eff} input into the full d_h-dimensional working space, so anisotropy does not bottleneck a trained head. *(b) Fixed, strongly anisotropic geometry.* Using the explicit (non-learned) construction with representations drawn from a power law covariance i.e. anisotropy driving d_{eff} from 124 to 13 at $d_h = 128$, capacity is instead governed by d_{eff}. Normalising by d_{eff}^2 collapses $T^\star$ far better than d_h^2 which is the coefficient of variation 0.43 versus 0.82). The dependence is, however, only directional, the fitted exponent is 1.61, and $T^\star/d_{\text{eff}}^2$ still drifts because under anisotropy capacity is a functional of the whole covariance spectrum that the scalar participation ratio only approximates. We therefore report the effective dimension as a practical correction for real models instead of a sharp capacity law.

The real models gpt2, google/gemma-2-2b-it, meta-llama/Llama-2-7b-hf have anisotropic embedding space as reported in Figure 5. Hence we empirically report in Figure 5 the scaling of memorization capacity ($T*$) as a function of head dimension and ambient head dimension. It is observed that the scaling depends more on (d_{eff}) as compared to (d_h).

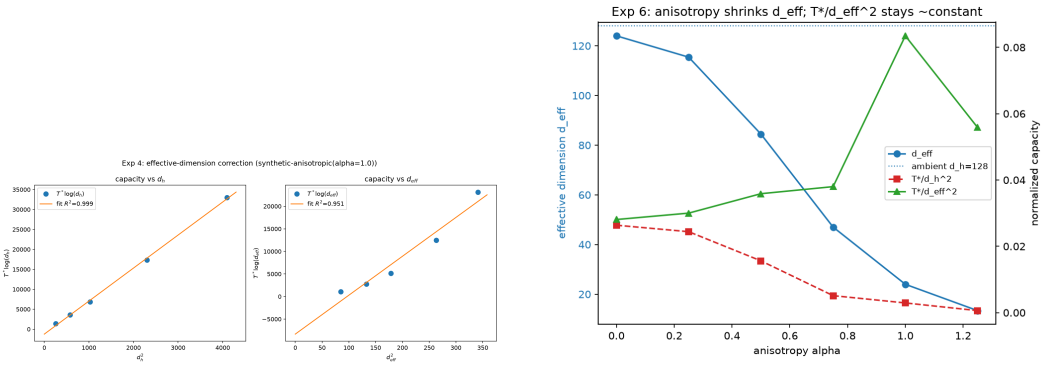

Figure 4: **Ambient versus effective dimension.** Left: with a learned head on pretrained embeddings, capacity tracks d_h^2 ($R^2 = 0.999$). Right: with a fixed anisotropic geometry, increasing anisotropy shrinks d_{eff} and $T^\star$; $T^\star/d_{\text{eff}}^2$ is far flatter than $T^\star/d_h^2$,

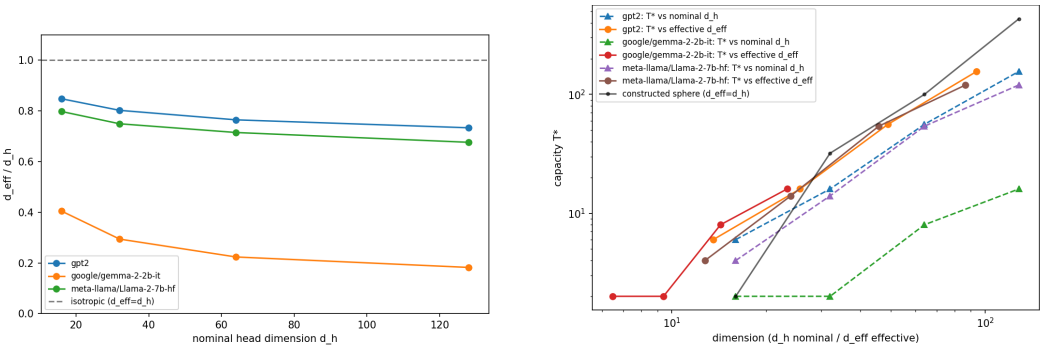

Figure 5: **Real models analysis** Left: Effective dimension (d_{eff}) and ambient dimension (d_h) ratio Right: Capacity (T^*) scaling graph for no. of sequences vs effective head dimension (d_{eff}) and ambient head dimension (d_h).

The capacity scale experiments are exact and hence bounded by the intrinsic cost of the object. At capacity the number of associations is $T^\star = \Theta(d_h^2/\log d_h)$, so a single point at $d_h = 10^4$ would require on the order of 10^7 associations, $O(d_h^4)$ construction, and $O(T^2)$ decoding. We therefore push each prediction as far as exact computation allows capacity to $d_h = 512$ and the operator-norm law to $d_h = 2048$ instead of an approximate point at larger d_h. Scaling further would require an approximate maximum inner product decoder and a sharded build, which we leave to future work.

8 Conclusion

We proposed a non- asymptotic account of the associative memory capacity of a single attention head. A construction from random spherical representations provably stores $\Omega(d_h^2/\log(d_h \max\{S, N\}))$ position specific associations, obtained by controlling query-key retrieval interference, decoder interference, and the operator-norm normalisation of the value output map with sub-Gaussian and sub-exponential concentration and a matrix Bernstein bound. Our experiments validate every component, the $\sqrt{d_h}$ growth of the normalisation constant, the sharp T/d_h^2 phase transition, and the end-to-end quadratic over logarithmic capacity, and show that gradient descent realises the same law. When the query/value maps are learned, capacity follows the ambient head dimension even on anisotropic real embeddings; when the attention geometry is fixed and strongly anisotropic, it shifts toward a spectral effective dimension, a practical correction rather than a sharp law. Establishing matching upper bounds, extending the analysis to multiple heads and layers, and characterising the spectral functional that governs capacity under anisotropy are natural directions for future work.

REFERENCES

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36: 1560–1588, 2023.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. In *International Conference on Learning Representations*, volume 2024, pp. 9718–9747, 2024.

Léo Dana, Muni Sreenivas Pydi, and Yann Chevaleyre. Memorization in attention-only transformers. *arXiv preprint arXiv:2411.10115*, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

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

Jerry Yao-Chieh Hu, Dennis Wu, and Han Liu. Provably optimal memory capacity for modern hopfield models: Transformer-compatible dense associative memories as spherical codes. *Advances in Neural Information Processing Systems*, 37:70693–70729, 2024.

