INFINITE GRAPHS AS LIMITS ON ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

A finite-dimensional attention head cannot realize the random graph. Modeling a hard-attention head as a threshold dot-product graph, we show that a head of dimension d fails the $(d+1)$-extension property at every threshold. The Rado graph, in which every relational query has a witness, therefore lies beyond every finite dimension. Forster's sign-rank bound turns the barrier into a rate: a generic relation on n tokens forces $d = \Omega(\sqrt{n})$, independently of how the token embeddings are chosen. Grading a head by the largest query it answers gives a hierarchy. Answering every query on k tokens needs at least 2^k tokens and dimension at least k. Conversely, with $n \geq 2^{Ck}$ tokens for a constant C, a head with independent Gaussian embeddings of dimension $O(k \log n)$ suffices with high probability. At an exponential token budget, that Gaussian head fails with high probability for each dimension $k \leq d \leq ck^2/\log k$: some k of its directions have a positive cone too small for any available token to hit. For directed heads, with queries and keys separate, the barrier holds on tournaments and the Gaussian bounds transfer for $d = \Omega(k)$. Whether some head answers all k-token queries at dimension linear in k is open.

1 INTRODUCTION

Thresholding the scores of a finite-dimensional attention head cannot produce the Rado graph. In the symmetric model studied here, token u is adjacent to token v when their embedding inner product clears a fixed threshold; the realizable graphs are the threshold dot-product graphs of that dimension (Reiterman et al., 1989; Fiduccia et al., 1998). The Rado graph R, the unique countable graph in which every finite relational query has a witness (Erdős & Rényi, 1963; Rado, 1964; Cameron, 1997), provides a benchmark for an in-context lookup relation that answers every such query. We show that no finite-dimensional head is even locally as expressive: a d-dimensional head fails the $(d+1)$-extension property, by a shattering count. The barrier comes from the inner-product threshold relation; finite dimension alone does not cause it. By contrast, for graphs built from distances the random geometric graph on a countable dense subset of ℓ_∞^d almost surely has an extension property (Bonato & Janssen, 2011).

This impossibility has quantitative consequences. Sign-rank gives a dimensional rate. Grading a head by the size of the relational query it answers then yields a hierarchy with costs in both tokens and embedding dimension. The tools are classical: sign-rank and Forster's bound (Forster, 2002), the VC dimension of halfspaces (Vapnik & Chervonenkis, 1971; Sauer, 1972; Shelah, 1972), and the geometry of Gaussian vectors.

1.1 OUR RESULTS

The Extension Barrier. The first result is qualitative and holds for every token set.

Theorem 1.1 (restated as Theorem 3.1)**.** *A d-dimensional threshold head fails the $(d+1)$-extension property, whatever its threshold. Consequently the Rado graph is not realizable by a head of any finite dimension.*

The shattering count transfers to the asymmetric head with separate queries and keys when the induced relation is a tournament.

Proposition 1.2 (restated as Theorem 3.3)**.** *No finite-dimensional directed head realizes the countable random tournament.*

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.

The Precision–Context Law. Sign-rank turns the barrier into a rate. A head of dimension d realizes a graph on n vertices only if some diagonal completion of its adjacency sign matrix has sign-rank at most $d+1$, and Forster's spectral bound then limits how much relational content a head can carry.

Theorem 1.3 (restated as Theorem 4.1)**.** *A graph on n vertices realized by a d-dimensional head satisfies $d+1 \geq n/\left(\|H\|+2\right)$ for every sign matrix H agreeing with its adjacency off the diagonal, where $\|\cdot\|$ is the spectral norm. In particular the graph defined off the diagonal by a symmetric Hadamard matrix forces $d \geq \sqrt{n}-3$, and a uniform random graph almost surely forces $d \geq \left(\frac{1}{2}-o(1)\right)\sqrt{n}$.*

Read the other way, a d-dimensional head carries a generic relation on at most $O\left(d^2\right)$ tokens; Theorem 4.3 says what the trade-off does and does not cover.

The Witnessing Budget. Between a fixed query size and a fixed token set lies a hierarchy that grades a head by the largest query it answers. Write C_k for the k-extension property, and call a head *Gaussian* when its threshold is zero and its embeddings are independent standard Gaussian vectors.

Theorem 1.4 (restated as Theorem 5.4)**.** *Let $k \geq 2$. A d-dimensional head with C_k has $d \geq k$, whatever its threshold, and needs at least $2^k + k$ tokens. Conversely there is an absolute constant C such that a Gaussian head on $n \geq 2^{Ck}$ tokens of any dimension $d \geq Ck \log n$ has C_k with probability at least $1 - 1/n$.*

At the minimal token budget $n = 2^{\Theta(k)}$ the upper bound reads $d = O\left(k^2\right)$ against a lower bound of k. The main technical result shows that the quadratic upper bound is tight up to a logarithmic factor for the Gaussian construction.

Theorem 1.5 (restated as Theorem 5.7)**.** *There are absolute constants γ_0 and $c > 0$ such that, for all large k, a Gaussian head on any even number $n \geq 2^{\gamma_0 k}$ of tokens fails C_k with probability $1 - o(1)$ at every dimension d with $2C_0 k \leq d \leq ck^2/\log k$, where C_0 is an absolute constant.*

Theorem 5.8 extends the failure down to $d \geq k$, the trivial bound of Theorem 1.4, so a Gaussian head of linear dimension at this budget has C_k with probability only $o(1)$. The mechanism is a cone degeneracy. Among $2^{\Theta(k)}$ random directions in dimension $\Theta(k)$ there is, with high probability, a k-subset in which one vector nearly opposes the average of the other $k-1$. The positive sign-cone of that subset, the set of directions at positive inner product with each of its members, has measure at most $2^{-\Omega(k^2)}$. That is far too small for any of the $2^{\Theta(k)}$ available tokens to fall inside, so the query asking for a common neighbor of the subset has no answer. The budget hierarchy transfers to the directed setting.