Tokio Kajitsuka and Issei Sato. On the optimal memorization capacity of transformers. In *International Conference on Learning Representations*, volume 2025, pp. 66783–66824, 2025.

Junghwan Kim, Michelle Kim, and Barzan Mozafari. Provable memorization capacity of transformers. In *The eleventh international conference on learning representations*, 2023.

Chong Li, Taiping Zeng, Xiangyang Xue, and Jianfeng Feng. Dynamic manifold hopfield networks for context-dependent associative memory. *arXiv preprint arXiv:2506.01303*, 2025.

Sadegh Mahdavi, Renjie Liao, and Christos Thrampoulidis. Memorization capacity of multi-head attention in transformers. In *International Conference on Learning Representations*, volume 2024, pp. 49942–49967, 2024.

ROBERTJ McEliece, Edwardc Posner, EUGENER Rodemich, and SANTOSHS Venkatesh. The capacity of the hopfield associative memory. *IEEE transactions on Information Theory*, 33(4): 461–482, 1987.

Eshaan Nichani, Jason Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories. In *International Conference on Learning Representations*, volume 2025, pp. 4166–4207, 2025.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=tL89RnzIiCd`.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

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

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Mert Vural, Alberto Bietti, Mahdi Soltanolkotabi, and Denny Wu. Learning to recall with transformers beyond orthogonal embeddings. In *International Conference on Learning Representations*, volume 2026, pp. 75057–75129, 2026.

A THEORY

Query Key Construction: We use the query key matrix $W_{QK} = \beta \sum_{r=1}^{T} x_S^r (x_{j_r}^r)^\top$, where $\beta > 0$ is a scalar amplification factor. For sequence t and distractor position $\ell \neq j_t$, we define the normalized score difference $D_{t,\ell} := \frac{1}{\beta}\left(\gamma_{j_t}^t - \gamma_\ell^t\right)$. Substituting the definition of W_{QK} results

$$D_{t,\ell} = \sum_{r=1}^{T} \langle x_S^t, x_S^r \rangle \left(\langle x_{j_r}^r, x_{j_t}^t \rangle - \langle x_{j_r}^r, x_\ell^t \rangle \right).$$

The contribution corresponding to $r = t$ is $\langle x_S^t, x_S^t \rangle \left(\langle x_{j_t}^t, x_{j_t}^t \rangle - \langle x_{j_t}^t, x_\ell^t \rangle \right) = 1 - \langle x_{j_t}^t, x_\ell^t \rangle$.

Accordingly, we have $D_{t,\ell} = A^t - C_\ell^t$, where $A^t := 1 + \sum_{r \neq t} \langle x_S^t, x_S^r \rangle \langle x_{j_r}^r, x_{j_t}^t \rangle$, and $C_\ell^t := \langle x_{j_t}^t, x_\ell^t \rangle + \sum_{r \neq t} \langle x_S^t, x_S^r \rangle \langle x_{j_r}^r, x_\ell^t \rangle$. The unit term in A^t is the desired retrieval signal. All other terms constitute random interference.

Concentration of the Target Score Interference: For $r \neq t$, we define $B_r^t := \langle x_S^t, x_S^r \rangle \langle x_{j_r}^r, x_{j_t}^t \rangle$. We condition on the sequence t vectors x_S^t and $x_{j_t}^t$. Under this conditioning, the random variables $\{B_r^t : r \neq t\}$ are independent across r, because the pairs $(x_S^r, x_{j_r}^r)$ are mutually independent across r and are independent of $(x_S^t, x_{j_t}^t)$.

Moreover, conditioned on x_S^t and $x_{j_t}^t$, both factors in B_r^t are centered spherical inner products. Hence the spherical concentration bound above gives $\|\langle x_S^t, x_S^r \rangle\|_{\psi_2} \leq \frac{C}{\sqrt{d_h}}$, and $\left\|\langle x_{j_r}^r, x_{j_t}^t \rangle\right\|_{\psi_2} \leq \frac{C}{\sqrt{d_h}}$. Therefore, by the standard product bound above, we have $\|B_r^t\|_{\psi_1} \leq \frac{C}{d_h}$ conditionally on $(x_S^t, x_{j_t}^t)$. Also, $\mathbb{E}\left[B_r^t \mid x_S^t, x_{j_t}^t\right] = 0$. Thus, conditional on $(x_S^t, x_{j_t}^t)$, Bernstein's inequality applies to the independent family $\{B_r^t : r \neq t\}$, yielding

$$\Pr\left(\left| \sum_{r \neq t} B_r^t \right| \geq a \;\middle|\; x_S^t, x_{j_t}^t \right) \leq 2 \exp\left[-c \min\left\{ \frac{d_h^2 a^2}{T-1}, d_h a \right\} \right].$$

Since the right hand side is deterministic and does not depend on the conditioning variables, the same bound holds unconditionally,

$$\boxed{\Pr\left(\left| \sum_{r \neq t} B_r^t \right| \geq a \right) \leq 2 \exp\left[-c \min\left\{ \frac{d_h^2 a^2}{T-1}, d_h a \right\} \right].}$$

In particular, let $\ell \geq 1$ satisfy $\ell \leq c_0(T-1)$ for a sufficiently small absolute constant $c_0 > 0$, and choose $a = C\frac{\sqrt{T\ell}}{d_h}$.Then, the sub-gaussian term in Bernstein's inequality dominates the sub-exponential term, and hence $\Pr\left(\left| \sum_{r \neq t} B_r^t \right| \geq C\frac{\sqrt{T\ell}}{d_h} \right) \leq 2e^{-c\ell}$.