Proposition 1.6 (restated as Theorem 5.12)**.** *Both the Gaussian upper bound of Theorem 1.4 and the quadratic failure of Theorem 1.5 hold for the directed Gaussian head, with the queries and the keys playing separate roles.*

1.2 Our Techniques

Shattering, and the Diagonal. The extension barrier is a shattering count once one observes that a witness is a halfspace with a *fixed* threshold. Passing to sign-rank needs two pieces of care. The threshold must be perturbed to make the signs strict, and the diagonal must be completed, since a graph fixes its sign matrix only off the diagonal while sign-rank is defined for a full matrix. Two completions of one off-diagonal pattern differ by a diagonal matrix of spectral norm at most 2, so the choice costs little.

Sign-Cone Measures. Much of Section 5 turns on one quantity: the probability that a random direction realizes a prescribed sign pattern against a fixed set of k vectors. For the upper bound, Theorem 5.3 bounds this probability below by e^{-k} when the k vectors are well conditioned. For the lower bound, Theorem 5.5 bounds it above by $e(C_\kappa \sin\theta)^{k-1}$ when one vector lies within angle θ of opposing the average of the others, where C_κ depends only on how well the others are conditioned.

Finding a Degenerate Subset. Theorem 5.7 needs a k-subset of tokens whose cone is tiny. Opposite the averages of $(k-1)$-subsets of half the tokens lie $\binom{n/2}{k-1} = 2^{\Theta\left(k^2\right)}$ candidate directions;

among them, one comes close enough to another token's direction for the cone estimate to apply. Turning the count into an existence proof requires a second moment over pairs of subsets. The case that decides it is the one in which two overlapping subsets are completed by the same vector: the overlap correlates their average directions, and that one vector must nearly oppose both. Theorem 5.6 says how much likelier two such directions are to point the same way than two independent ones: at correlation ρ their joint density is at most $((1+\rho)/(1-\rho))^{d/2}$.

1.3 Other Related Work

Dot-Product Representations. Theorem 2.7 is the dot product representation of a graph, and the least dimension realizing a given graph is its dot product dimension, introduced by Reiterman et al. (1989) and developed by Fiduccia et al. (1998). Lee et al. (2024) study the same representation as a graph decoder and prove sign-rank lower bounds for stars and grids.

Expressivity of Attention. Sanford et al. (2023) show that their triple-matching task on length-N sequences over an alphabet of size at least $N+1$ requires $mpH = \Omega(N/\log\log N)$ in a one-layer transformer, where m is embedding dimension, p bit precision and H head count (Theorem 7). Their barrier and ours are complementary: theirs concerns a specific task and the full architecture, ours a relational completeness property of a single head. In the query–key model of Adler (2025), a layer that recovers the in-context neighbors of every directed graph with m' edges on m items needs total key dimension $\Omega\left((m'/d_{\text{model}})\log(m^2/m')\right)$, where d_{model} is the embedding dimension. The bound assumes $O(1)$ effective bits per parameter, or a fixed positive margin and bounded parameter and embedding norms. That bound counts over all graphs with m' edges, where Theorem 4.1 certifies the dimension a single named graph needs. Rajaraman et al. (2026) bound the number of heads rather than their dimension: in a one-layer attention-only transformer, k heads solve k-bit parity and cannot solve $(k+1)$-bit parity, whatever the embedding dimension and precision.

Sign-Rank. Forster's theorem (2002) remains the workhorse, and the relation between sign-rank and VC dimension is by now well mapped (Alon et al., 2016). Weller et al. (2026) relate dimension to sign-rank as Theorem 1.3 does, for embedding-based retrieval. With one global threshold, the least dimension at which inner products reproduce a relevance matrix A lies between the sign-rank of $2A - J$ minus one and that sign-rank, where J is the all-ones matrix. The known methods also have limits. Hatami et al. (2022) exhibit matrices of polynomially large sign-rank for which bounds through VC dimension, Forster's margin bound, and the density of the largest monochromatic rectangle are all constant. Section 1.5 explains how this bears on Theorem 5.10.

Geometric Graphs and the Rado Property. For graphs built from distances rather than inner products, geometry is known to obstruct the extension property in qualitative form. In every finite-dimensional normed space not isometric to ℓ_∞, almost every dense set is strongly non-Rado, meaning that no single graph arises almost surely (Balister et al., 2018). In ℓ_∞^d the random geometric graph on a countable dense set is almost surely one fixed graph, and that graph has an extension property (Bonato & Janssen, 2011); this is the contrast drawn at the start of the introduction. Theorems 3.1 and 5.7 are quantitative where these verdicts are qualitative.

Existentially Closed Graphs. The finite analogue of the extension property is k-existential closure. Bonato (2009) surveys it for graphs, Bonato et al. (2010) give bounds and constructions for tournaments, and Paley graphs are the standard explicit family with it (Blass et al., 1981). Theorem 5.11 uses Paley graphs to show that the sign-rank of a k-e.c. graph can be as large as $2^{\Omega(k)}$.

1.4 Discussion

Two readings of these results are worth separating. The first is a capacity statement, made precise in Theorem 3.2. The VC dimension of the head's halfspace family bounds which relational queries it can answer, regardless of the parameter count or context length. Theorem 3.1 shows, through the Rado graph, that this capacity, finite for every head, falls short of relational universality, and Theorem 4.1 makes the shortfall quantitative for a named graph, certifying $d = \Omega(\sqrt{n})$. The second is about scaling. By Theorems 5.4 and 5.7, relational completeness at level k costs a token budget exponential in k and, for the Gaussian construction at that budget, dimension between $\Omega(k^2/\log k)$

and $O(k^2)$. The near-quadratic lower bound is specific to the Gaussian construction; Theorem 5.10 asks whether a structured head can achieve the same completeness at linear dimension.

1.5 Open Problems

A Structured Head at Linear Dimension. Theorem 5.10 asks for a head at threshold 0, of dimension $O(k)$, with the k-extension property on $2^{\Theta(k)}$ tokens. By Theorem 5.11, such a head would give a k-existentially closed graph, one with the k-extension property, on $2^{\Theta(k)}$ vertices with a diagonal completion of sign-rank $O(k)$, so a negative answer would follow from showing that no such graph exists. The known sign-rank lower bounds may not reach that question: the three known methods of Section 1.3 fail together on some matrices of polynomially large sign-rank (Hatami et al., 2022). A positive answer needs a construction, and none of the three candidates we examined is known to give one. Gaussian embeddings provably produce a degenerate cone by Theorems 5.7 and 5.8. The two structured candidates, points on the moment curve and embeddings built from algebraic codes, appear to fail for separate reasons described in Appendix A.

The Logarithmic Gap in the Gaussian Bound. Theorems 5.4 and 5.7 pin the Gaussian dimension at the minimal budget between $ck^2/\log k$ and $O(k^2)$. The logarithm enters the proof of Theorem 5.7 in two places: the angle hypothesis $\sin\theta \le \sqrt{\kappa/m}/16$ of Theorem 5.5, and the pairs of $(k-1)$-subsets that overlap in almost all of their elements in the second moment. Removing it would need a cone bound valid at larger angles and a different treatment of heavily overlapping pairs. Where in the range from $k^2/\log k$ to k^2 the Gaussian dimension at the minimal budget lies is open.

An Invariant for the Directed Case. Theorem 5.13 shows that order dimension does not bound the dimension of a directed head from below, since a directed head of dimension 4 already realizes posets of every order dimension. A second candidate comes from Rajkumar et al. (2022). The *flip class* of a tournament T is the set of tournaments obtained from T by reversing every arc across some cut of its vertex set and then relabelling the vertices. For each tournament in the flip class of T and each ordering of its vertices, count the vertices that are the tail of at least one arc pointing backward in the ordering; $\mu(T)$ is the least such count.

They prove that every tournament T is realizable in dimension $2(\mu(T)+1)$ in the skew-symmetric specialization of Theorem 2.8. There each key is the image of its query under a fixed skew-symmetric linear map and the threshold is 0, so the induced digraph is a tournament whenever no score $\langle x_u, y_v\rangle$ with $u \neq v$ vanishes. This bound does not reverse: for every $m \ge 2$ a skew-symmetric head of dimension 6 realizes a tournament on $n = 4m$ vertices with $\mu(T) \ge n/4$, as Appendix A shows. From below, the dimension of a directed head is bounded by the extension barrier of Theorem 3.3 and by sign-rank, through the rank argument in the proof of Theorem 4.1. We know of no combinatorial invariant of a tournament that determines it up to a constant factor.

2 Preliminaries

2.1 Graphs with the Extension Property

A relational query names some tokens a witness must be adjacent to and some it must not be; the extension property asks that every such query of bounded size have a witness.

Definition 2.1. A graph G on vertex set V has the *k-extension property* if for every pair of disjoint sets $A, B \subseteq V$ with $|A \cup B| \le k$ there is a vertex $z \in V \setminus (A \cup B)$ adjacent to every vertex of A and to no vertex of B. It has the *extension property* if it has the k-extension property for every $k \in \mathbb{N}$.

A finite graph with the k-extension property is also called k-existentially closed, or k-e.c., and the k-extension property implies the j-extension property for every $j \le k$. A graph is *relationally universal* if it has the extension property; this is the property Theorem 3.1 shows no finite-dimensional head can have.

Fact 2.2 (Erdős & Rényi, 1963; Rado, 1964; Cameron, 1997; Hodges, 1993)**.** *Up to isomorphism there is a unique countable graph with the extension property, the* Rado graph R*. It is the Fraïssé limit of the class of finite graphs and is homogeneous: every isomorphism between finite induced*

subgraphs extends to an automorphism of R. A countable random graph, each edge included independently with probability $\frac{1}{2}$, is almost surely isomorphic to R.

The almost sure statement is due to Erdős & Rényi (1963), the explicit construction to Rado (1964), and uniqueness and homogeneity are Fraïssé's theorem (Hodges, 1993); Cameron (1997) surveys the graph. The directed counterpart is the countable random tournament. A digraph D on V has the *directed k-extension property* if for every pair of disjoint $A, B \subseteq V$ with $|A \cup B| \leq k$ there is a vertex $z \notin A \cup B$ with $z \to a$ for every $a \in A$ and $b \to z$ for every $b \in B$. A *tournament* on V is a digraph that orients exactly one of uv and vu for every pair of distinct $u, v \in V$.

Fact 2.3 (Hodges, 1993; Lachlan, 1984; Macpherson, 2011)**.** *Up to isomorphism there is a unique countable tournament with the directed k-extension property for every k, the* countable random tournament. *It is the Fraïssé limit of the class of finite tournaments, is homogeneous, and is almost surely isomorphic to the tournament obtained by orienting each pair independently and uniformly.*

Existence and uniqueness are again Fraïssé's theorem (Hodges, 1993); Lachlan (1984) classifies the countable homogeneous tournaments, of which this is the generic one, and Macpherson (2011) surveys the area. The following classical count is what makes a token budget exponential in k the minimal one throughout Section 5.