Concentration of the Distractor Score: We recall that the distractor contribution is decomposed as $C_\ell^t = R_{t,\ell} + \sum_{r \neq t} I_r^{t,\ell}$, where $R_{t,\ell} := \langle x_{j_t}^t, x_\ell^t \rangle$ is the target distractor representation correlation, and

$$I_r^{t,\ell} := \langle x_S^t, x_S^r \rangle \langle x_{j_r}^r, x_\ell^t \rangle, \qquad r \neq t,$$

is the cross sequence interference contributed by sequence r. We first control the correlation term. By the spherical concentration bound above, $\|R_{t,\ell}\|_{\psi_2} \leq \frac{C}{\sqrt{d_h}}$,and therefore, for every $a \geq 0$, $\Pr(|R_{t,\ell}| \geq a) \leq 2\exp(-cd_h a^2)$. Then, we consider the cross sequence interference terms $I_r^{t,\ell}$ for $r \neq t$:

$$I_r^{t,\ell} := \langle x_S^t, x_S^r \rangle \langle x_{j_r}^r, x_\ell^t \rangle.$$

For fixed t and ℓ, if $\ell \neq S$, condition on the sequence-t vectors x_S^t and x_ℓ^t. Under this conditioning, the family $\{I_r^{t,\ell} : r \neq t\}$ is independent across r, because the pairs $(x_S^r, x_{j_r}^r)$ are mutually independent across r and are independent of (x_S^t, x_ℓ^t). Also, conditioned on (x_S^t, x_ℓ^t), the two factors in

$I_r^{t,\ell}$ are centered spherical inner products. Hence, the spherical concentration bound above results $\|\langle x_S^t, x_S^r\rangle\|_{\psi_2} \le \frac{C}{\sqrt{d_h}}$, and $\left\|\langle x_{j_r}^r, x_\ell^t\rangle\right\|_{\psi_2} \le \frac{C}{\sqrt{d_h}}$.

If $\ell = S$, then $x_\ell^t = x_S^t$ is already included in the conditioning, conditional on x_S^t, the random vectors x_S^r and $x_{j_r}^r$ are independent for each $r \neq t$, and the two factors are again centered spherical inner products.

Then, by the standard product bound above, $\|I_r^{t,\ell}\|_{\psi_1} \le \frac{C}{d_h}$ conditionally on the indicated conditioning variables. Also, $\mathbb{E}\left[I_r^{t,\ell} \mid x_S^t, x_\ell^t\right] = 0$, for $\ell \neq S$, while for $\ell = S$ the same conclusion holds conditional on x_S^t alone.

Therefore, conditional on (x_S^t, x_ℓ^t), Bernstein's inequality gives

$$\Pr\left(\left|\sum_{r\neq t} I_r^{t,\ell}\right| \ge a \;\middle|\; x_S^t, x_\ell^t\right) \le 2\exp\left[-c\min\left\{\frac{d_h^2 a^2}{T-1}, d_h a\right\}\right].$$

Upon relaxation of the conditioning,

$$\boxed{\Pr\left(\left|\sum_{r\neq t} I_r^{t,\ell}\right| \ge a\right) \le 2\exp\left[-c\min\left\{\frac{d_h^2 a^2}{T-1}, d_h a\right\}\right].}$$

Thus, the distractor score is controlled by the sum of two components, the spherical correlation $R_{t,\ell}$ and the cross sequence interference $\sum_{r\neq t} I_r^{t,\ell}$.

A.1 Proof of Proposition 1

Proof. The concentration bounds above control the target score interference and distractor score interference. Applying a union bound over all relevant pairs (t,ℓ) gives, with probability at least $1-\delta_1$, $\left|R_{t,\ell} + \sum_{r\neq t} I_r^{t,\ell}\right| \le \tau_1$. Simultaneously for all $t \in [T]$ and all $\ell \neq j_t$.

Then, $A^t \ge 1-\tau_1$ and $C_\ell^t \le \tau_1$. Therefore, $D_{t,\ell} = A^t - C_\ell^t \ge 1 - 2\tau_1$. Multiplying by β gives $\gamma_{j_t}^t - \gamma_\ell^t \ge \beta(1-2\tau_1)$. □

B Concentration of the Unembedding Correlation

For fixed t and $j \neq y_t$, $R_{t,j} = \langle u_j, u_{y_t}\rangle$. By the spherical concentration bound above, $\|R_{t,j}\|_{\psi_2} \le \frac{C}{\sqrt{d_h}}$, and therefore $\Pr\left(|R_{t,j}| \ge a\right) \le 2\exp(-cd_h a^2)$. This controls the correlation between an incorrect unembedding vector and the target unembedding vector.

B.1 Concentration of the Decoder Interference

Consider $I_r^{t,j} = \langle x_r, x_t\rangle\left(\langle u_{y_t}, u_{y_r}\rangle - \langle u_j, u_{y_r}\rangle\right),\ r \neq t$. For fixed t and $j \neq y_t$, we distinguish between a possible exceptional index and the regular indices. If there exists an index $r^\star \neq t$ such that $y_{r^\star} = j$, then this index is unique because the target tokens $y_1, \ldots, y_T$ are pairwise distinct. In this case, $\sum_{r\neq t} I_r^{t,j} = I_{r^\star}^{t,j} + \sum_{\substack{r\neq t \\ r\neq r^\star}} I_r^{t,j}$. If no such index exists, then all indices $r \neq t$ are regular and the second sum below contains all terms. For every regular index $r \in [T]\setminus\{t, r^\star\}$, we define $I_{r,1}^{t,j} := \langle x_r, x_t\rangle\langle u_{y_t}, u_{y_r}\rangle$ and $I_{r,2}^{t,j} := \langle x_r, x_t\rangle\langle u_j, u_{y_r}\rangle$. Thus, $I_r^{t,j} = I_{r,1}^{t,j} - I_{r,2}^{t,j}$. For fixed t and j, the condition on $x_t, \quad u_{y_t}, \quad u_j$. Under this conditioning, for every regular index r we have $y_r \neq y_t$ and $y_r \neq j$. Hence the random vectors $x_r, \quad u_{y_r}$, are independent of the conditioning variables and are independent across different regular indices r. By the spherical concentration bound above, $\|\langle x_r, x_t\rangle\|_{\psi_2} \le \frac{C}{\sqrt{d_h}}$, $\|\langle u_{y_t}, u_{y_r}\rangle\|_{\psi_2} \le \frac{C}{\sqrt{d_h}}$,, and $\|\langle u_j, u_{y_r}\rangle\|_{\psi_2} \le \frac{C}{\sqrt{d_h}}$. Therefore, by the standard product bound above, $\left\|I_{r,1}^{t,j}\right\|_{\psi_1} \le \frac{C}{d_h}$, and $\left\|I_{r,2}^{t,j}\right\|_{\psi_1} \le \frac{C}{d_h}$. Also, by the triangle inequality for the ψ_1 norm, $\left\|I_r^{t,j}\right\|_{\psi_1} \le \frac{C}{d_h}$ for every regular index $r \in [T]\setminus\{t, r^\star\}$.