Fact 2.4 (Bonato, 2009; proved in Appendix A)**.** *If a graph G on vertex set V has the k-extension property then $|V| \geq 2^k + k$. The same bound holds for a tournament with the directed k-extension property.*

The tournament hypothesis in the directed half of Theorem 2.4 cannot be dropped. The complete digraph on $k+1$ vertices, with both arcs between every pair, has the directed k-extension property: any A, B with $|A \cup B| \leq k$ leave a vertex z outside them, and every arc to and from z is present.

2.2 VC Dimension and Sign-Rank

For a class $\mathcal{H}$ of subsets of a ground set X, a finite $S \subseteq X$ is *shattered* if every subset of S arises as $H \cap S$ for some $H \in \mathcal{H}$. The VC dimension $\mathrm{VC}(\mathcal{H})$ is the size of the largest shattered set (Vapnik & Chervonenkis, 1971; Sauer, 1972; Shelah, 1972). The head's witnesses are halfspaces at a threshold that is fixed once and for all, with only the normal vector free, and the VC dimension of that class is d at every threshold.

Fact 2.5 (Wenocur & Dudley, 1981; proved in Appendix A)**.** *Let $d \geq 1$, fix $\tau \in \mathbb{R}$, and put $H_w = \{x \in \mathbb{R}^d : \langle x, w \rangle \geq \tau\}$ and $\mathcal{H}_\tau = \{H_w : w \in \mathbb{R}^d\}$. Then $\mathrm{VC}(\mathcal{H}_\tau) = d$, so no set of $d+1$ points of $\mathbb{R}^d$ is shattered by $\mathcal{H}_\tau$. The larger class of all $H_{w,\tau'} = \{x : \langle x, w \rangle \geq \tau'\}$, with w and τ' both free, has VC dimension $d+1$.*

The upper bound is the affine case of a theorem of Dudley (Wenocur & Dudley, 1981), and the second statement is the count of Cover (1965) and of Vapnik & Chervonenkis (1971).

The *sign-rank* $\mathrm{sr}(M)$ of a sign matrix $M \in \{-1, +1\}^{m \times n}$ is the least rank of a real matrix N with $\mathrm{sign}(N_{ij}) = M_{ij}$ for all i, j (Paturi & Simon, 1986; Alon et al., 1985). Forster's theorem bounds sign-rank below in terms of the spectral norm.

Fact 2.6 (Forster, 2002)**.** *For every $M \in \{-1, +1\}^{n \times n}$ one has $\mathrm{sr}(M) \geq n / \|M\|$, where $\|M\|$ is the spectral norm.*

Theorem 2.6 is strongest where $\|M\|$ is smallest, and every sign matrix has $\|M\|^2 \geq \|M\|_F^2 / n = n$; a Hadamard matrix attains $\|M\| = \sqrt{n}$, and a uniform random symmetric sign matrix has $\|M\| = (2 + o(1))\sqrt{n}$ almost surely, by the Bai–Yin theorem (1988).

2.3 Attention Heads as Threshold Dot-Product Graphs

We model a hard-attention head by the graph that its thresholded scores select.

Definition 2.7. A *d-dimensional threshold head* on a token set V is a map $\varphi \colon V \to \mathbb{R}^d$ together with a threshold $\tau \in \mathbb{R}$. It induces the graph $G(\varphi, \tau)$ on V with $u \sim v$, for $u \neq v$, if and only if $\langle \varphi(u), \varphi(v) \rangle \geq \tau$. A graph G is *d-realizable* if $G \cong G(\varphi, \tau)$ for some d-dimensional threshold head.

This is the symmetric idealization of one hard-attention head, in which query and key coincide; the asymmetric case, where query and key are separate linear maps, is the directed head of Theorem 2.8.

Definition 2.8. A *d-dimensional directed head* on a token set V assigns to each $v \in V$ a query $x_v \in \mathbb{R}^d$ and a key $y_v \in \mathbb{R}^d$, together with a threshold $\tau \in \mathbb{R}$. It induces the digraph on V with $u \to v$, for $u \neq v$, if and only if $\langle x_u, y_v \rangle \geq \tau$. A digraph D is *d-realizable*, and the head *realizes* it, if D is isomorphic to the digraph induced by some d-dimensional directed head, and the same words apply to a threshold head of Theorem 2.7 and a graph.

3 The Extension Barrier, Symmetric and Directed

The barrier is local: a head of dimension d already fails at queries of size $d + 1$.

Theorem 3.1 (restatement of Theorem 1.1; proved in Appendix A)**.** *Let $d \geq 1$ and let φ and τ be a d-dimensional threshold head on a token set V. Then $G(\varphi, \tau)$ fails the $(d+1)$-extension property, whatever the threshold τ. Consequently, for every d the Rado graph R is not d-realizable, and no finite-dimensional threshold head is relationally universal.*

Remark 3.2. The failure in Theorem 3.1 is local, and that is what makes it a capacity statement. Fix any $d + 1$ tokens whatever: no d-dimensional head answers every query on them. Either their images cannot be shattered, by Theorem 2.5, or two of them coincide, and then no token separates that pair. Enlarging the token set does not help; what is bounded is not the parameter count but the VC dimension of the halfspace family the head expresses, which Theorem 2.5 puts at exactly d whatever the threshold.

Dropping symmetry replaces $G(\varphi, \tau)$ by the digraph of Theorem 2.8 and R by the countable random tournament of Theorem 2.3; the barrier transfers by the same count, now applied to the key embeddings alone.