Also, $\mathbb{E}\left[I_r^{t,j} \mid x_t, u_{y_t}, u_j\right] = 0$. Thus, conditional on (x_t, u_{y_t}, u_j), the regular random variables $\left\{I_r^{t,j} : r \in [T] \setminus \{t, r^\star\}\right\}$ are independent and centered. Hence, by Bernstein's inequality, for every $a > 0$,

$$\Pr\left(\left|\sum_{\substack{r\neq t\\ r\neq r^\star}} I_r^{t,j}\right| \geq a \;\middle|\; x_t, u_{y_t}, u_j\right) \leq 2\exp\left[-c\min\left\{\frac{d_h^2 a^2}{T-2}, d_h a\right\}\right].$$

Since the right-hand side does not depend on the conditioning variables, the same bound holds unconditionally $\Pr\left(\left|\sum_{\substack{r\neq t\\ r\neq r^\star}} I_r^{t,j}\right| \geq a\right) \leq 2\exp\left[-c\min\left\{\frac{d_h^2 a^2}{T-2}, d_h a\right\}\right]$.

It remains to control the possible exceptional term. If an $r^\star$ exists, then $y_{r^\star} = j$, and therefore $\langle u_j, u_{y_{r^\star}}\rangle = 1$. Hence, $I_{r^\star}^{t,j} = \langle x_{r^\star}, x_t\rangle\left(\langle u_{y_t}, u_{y_{r^\star}}\rangle - 1\right)$. Since both vectors u_{y_t} and $u_{y_{r^\star}}$ have unit norm, $|\langle u_{y_t}, u_{y_{r^\star}}\rangle - 1| \leq 2$.Therefore, $|I_{r^\star}^{t,j}| \leq 2\,|\langle x_{r^\star}, x_t\rangle|$. By the spherical concentration bound above, $\|\langle x_{r^\star}, x_t\rangle\|_{\psi_2} \leq \frac{C}{\sqrt{d_h}}$, and hence $\Pr\left(|I_{r^\star}^{t,j}| \geq a\right) \leq 2\exp(-cd_h a^2)$. When no exceptional index $r^\star$ exists, this term is absent. We combine the regular contribution and the possible exceptional contribution, there exists an absolute constant $C > 0$ such that, for every $\ell \geq 1$, $\left|\sum_{r\neq t} I_r^{t,j}\right| \leq C\left(\frac{\sqrt{T\ell}}{d_h} + \sqrt{\frac{\ell}{d_h}}\right)$ with probability at least $1 - 4e^{-c\ell}$. The first term corresponds to the sum of the regular sub-exponential interference terms, while the second term accounts for the possible unique exceptional index satisfying $y_{r^\star} = j$.

B.2 Simultaneous Decoder Margin

We define $\ell_2 := \log\left(\frac{c_2 TN}{\delta_2}\right)$, where $c_2 > 0$ is a sufficiently large absolute constant, and define $\tau_2 := C_2\left(\frac{\sqrt{T\ell_2}}{d_h} + \sqrt{\frac{\ell_2}{d_h}}\right)$, where $C_2 > 0$ is an absolute constant.

The first term controls the cumulative decoder interference, while the second controls the direct unembedding correlation.

Proposition 3 (Simultaneous decoder margin)**.** *With probability at least $1 - \delta_2$, simultaneously for every $t \in [T]$ and every incorrect token $j \in [N] \setminus \{y_t\}$, $D_{t,j} \geq 1 - 2\tau_2$. Equivalently, $\langle u_{y_t}, W'_{OV} x_t\rangle - \langle u_j, W'_{OV} x_t\rangle \geq 1 - 2\tau_2$.*

Proof. We choose $a_1 = C\left(\frac{\sqrt{T\ell_2}}{d_h} + \sqrt{\frac{\ell_2}{d_h}}\right)$ for the cumulative decoder interference, and $a_2 = C\sqrt{\frac{\ell_2}{d_h}}$ for the direct unembedding correlation.

By the concentration bounds established above and a union bound over all relevant pairs (t, j), after choosing the numerical constant in ℓ_2 sufficiently large, with probability at least $1 - \delta_2$, $|R_{t,j}| \leq a_2$ and $\left|\sum_{r\neq t} I_r^{t,j}\right| \leq a_1$ simultaneously for all $t \in [T]$ and $j \in [N] \setminus \{y_t\}$. Hence, $D_{t,j} = 1 - R_{t,j} + \sum_{r\neq t} I_r^{t,j} \geq 1 - a_2 - a_1$. Absorbing the absolute constants into C_2 results $D_{t,j} \geq 1 - 2\tau_2$. □

B.3 Decoder Based Feasibility Condition

The decoder margin is strictly positive whenever $\tau_2 < \frac{1}{2}$. Then, Proposition 3 gives $D_{t,j} > 0$ for every $t \in [T]$ and every $j \neq y_t$.

Thus the retrieval and decoding analyses impose the two conditions $\tau_1 < \frac{1}{2}$, $\tau_2 < \frac{1}{2}$. Using the definition of τ_2, a sufficient condition for the decoder based concentration is $C_2\left(\frac{\sqrt{T\ell_2}}{d_h} + \sqrt{\frac{\ell_2}{d_h}}\right) < \frac{1}{2}$, where $\ell_2 = \log\left(\frac{c_2 TN}{\delta_2}\right)$. In particular, retaining the dominant cumulative interference term results $T\log\left(\frac{c_2 TN}{\delta_2}\right) \leq d_h^2$. The combination of this constraint with the retrieval based constraint from Section 5.2 determines the admissible scaling of T.

C Operator Norm Control

We now establish a high probability upper bound on the operator norm of the decoder matrix $W'_{OV} = \sum_{t=1}^{T} u_{y_t}(x_{j_t}^t)^\top$. The operator norm is required because the normalized decoder $W_{OV} = \frac{W'_{OV}}{\kappa}$ must act as a uniformly bounded linear map on all unit vectors. Instead of controlling $W'_{OV}e$ separately for every direction e, we use the rectangular Matrix Bernstein inequality of Tropp (2012).

C.1 Decomposition into Independent Random Matrices

We define $X_t := u_{y_t}(x_{j_t}^t)^\top, t \in [T]$. Then $W'_{OV} = \sum_{t=1}^{T} X_t$. By the random construction and the distinct target assumption, the vectors $u_{y_1}, \ldots, u_{y_T}$ are mutually independent. They are also independent of the position specific representations $x_{j_1}^1, \ldots, x_{j_T}^T$. Also, $X_1, \ldots, X_T$ are independent random matrices.

Since u_{y_t} and $x_{j_t}^t$ are independent and uniformly distributed on $\mathbb{S}^{d_h-1}$, $\mathbb{E}[u_{y_t}] = \mathbb{E}[x_{j_t}^t] = 0$. Therefore, $\mathbb{E}[X_t] = \mathbb{E}\left[u_{y_t}(x_{j_t}^t)^\top\right] = \mathbb{E}[u_{y_t}]\mathbb{E}[x_{j_t}^t]^\top = 0$. Hence $\mathbb{E}[X_t] = 0$. Thus the summands satisfy the mean zero condition of the Matrix Bernstein inequality.

For any vectors u, e, $\|ue^\top\|_{\text{op}} = \|u\|_2\|e\|_2$. Since $\|u_{y_t}\|_2 = \|x_{j_t}^t\|_2 = 1$, we have $\|X_t\|_{\text{op}} = 1$ almost surely. Thus, the uniform bound in the Matrix Bernstein inequality can be taken to be $R = 1$.

C.2 Matrix Variance Parameter

For the rectangular Matrix Bernstein inequality, we define $\sigma^2 := \max\left\{\left\|\sum_{t=1}^{T}\mathbb{E}[X_tX_t^\top]\right\|_{\text{op}}, \left\|\sum_{t=1}^{T}\mathbb{E}[X_t^\top X_t]\right\|_{\text{op}}\right\}$. We compute the two terms separately. First, $X_tX_t^\top = u_{y_t}(x_{j_t}^t)^\top x_{j_t}^t u_{y_t}^\top = u_{y_t}u_{y_t}^\top$, because $\|(x_{j_t}^t)\|_2^2 = 1$. By rotational invariance of the uniform distribution on the sphere, $\mathbb{E}\left[u_{y_t}u_{y_t}^\top\right] = \frac{1}{d_h}I_{d_h}$. Therefore, $\sum_{t=1}^{T}\mathbb{E}[X_tX_t^\top] = \frac{T}{d_h}I_{d_h}$, and hence $\left\|\sum_{t=1}^{T}\mathbb{E}[X_tX_t^\top]\right\|_{\text{op}} = \frac{T}{d_h}$. Similarly, $X_t^\top X_t = x_{j_t}^t u_{y_t}^\top u_{y_t}(x_{j_t}^t)^\top = x_{j_t}^t(x_{j_t}^t)^\top$.

Again, rotational invariance gives $\mathbb{E}\left[x_{j_t}^t(x_{j_t}^t)^\top\right] = \frac{1}{d_h}I_{d_h}$. Also,

$$\left\|\sum_{t=1}^{T}\mathbb{E}[X_t^\top X_t]\right\|_{\text{op}} = \frac{T}{d_h}.$$

It follows that $\sigma^2 = \frac{T}{d_h}$.

C.3 Matrix Bernstein Bound

We now apply the rectangular Matrix Bernstein inequality Tropp (2012). For independent, mean-zero random matrices $X_t \in \mathbb{R}^{d_h\times d_h}$ satisfying $\|X_t\|_{\text{op}} \le R$ and variance parameter σ^2, the inequality gives, for every $a \ge 0$, $\Pr\left(\left\|\sum_{t=1}^{T}X_t\right\|_{\text{op}} \ge a\right) \le 2d_h\exp\left(-\frac{a^2}{2\sigma^2+\frac{2}{3}Ra}\right)$.

Using $R = 1$ and $\sigma^2 = \frac{T}{d_h}$, we obtain $\Pr\left(\|W'_{OV}\|_{\text{op}} \ge a\right) \le 2d_h\exp\left(-\frac{a^2}{2T/d_h+\frac{2}{3}a}\right)$. Equivalently, $\Pr\left(\|W'_{OV}\|_{\text{op}} \ge a\right) \le 2d_h\exp\left(-\frac{a^2/2}{T/d_h+a/3}\right)$.

C.4 High Probability Choice of κ

Let $L_3 := \log\left(\frac{2d_h}{\delta_3}\right)$. We choose $a = \kappa$ so that $2d_h\exp\left(-\frac{\kappa^2/2}{T/d_h+\kappa/3}\right) \le \delta_3$. It is sufficient that $\frac{\kappa^2/2}{T/d_h+\kappa/3} \ge L_3$. Equivalently, $\kappa^2 - \frac{2L_3}{3}\kappa - \frac{2TL_3}{d_h} \ge 0$. The positive root of the correspond-

ing quadratic equation is $\kappa = \frac{L_3}{3} + \sqrt{\frac{L_3^2}{9} + \frac{2TL_3}{d_h}}$. Therefore, with probability at least $1 - \delta_3$, $\|W'_{OV}\|_{\mathrm{op}} \le \frac{L_3}{3} + \sqrt{\frac{L_3^2}{9} + \frac{2TL_3}{d_h}}$.

Using $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$, we obtain the simpler bound $\|W'_{OV}\|_{\mathrm{op}} \le \sqrt{\frac{2TL_3}{d_h}} + \frac{2L_3}{3}$. Hence, for an absolute constant $C > 0$, $\kappa = C\left(\sqrt{\frac{T}{d_h}\log\frac{2d_h}{\delta_3}} + \log\frac{2d_h}{\delta_3}\right)$ is a valid high probability upper bound satisfying $\Pr\left(\|W'_{OV}\|_{\mathrm{op}} \le \kappa\right) \ge 1 - \delta_3$.

C.5 Normalized Decoder

On the event $\|W'_{OV}\|_{\mathrm{op}} \le \kappa$, the normalized decoder $W_{OV} = \frac{W'_{OV}}{\kappa}$ satisfies $\|W_{OV}\|_{\mathrm{op}} \le 1$. Therefore, for every unit vector $e \in \mathbb{R}^{d_h}$, $\|Ae\|_2 \le \|W_{OV}\|_{\mathrm{op}}\|e\|_2 \le 1$. In particular, $\|v_{j_t}^t\|_2 = \|W_{OV}x_{j_t}^t\|_2 \le 1$.