Proposition 3.3 (restatement of Theorem 1.2; proved in Appendix A)**.** *Let $d \geq 1$. No d-dimensional directed head realizes the countable random tournament. More precisely, if the digraph induced by a d-dimensional directed head is a tournament, then it does not have the directed $(d+1)$-extension property, whatever the threshold.*

4 Sign-Rank Capacity: A Precision–Context Law

Theorem 3.1 names a query size a head cannot reach and says nothing about how large a token set a head of a given dimension can carry; sign-rank supplies the rate. A graph fixes a sign matrix only off the diagonal, whereas sign-rank is defined for a full sign matrix. The statement is therefore phrased so that the head chooses the diagonal, and the spectral norm may be read off any completion whatever, at a cost of 2 in the norm.

Theorem 4.1 (restatement of Theorem 1.3; proved in Appendix A)**.** *Let G be a graph on n vertices $v_1, \ldots, v_n$ and let $H \in \{-1, +1\}^{n \times n}$ be any sign matrix with $H_{ij} = +1$ when $v_i \sim v_j$ and $H_{ij} = -1$ when $v_i \not\sim v_j$, for all $i \neq j$; the diagonal of H is arbitrary. If G is d-realizable then $d + 1 \geq n/\left(\|H\| + 2\right)$. In particular, if H is a symmetric Hadamard matrix then $d \geq \sqrt{n} - 3$, and a uniform random graph on n vertices almost surely requires $d \geq \left(\frac{1}{2} - o(1)\right)\sqrt{n}$.*

Remark 4.2. For the random graph the bound of Theorem 4.1 is not the best available, and a counting argument gives more. A head of dimension d on n tokens is described by $nd + 1$ real parameters, including its threshold. For $d \leq n/4$, the Milnor–Thom bound on the sign patterns of $\binom{n}{2}$ quadratic forms in those parameters puts the number of realizable graphs at $2^{O(nd \log(n/d))}$, against $2^{n(n-1)/2}$ graphs in all. Taking $d = cn$ for a sufficiently small constant $c > 0$ shows that almost every graph needs $d = \Omega(n)$. The counting argument is the one Alon et al. (1985) give for set systems. Reiterman et al. (1989) give the explicit bound $d \geq n/15$ for almost every graph, for exactly the threshold dot-product dimension used here. What Theorem 4.1 adds is a *certificate*: it applies to a named graph through a computable spectral quantity, which is what Section 5 can use against an adversarial construction. The price is that $\sqrt{n}$ is the ceiling of the spectral route rather than an artifact of this proof (Hatami et al., 2022).

Remark 4.3. Read the other way, Theorem 4.1 says that a d-dimensional head can carry a generic relation on at most $O(d^2)$ tokens. Embedding dimension and context length therefore trade off

independently of parameters outside the fixed threshold head: adding those parameters does not change the graphs that head can realize. The rate this certifies is $d = \Omega(\sqrt{n})$. For a generic relation the theorem of Reiterman et al. (1989) recorded in Theorem 4.2 improves it to $d = \Omega(n)$, so a head then carries a generic relation on only $O(d)$ tokens. The bound is a statement about the induced graph alone, so any mechanism whose realized adjacency is $\langle \varphi(u), \varphi(v) \rangle \geq \tau$ inherits it; what it does not cover is a soft head whose scores are never thresholded into a selection.

5 The Witnessing Budget

Theorem 2.1 grades a head by how large a relational query it answers. Write C_k for the statement that the head has the k-extension property, and read k as a witnessing budget: C_k says that every query touching at most k tokens has an answer inside the token set. Theorem 3.1 gives the lower half of the hierarchy at once; the work is the converse, and then the question of whether the converse is tight.

Throughout this section the threshold is $\tau = 0$ and the embeddings are random. Rescaling an embedding by a positive scalar leaves the induced graph unchanged, so $\mathcal{N}(0, I_d)$ embeddings and uniform embeddings on S^{d-1} induce the same distribution on graphs, and we use whichever is convenient. The *Gaussian head* of dimension d on n tokens has $\tau = 0$ and n independent $\mathcal{N}(0, I_d)$ embeddings. We write x_v for $\varphi(v)$ throughout this section, and $x_1, \ldots, x_n$ once the tokens are indexed, since the tokens play no role beyond their embeddings.

5.1 Sign-Cone Measures and Gram Conditioning

For a finite S and unit vectors $(\hat{x}_i)_{i \in S}$ with Gram matrix $G = (\langle \hat{x}_i, \hat{x}_j \rangle)_{i,j \in S}$, and for a sign pattern $s \in \{-1, +1\}^S$, write

$$q_s(G) = \mathbf{Pr}_{g \sim \mathcal{N}(0, I_d)}[\text{sign} \langle \hat{x}_i, g \rangle = s_i \text{ for every } i \in S] \,,$$

the *sign-cone measure* of the pattern s. We write $q_+(G)$ for the all-positive pattern. Theorem 5.2 below shows the notation is legitimate, in that the probability depends on G and not on the vectors themselves.

For $\theta \in (0, \pi/2]$ let $p(\theta)$ denote the normalized surface measure of a cap of angular radius θ in S^{d-1}, so that a uniformly random unit vector lies within angle θ of a fixed unit vector with probability $p(\theta)$.

Fact 5.1 (Baraniuk et al., 2008; Candès & Tao, 2005; Foucart & Rauhut, 2013; Vershynin, 2018)**.** *There are absolute constants $C_0, c_0, \kappa > 0$ such that if $x_1, \ldots, x_m$ are independent $\mathcal{N}(0, I_d)$ vectors with $d \geq C_0 m$, then with probability at least $1 - e^{-c_0 d}$ the normalized Gram matrix G of $\hat{x}_i = x_i / \|x_i\|$ satisfies $\lambda_{\min}(G) \geq \kappa$, and moreover $\frac{1}{2}\sqrt{d} \leq \|x_i\| \leq 2\sqrt{d}$ for every i and $\|x_1 + \cdots + x_m\| \leq 2\sqrt{md}$. The same holds with κ replaced by $\frac{1}{2}$ and C_0 by a larger absolute constant $C_1 \geq C_0$.*

5.2 The Hierarchy

The upper bound of the hierarchy rests on sign-cone measures, and the first step reduces such a measure to a Gram matrix.

Proposition 5.2 (Gram reduction; proved in Appendix A)**.** *Let $\hat{x}_i$, $i \in S$, be unit vectors in $\mathbb{R}^d$ with Gram matrix G, let $|S| = m$, and let $s \in \{-1, +1\}^S$. Then $q_s(G)$ depends only on G, and not on the vectors or on the ambient dimension d. Moreover $q_s(G) = q_+(D_s G D_s)$, where $D_s = \text{diag}(s)$, and $D_s G D_s$ has the same unit diagonal and the same eigenvalues as G.*

After that reduction, a lower bound of $\frac{1}{2}$ on the least eigenvalue of the Gram matrix alone bounds every sign-cone measure from below.

Lemma 5.3 (sign-cone measure under bounded conditioning; proved in Appendix A)**.** *Let $\hat{x}_1, \ldots, \hat{x}_m \in \mathbb{R}^d$ be unit vectors whose Gram matrix G satisfies $\lambda_{\min}(G) \geq \frac{1}{2}$. Then $q_s(G) \geq e^{-m} \geq 2^{-2m}$ for every sign pattern $s \in \{-1, +1\}^m$.*

The lower half of the hierarchy follows from Theorem 3.1 and Theorem 2.4. The upper half applies Theorem 5.3 to every set of at most k tokens of a Gaussian head. Theorem 5.1 keeps those sets well conditioned.

Theorem 5.4 (restatement of Theorem 1.4; proved in Appendix A)**.** *Fix $k \geq 2$.*

1. (Lower bound.) *If a d-dimensional threshold head on n tokens satisfies C_k then $d \geq k$, whatever the threshold, and $n \geq 2^k + k$.*
2. (Upper bound.) *There is an absolute constant C such that, for $\tau = 0$ and independent $\mathcal{N}(0, I_d)$ embeddings of n tokens, if $d \geq Ck \log n$ and $n \geq 2^{Ck}$ then the head satisfies C_k with probability at least $1 - 1/n$.*

The two halves meet at $n = 2^{\Theta(k)}$, which by part (1) is the smallest token budget at which C_k is possible at all, and there the upper bound reads $d = O(k \log n) = O(k^2)$ against a lower bound of k. The constraint $n \geq 2^{Ck}$ forces $\log n = \Omega(k)$, so the bound never falls below $\Omega(k^2)$, and Theorem 5.7 shows this is not an artifact of the analysis.

5.3 The Gaussian Head Is Quadratic

The upper bound of Theorem 5.4 rests on Theorem 5.3, which sees only $\lambda_{\min} \geq \frac{1}{2}$ and so pays for making *every* k-subset well conditioned. It is natural to hope this is wasteful, since an ill-conditioned subset has small sign-cones, not absent ones. This subsection shows the hope is unfounded for the Gaussian construction. A k-set in which one vector nearly opposes the average of the others has a cone too small for $2^{\Theta(k)}$ tokens to hit, and $2^{\Theta(k)}$ random directions in dimension $\Theta(k)$ contain such a set with high probability. Two ingredients carry the argument: a bound on the cone of such a set, and a two-point estimate for two subset averages.

Lemma 5.5 (cone of a nearly opposed set; proved in Appendix A)**.** *Let $\kappa \in (0, 1]$ and let $x_1, \ldots, x_m, x_w$ be nonzero vectors of $\mathbb{R}^d$, with normalizations $\hat{x}_i = x_i / \|x_i\|$. Put $B = \{1, \ldots, m\}$, $S_B = \sum_{i \in B} x_i$, and $v_B = -S_B / \|S_B\|$, and suppose that $\lambda_{\min}(G_B) \geq \kappa$, that $\|x_i\| \geq \frac{1}{2}\sqrt{d}$ for $i \in B$, and that $\|S_B\| \leq 2\sqrt{md}$, where G_B is the Gram matrix of $(\hat{x}_i)_{i \in B}$. If the angle θ between $\hat{x}_w$ and v_B satisfies $\theta \leq \frac{\pi}{2}$ and $\sin\theta \leq \sqrt{\kappa/m}/16$, then the set $S = B \cup \{w\}$ satisfies $q_+(G_S) \leq e\,(C_\kappa \sin\theta)^m$, where $C_\kappa = 8e/\sqrt{2\pi\kappa}$.*

The second ingredient controls pairs of subsets whose average directions are correlated.

Lemma 5.6 (two-point cap estimate; proved in Appendix A)**.** *Let u, u' be jointly Gaussian random vectors of $\mathbb{R}^d$, each distributed as $\mathcal{N}(0, I_d)$, with $\mathrm{Cov}(u, u') = \rho I_d$ for some $\rho \in [0, 1)$. Then $\mathbf{Pr}[\angle(u, x) \leq \theta \text{ and } \angle(u', x) \leq \theta] \leq p(\theta)^2\,((1+\rho)/(1-\rho))^{d/2}$ for every fixed unit vector x and every $\theta \in (0, \pi/2]$.*

Theorems 5.5 and 5.6 together give the main theorem of this section.