Order of the Bound at the Target Capacity Scale: Suppose that the number of stored sequences is chosen at the retrieval scale $T = \Theta\left(\frac{d_h^2}{\log(d_h \max\{S,N\})}\right)$, and δ_3 is fixed. Then $\log\left(\frac{2d_h}{\delta_3}\right) = \Theta(\log d_h)$, whenever S and N are polynomial in d_h. Substitution into the preceding bound gives $\kappa = O\left(\sqrt{\frac{d_h^2}{d_h \log d_h}\log d_h} + \log d_h\right)$, and therefore $\kappa = O(\sqrt{d_h})$. This bound is sufficient to guarantee $\|W_{OV}\|_{\mathrm{op}} \le 1$.

D Robust Decoding

We now combine the attention retrieval guarantee with the decoder margin. The operator norm bound from Section C ensures that the normalized decoder $W_{OV} = \frac{W'_{OV}}{\kappa}$ maps every unit representation vector to a vector of norm at most one. The retrieval result from Section 5.2 then implies that the actual attention output is close to the ideal target value. Finally, we show that this perturbation does not change the identity of the maximizing output token whenever the perturbation is smaller than half the ideal decoding margin.

Bounded Value Norms: We eecall that $W_{OV} = \frac{W'_{OV}}{\kappa}$ and that, on the event $\|W'_{OV}\|_{\mathrm{op}} \le \kappa$, we have $\|W_{OV}\|_{\mathrm{op}} \le 1$. For every sequence t and position s, we define the corresponding value vector by $v_s^t := W_{OV}x_s^t$. Since $\|x_s^t\|_2 = 1$, we obtain $\|v_s^t\|_2 \le \|W_{OV}\|_{\mathrm{op}}\|x_s^t\|_2 \le 1$. In particular, the ideal target value $v_{j_t}^t = W_{OV}x_{j_t}^t$ also satisfies $\|v_{j_t}^t\|_2 \le 1$.

Attention Output Is Close to the Ideal Value: For sequence t, the actual attention output is $z^t = \sum_{s=1}^S \alpha_s^t v_s^t$. On the retrieval event, $\alpha_{j_t}^t \ge 1 - \epsilon_1$, while on the operator norm event $\|v_s^t\|_2 \le 1$ for every s. Hence, $\|z^t - v_{j_t}^t\|_2 \le (1 - \alpha_{j_t}^t)\|v_{j_t}^t\|_2 + \sum_{s \ne j_t} \alpha_s^t \|v_s^t\|_2 \le 2\epsilon_1$.

Sufficient Condition for Decoding: By the decoder margin, $\Delta_{\mathrm{dec}} = \frac{1-2\tau_2}{\kappa}$. For any $j \ne y_t$, $\langle u_{y_t} - u_j, z^t\rangle = \langle u_{y_t} - u_j, v_{j_t}^t\rangle + \langle u_{y_t} - u_j, z^t - v_{j_t}^t\rangle$. Since $\|u_{y_t} - u_j\|_2 \le 2$ and $\|z^t - v_{j_t}^t\|_2 \le 2\epsilon_1$, $\langle u_{y_t} - u_j, z^t\rangle \ge \Delta_{\mathrm{dec}} - 4\epsilon_1$. Therefore decoding is guaranteed whenever $\tau_2 < \frac{1}{2}$ and $\epsilon_1 < \frac{1-2\tau_2}{4\kappa}$.

Simultaneous Decoding Guarantee: Combining the retrieval event, decoder margin event, and operator norm event, the preceding bounds imply decoding simultaneously for every $t \in [T]$ whenever $\tau_1 < \frac{1}{2}$, $\tau_2 < \frac{1}{2}$, $\epsilon_1 < \frac{1-2\tau_2}{4\kappa}$. A union bound over the three failure events gives probability at least $1 - \delta_1 - \delta_2 - \delta_3$.

E Capacity Scaling

We now combine the retrieval and decoder concentration bounds to determine the admissible scaling of the number of stored sequences T with the head dimension d_h. The retrieval analysis requires $\tau_1 < \frac{1}{2}$, where $\tau_1 = C_1\left(\frac{\sqrt{T\ell_1}}{d_h} + \sqrt{\frac{\ell_1}{d_h}}\right)$, $\ell_1 = \log\left(\frac{c_1 TS}{\delta_1}\right)$.

Similarly, the decoder analysis requires $\tau_2 < \frac{1}{2}$, where $\tau_2 = C_2\left(\frac{\sqrt{T\ell_2}}{d_h} + \sqrt{\frac{\ell_2}{d_h}}\right), \qquad \ell_2 = \log\left(\frac{c_2 TN}{\delta_2}\right)$.

E.1 Retrieval Based Constraint

The condition $\tau_1 < \frac{1}{2}$ requires $C_1\left(\frac{\sqrt{T\ell_1}}{d_h} + \sqrt{\frac{\ell_1}{d_h}}\right) < \frac{1}{2}$. Since both terms are nonnegative, a sufficient condition is that each term is bounded by a sufficiently small absolute constant. Also, there exist absolute constants $c, c' > 0$ such that $\frac{T\ell_1}{d_h^2} \le c$ and $\frac{\ell_1}{d_h} \le c'$. The first condition is the dominant constraint when T grows with d_h. Thus, $T \log\left(\frac{c_1 TS}{\delta_1}\right) \le c d_h^2$.

E.2 Decoder Based Constraint

Similarly, the condition $\tau_2 < \frac{1}{2}$ is implied by $\frac{T\ell_2}{d_h^2} \le c$ and $\frac{\ell_2}{d_h} \le c'$ for sufficiently small absolute constants. Hence, $T \log\left(\frac{c_2 TN}{\delta_2}\right) \le c d_h^2$. Thus both retrieval and decoding require a common admissible value of T satisfying $T \log\left(\frac{c_1 TS}{\delta_1}\right) \le d_h^2$ and $T \log\left(\frac{c_2 TN}{\delta_2}\right) \le d_h^2$.