Theorem 5.7 (restatement of Theorem 1.5; proved in Appendix A)**.** *There are absolute constants γ_0, $c_2 > 0$, and k_0 such that the following holds for every $k \geq k_0$ and every even integer $n \geq 2^{\gamma_0 k}$. Let a threshold head have $\tau = 0$ and n tokens with independent $\mathcal{N}(0, I_d)$ embeddings, where $2C_0 k \leq d \leq c_2 k^2 / \log_2 k$, with C_0 the constant of Theorem 5.1. Then the head fails C_k with probability at least $1 - \varepsilon(k)$, where $\varepsilon(k) \to 0$ as $k \to \infty$ and ε depends only on k and the absolute constants, so that the bound is uniform over admissible n and d. In fact, with probability at least $1 - \varepsilon(k)$ there is a set S of k tokens such that no token is adjacent to every token of S.*

The theorem leaves a window: it asks for $d \geq 2C_0 k$, while Theorem 5.4 rules out only $d < k$, so the range $k \leq d < 2C_0 k$ is reached by neither. Theorem 5.8 closes it.

Corollary 5.8 (proved in Appendix A)**.** *In the setting of Theorem 5.7, and for all sufficiently large k, the Gaussian head fails C_k with probability $1 - o(1)$ for every dimension d with $k \leq d \leq c_2 k^2 / \log_2 k$, and not only for $d \geq 2C_0 k$. Together with Theorem 5.4, which rules out $d < k$ outright, this means that at the minimal token budget the Gaussian head requires $d > c_2 k^2 / \log_2 k$ with high probability.*

Remark 5.9. The hope recorded at the head of this subsection fails for the Gaussian head: its degenerate subsets occur with probability $1 - o(1)$, not as rare events a union bound could dismiss.

A typical Gaussian head therefore cannot reach linear dimension; Theorem 5.10 asks whether some head can.

Conjecture 5.10. There are constants C_{lin} and γ such that for every k there is a d-dimensional threshold head with $\tau = 0$, of dimension $d \le C_{\mathrm{lin}} k$, on $n = 2^{\gamma k}$ tokens, with the k-extension property. For $\gamma \ge \gamma_0$, by Theorems 5.7 and 5.8, the Gaussian head has this property only with probability $o(1)$.

Remark 5.11. Theorem 5.10 implies an extremal sign-rank question. By the proof of Theorem 4.1, a d-dimensional head with C_k on n tokens is a k-existentially closed graph on n vertices, some diagonal completion of whose sign matrix has sign-rank at most $d + 1$. A positive answer would therefore produce a k-e.c. graph on $2^{\Theta(k)}$ vertices at sign-rank $O(k)$. The converse, from such a graph back to a head, does not follow: realizability at $\tau = 0$ asks for a sign-agreeing matrix of the form $X^\top X$. So the sign-rank reading can bear only on the negative side, and only through a lower bound on sign-rank.

The Paley graph P_q, on a prime power $q \equiv 1 \pmod 4$ vertices, is k-e.c. once q is large enough in terms of k (Blass et al., 1981). Its sign matrix is $Q + I$, where Q is the Jacobsthal matrix with $QQ^\top = qI - J$ and $Q\mathbf{1} = 0$. So the eigenvalues of $Q + I$ lie in $\{1, 1 \pm \sqrt{q}\}$, every diagonal completion has norm at most $\sqrt{q} + 3$, and Theorem 2.6 gives every completion sign-rank at least $q/(\sqrt{q}+3) \ge \sqrt{q} - 3$. Since a k-e.c. graph has more than 2^k vertices by Theorem 2.4, that bound is $2^{\Omega(k)}$. At the other end, the Gaussian head of Theorem 5.4 at $n = 2^{Ck}$ and $d = C^2k^2$, with C taken to be an integer, has C_k with positive probability, so by the first paragraph some k-e.c. graph on $2^{\Theta(k)}$ vertices has a completion of sign-rank $O(k^2)$. No completion of a k-e.c. graph has sign-rank below k: on any k columns the witness rows show every sign pattern, and a real matrix of rank r realizing those signs writes them as signs of inner products in $\mathbb{R}^r$, so the k columns are shattered at threshold 0 and $k \le r$ by Theorem 2.5. So the least sign-rank of a completion of a k-e.c. graph on $2^{\Theta(k)}$ vertices lies between k and $O(k^2)$, and Theorem 5.10 asks for $O(k)$. The spectral bound certifies exponential values such as the Paley graph's, but a bound for one graph says nothing by itself about the least value over all k-e.c. graphs.

Appendix A records why we expect a negative answer, if there is one, to come from a covering phenomenon, and what the three constructions we examined do.

5.4 The Directed Case

The upper bound of Theorem 5.4 and the failure of Theorem 5.7 carry over to directed heads with independent queries and keys. The lower bound does not: the complete digraph on $k + 1$ vertices, realized at $d = 1$ by $x_v = y_v = 1$ and $\tau = 0$, has the directed k-extension property.