A Common Sufficient Scaling: Let $M := \max\{S, N\}$ and define $L := \log\left(\frac{C_0 d_h M}{\delta}\right)$ for a sufficiently large absolute constant $C_0 > 0$, where $\delta := \min\{\delta_1, \delta_2\}$. Suppose that $T \le c_0 \frac{d_h^2}{L}$ for a sufficiently small absolute constant $c_0 > 0$. We first verify that the logarithmic factors ℓ_1 and ℓ_2 are of order L under this choice. Since $T \le c_0 \frac{d_h^2}{L}$, we have $\ell_1 = \log\left(\frac{c_1 TS}{\delta_1}\right) \le \log\left(\frac{c_1 c_0 d_h^2 S}{\delta_1 L}\right)$. Similarly, $\ell_2 = \log\left(\frac{c_2 TN}{\delta_2}\right) \le \log\left(\frac{c_2 c_0 d_h^2 N}{\delta_2 L}\right)$. Hence, up to absolute constants and lower order logarithmic terms, $\ell_1, \ell_2 = O\left(\log\left(\frac{d_h^2 M}{\delta}\right)\right)$. In particular, for polynomially bounded S and N in d_h and fixed failure probabilities, $\ell_1, \ell_2 = O(\log d_h)$. Substitution into the definitions of τ_1 and τ_2 yields $\tau_1 = O\left(\frac{\sqrt{T \log(d_h M/\delta)}}{d_h} + \sqrt{\frac{\log(d_h M/\delta)}{d_h}}\right)$, and $\tau_2 = O\left(\frac{\sqrt{T \log(d_h M/\delta)}}{d_h} + \sqrt{\frac{\log(d_h M/\delta)}{d_h}}\right)$. For sufficiently small c_0, the first terms are bounded by a constant strictly smaller than $1/2$, while the second terms vanish whenever $\log\left(\frac{d_h M}{\delta}\right) = o(d_h)$. Thus, in this regime, both $\tau_1 < \frac{1}{2}$ and $\tau_2 < \frac{1}{2}$ holds simultaneously.

Memorization Scale The preceding argument shows that the construction supports a number of stored sequences of order $T = \Omega\left(\frac{d_h^2}{\log(d_h M/\delta)}\right)$ is achievable by the explicit construction, up to absolute constants, provided the logarithmic factor satisfies $\log\left(\frac{d_h M}{\delta}\right) = o(d_h)$ and the remaining feasibility conditions hold. We recall that $M = \max\{S, N\}$, we obtain the scaling $T = \Omega\left(\frac{d_h^2}{\log(d_h \max\{S,N\}/\delta)}\right)$ is achievable by the construction.

More precisely, the concentration inequalities impose an upper bound on admissible T of this order. Constructing the Transformer at any constant factor smaller value of T therefore yields the corresponding lower bound on the memorization capacity.

Finally, the pairwise distinct target token condition requires $T \le N$. Thus, the overall construction also obeys the trivial vocabulary constraint

$$T \le \min\left\{N,\, C\frac{d_h^2}{\log(d_h \max\{S,N\}/\delta)}\right\}.$$

Polynomial Sequence Length and Vocabulary: When S and N are polynomial in d_h and the failure probability is fixed, we have $\log(d_h \max\{S, N\}/\delta) = \Theta(\log d_h)$. Then, $T = \Omega\left(\frac{d_h^2}{\log d_h}\right)$ is achievable up to constant factors, subject to the vocabulary constraint $T \le N$. Thus the attention and decoder concentration arguments establish an achievable quadratic over logarithmic memorization scale in the head dimension.

F MEMORIZATION THEOREM

Proof. We consider the three probabilistic events established in the preceding sections. First, by Proposition 1 and the softmax bound above, with probability at least $1-\delta_1$, $\alpha_{j_t}^t \geq 1-\epsilon_1$ simultaneously for every $t \in [T]$.

Second, by Proposition 3, with probability at least $1-\delta_2$, $\langle u_{y_t}, W'_{OV} x_{j_t}^t \rangle - \langle u_j, W'_{OV} x_{j_t}^t \rangle \geq 1-2\tau_2$ simultaneously for every $t \in [T]$ and every $j \neq y_t$. Third, by the operator norm result of Section C with probability at least $1-\delta_3$, $\|W'_{OV}\|_{\mathrm{op}} \leq \kappa$. On the third event, $W_{OV} = \frac{W'_{OV}}{\kappa}$ satisfies $\|W_{OV}\|_{\mathrm{op}} \leq 1$. Also, $\|v_s^t\|_2 \leq 1$ for every t, s. Together with the retrieval event, this results $\|z^t - v_{j_t}^t\|_2 \leq 2\epsilon_1$. Moreover, the decoder margin event implies $\langle u_{y_t}, v_{j_t}^t \rangle - \langle u_j, v_{j_t}^t \rangle \geq \frac{1-2\tau_2}{\kappa}$ for every $j \neq y_t$. Hence, for every such j,

$$\begin{aligned}\langle u_{y_t} - u_j, z^t \rangle &\geq \frac{1-2\tau_2}{\kappa} - 2\|u_{y_t} - u_j\|_2 \epsilon_1 \\ &\geq \frac{1-2\tau_2}{\kappa} - 4\epsilon_1 > 0.\end{aligned}$$

Therefore $y_t = \arg\max_{j \in [N]} \langle u_j, z^t \rangle$ for every $t \in [T]$. Finally, applying the union bound to the three failure events gives $\Pr(\text{all three events hold}) \geq 1-\delta_1-\delta_2-\delta_3 \geq 1-\delta_{\mathrm{fail}}$.. □

F.1 A CAPACITY LOWER BOUND

The preceding theorem reduces the existence of a memorizing Transformer to the simultaneous feasibility of the retrieval and decoder conditions. In particular, suppose that $T \leq c\frac{d_h^2}{\log(d_h \max\{S,N\}/\delta)}$ for a sufficiently small absolute constant $c > 0$, and that the remaining parameters ϵ_1, β, and κ satisfy the conditions of Theorem 2. Then the construction memorizes all T stored associations with probability at least $1-\delta_{\mathrm{fail}}$. Also, if $\mathcal{C}(d_h, S, N, \delta_{\mathrm{fail}})$ denotes the memorization capacity of the considered attention-only Transformer, then $\mathcal{C}(d_h, S, N, \delta_{\mathrm{fail}}) \geq c\frac{d_h^2}{\log(d_h \max\{S,N\}/\delta)}$ whenever the feasibility conditions of Theorem 2 hold.

When S and N are polynomially bounded in d_h and the failure probability is fixed, this reduces to $\mathcal{C}(d_h) = \Omega\left(\frac{d_h^2}{\log d_h}\right)$. The bound above is a lower bound on the memorization capacity, the concentration analysis determines a scale of T for which theexplicit construction succeeds, and therefore establishes that the model can memorize at least that many associations.

Table 1: Comparison of memorization-capacity results for attention and Transformers. Unified notation: T = number of memorized examples/associations, S = context length, N = vocabulary size, d = embedding dimension, H = number of heads, d_h = head dimension, d_v, d_{out} = value and output dimensions, m = MLP width, C = number of classes, n = training-sample size. For Nichani et al. (2025): $\mathcal{S}, \mathcal{R}$ = subject and relation sets. For Cabannes et al. (2024): M = number of classes, $p(x)$ = token frequency, η = step size, α = input-embedding correlation. $\widetilde{O}, \widetilde{\Omega}, \gtrsim, \lesssim$ hide polylog factors.

Model	Setting	Data / embedding assumption	Memorization capacity	#Parameters	Lower bound / tightness
Transformer (attn. + FFN); depth $\widetilde{O}(S+\sqrt{ST})$, width 16 (1 head, d_h=1) Kim et al. (2023)	Seq-to-seq (label per token); seq. classification	Token-wise (r,δ)-separated; distinct inputs; perm.-equivariant case also needs distinct contexts and consistent labels	T sequences of length S	$\widetilde{O}(d+S+\sqrt{ST})$; $\widetilde{O}(d+S+\sqrt{T})$ for classification; bit compl. $\widetilde{O}(\sqrt{ST})$	None on #params; optimal in total bits, $\Omega(ST)$, when $d, S = O(T)$
1-layer multi-head attention, no FFN Mahdavi et al. (2024)	Exact real-valued label $y \in \mathbb{R}^{d_{\text{out}}}$ per example; self- or cross-attention	Query set has Kruskal rank $\geq S$; each context has rank S; $S < d$	$T = H(\min(S, d_h)-1)+1$	$2Hdd_h + Hdd_v + d\,d_{\text{out}}$ ($\Theta(Hd^2)$ for d_h=d)	Shared contexts: $\operatorname{rank}(Z) \leq H(S-1)+1$, so the rank argument is tight for $d_h \geq S$; general capacity upper bound open
1-layer attention-only, H heads, skip connection Dana et al. (2024)	Next-token (last position); any S	Deterministic next token; $d_h \leq d$; embeddings *constructed* from algebraically independent reals, not random	$T = Hd_h + d$ for $d \geq 2$ (Thm. 1 with ε=0; Cor. 1 states d=2); KL bound for distribution memorization	$d(S+2N+4Hd_h)$; $d(S+2N+2(d_h+1)H)$ with rank-1 W_{QK}	No matching bound. Empirically, accuracy is linear in H; in d_h it is quadratic at d=10, reported linear at d=50
1-layer multi-head self-attention (+ optional MLP) Nichani et al. (2025)	Factual recall: $(s,r) \mapsto a^*$ among noise tokens; query = fixed EOS token	i.i.d. uniform-sphere embedding per vocabulary token; disjoint answer sets, $D = \max_r \lvert\mathcal{A}_r\rvert$	All $\lvert\mathcal{S}\rvert\lvert\mathcal{R}\rvert$ facts. Attn-only: $d \geq \widetilde{\Omega}(\max(\lvert\mathcal{R}\rvert, D))$, $Hd_h \geq \widetilde{\Omega}(\lvert\mathcal{S}\rvert + \lvert\mathcal{R}\rvert)$. With MLP: $Hd_h \geq \widetilde{\Omega}(\lvert\mathcal{S}\rvert + \lvert\mathcal{R}\rvert)$, $m \geq \widetilde{\Omega}(\max_a \lvert a^{*-1}(a)\rvert)$, $md \geq \widetilde{\Omega}(\lvert\mathcal{S}\rvert\lvert\mathcal{R}\rvert)$, polynomial activation	$4Hdd_h$ (attn.) $+ 2md$ (MLP)	Zero loss needs $\geq \lvert\mathcal{S}\rvert\lvert\mathcal{R}\rvert \log D + \Omega(\lvert\mathcal{R}\rvert D \log(N/D))$ bits $\Rightarrow$ tight up to logs (attn-only: when $\lvert\mathcal{R}\rvert \geq D$)
1 self-attention layer (used as averaging) between deep FFNs; width 14, depth $\widetilde{O}(\sqrt{T})$ (next-token), $\widetilde{O}(\sqrt{ST})$ (seq-to-seq) Kajitsuka & Sato (2025)	Next-token; seq-to-seq	Token-wise (r,δ)-separated; consistent labels; $S, C, r/\delta = T^{O(1)}$	T sequences	Next-token: $\widetilde{O}(d+\sqrt{T})$; seq-to-seq: $\widetilde{O}(\sqrt{ST})$ for $d = \widetilde{O}(\sqrt{ST})$	$\Omega(\sqrt{T})$ (next-token); $\Omega\big(\sqrt{ST/\log(ST)}\big)$ (seq-to-seq, hardmax) $\Rightarrow$ FFN is the bottleneck
1-layer, single head, attention-only or attention + MLP; trained by 3-step empirical GD Vural et al. (2026)	Retrieve trigger-marked token among S i.i.d. tokens; query = EOS; learn permutation	Gaussian $\mathcal{N}(0, 1/d)$ embeddings; $S = N^c$, $c \in (0,1)$; $N \log N \lesssim n = o(NS)$	N associations learned (acc. $1-o(1)$) if attn-only $d^2 \gtrsim \max\{N,\ N^{2/3}S^{8/3}/n^{4/3}\}$; attn. + MLP needs $md \gtrsim N$ plus extra conditions	Parameter size d^2 (attn-only) or md (MLP); trainable W_{KQ}; W_{in} fixed	Gradient-query methods fail if $n \lesssim N \min\{1, S/d^2\}$ (sample-complexity bound; not matching)
Linear associative memory $W \in \mathbb{R}^{d\times d}$; GD/GF on cross-entropy Cabannes et al. (2024)	N input tokens $\to M$ classes, frequencies $p(x)$	Fixed embeddings: orthogonal, or α-correlated (two-token analysis); random spherical in numerics	$d \geq N$, lin. indep. e_x: all N learned, margins $\sim \log t$; loss $O(1/\eta t)$ (orthogonal, M=2, W_0=0); $d < N$: CE minimizer can forget the most frequent association	d^2	Dynamics only; one-step loss spike $\geq \eta(\alpha p_1 - p_2)p_2$ when $\alpha p_1 > p_2$
Single attention head [**Ours**]		i.i.d. random spherical representation per (sequence, position); independent query per sequence	$T = \Omega(d_h^2/\log d_h)$ when $S, N = \operatorname{poly}(d_h)$		