Proposition 5.12 (restatement of Theorem 1.6; proved in Appendix A)**.** *Let the directed head of Theorem 2.8 have $\tau = 0$ and independent $\mathcal{N}(0, I_d)$ queries and keys. The upper bound of Theorem 5.4 and the failure of Theorem 5.7 both transfer, in the following sense. There is an absolute constant C' such that the directed k-extension property holds with probability at least $1 - 1/n$ once $n \ge 2^{C'k}$ and $d \ge C'k \log n$, and it fails with probability $1 - o(1)$ in the regime of Theorem 5.7.*

Remark 5.13. The directed setting has one invariant the symmetric one lacks, order dimension, the least number of total orders whose intersection is a given partial order, and it gives no obstruction at all. Read a poset as the digraph of its strict order, with $u \to v$ exactly when $u < v$. Already at $d = 4$ and $\tau = 0$ a directed head realizes, for every $n \ge 2$, the standard poset of order dimension n, with elements $a_1, \dots, a_n, b_1, \dots, b_n$ and $a_i < b_j$ exactly when $i \ne j$. Appendix A gives the queries and keys. For tournaments, the flip-class measure of Rajkumar et al. (2022) bounds from above the least dimension of a skew-symmetric head realizing a tournament, but not from below, as Appendix A shows. Section 1.5 returns to the question.

AI use statement

Claude (Anthropic) helped develop the proofs, survey prior work, and draft the manuscript. Codex (OpenAI) assisted with reviewing the mathematical arguments, citations, and prose of the ICLR submission. The authors posed the problem, reviewed the AI-assisted work, and take responsibility for the paper and any inaccuracies.

Under the ICLR 2027 AI Policy for Authors, model development, claim formulation, proof ingredients, and proof writing require disclosure; drafting, literature search, and reference formatting are recommended for disclosure. Data, experiments, implementation, and translation do not arise in this theoretical paper.

Reproducibility statement

The paper is theoretical and reports no experiments. Every result is a numbered statement whose hypotheses are stated in full where it appears, every proof is given in full in Appendix A, and every result imported from the literature is a numbered fact carrying its citation.

References

Micah Adler. A capacity-based rationale for multi-head attention. arXiv:2509.22840, 2025.

Noga Alon, Peter Frankl, and Vojtěch Rödl. Geometrical realization of set systems and probabilistic communication complexity. In *26th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 277–280, 1985.

Noga Alon, Shay Moran, and Amir Yehudayoff. Sign rank versus VC dimension. In *Annual Conference on Learning Theory (COLT)*, volume 49 of *Proceedings of Machine Learning Research*, pp. 47–80, 2016.

Zhi-Dong Bai and Yong-Quan Yin. Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a Wigner matrix. *The Annals of Probability*, 16(4):1729–1741, 1988.

Paul Balister, Béla Bollobás, Karen Gunderson, Imre Leader, and Mark Walters. Random geometric graphs and isometries of normed spaces. *Transactions of the American Mathematical Society*, 370(10):7361–7389, 2018.

Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

Andreas Blass, Geoffrey Exoo, and Frank Harary. Paley graphs satisfy all first-order adjacency axioms. *Journal of Graph Theory*, 5(4):435–439, 1981.

Anthony Bonato. The search for n-e.c. graphs. *Contributions to Discrete Mathematics*, 4(1):40–53, 2009.

Anthony Bonato and Jeannette Janssen. Infinite random geometric graphs. *Annals of Combinatorics*, 15(4):597–617, 2011.

Anthony Bonato, Przemysław Gordinowicz, and Paweł Prałat. Bounds and constructions for n-e.c. tournaments. *Contributions to Discrete Mathematics*, 5(2):52–66, 2010.

Peter J. Cameron. The random graph. In R. L. Graham and J. Nešetřil (eds.), *The Mathematics of Paul Erdős, II*, volume 14 of *Algorithms and Combinatorics*, pp. 333–351. Springer, 1997.

Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

Thomas M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965.

Paul Erdős and Alfréd Rényi. Asymmetric graphs. *Acta Mathematica Academiae Scientiarum Hungaricae*, 14:295–315, 1963.

Charles M. Fiduccia, Edward R. Scheinerman, Ann Trenk, and Jennifer S. Zito. Dot product representations of graphs. *Discrete Mathematics*, 181(1–3):113–138, 1998.

Jürgen Forster. A linear lower bound on the unbounded error probabilistic communication complexity. *Journal of Computer and System Sciences*, 65(4):612–625, 2002.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.

Prashant Gokhale, Piotr Indyk, Yuhao Liu, Sandeep Silwal, Tony Wang, and Haike Xu. Compact geometric representations of hierarchies. In *Annual Conference on Learning Theory (COLT)*, volume 336 of *Proceedings of Machine Learning Research*, pp. 2854–2877, 2026.

Hamed Hatami, Pooya Hatami, William Pires, Ran Tao, and Rosie Zhao. Lower bound methods for sign-rank and their limitations. In *Approximation, Randomization, and Combinatorial Optimization . Algorithms and Techniques (APPROX/RANDOM 2022)*, volume 245 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 22:1–22:24, 2022.

Wilfrid Hodges. *Model Theory*, volume 42 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1993.

A. H. Lachlan. Countable homogeneous tournaments. *Transactions of the American Mathematical Society*, 284(2):431–461, 1984.

Su Hyeong Lee, Qingqi Zhang, and Risi Kondor. Sign rank limitations for inner product graph decoders. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27118–27136, 2024.

Dugald Macpherson. A survey of homogeneous structures. *Discrete Mathematics*, 311(15):1599–1634, 2011.

Ramamohan Paturi and Janos Simon. Probabilistic communication complexity. *Journal of Computer and System Sciences*, 33(1):106–123, 1986.

Richard Rado. Universal graphs and universal functions. *Acta Arithmetica*, 9:331–340, 1964.

Rajmohan Rajaraman, Ravi Sundaram, and Amanuel Tesfaye. The head complexity of Boolean functions in single-layer attention. arXiv:2609.04046, 2026.

Arun Rajkumar, Vishnu Veerathu, and Abdul Bakey Mir. A theory of tournament representations. In *International Conference on Learning Representations (ICLR)*, 2022.

Jan Reiterman, Vojtěch Rödl, and Edita Šiňajová. Embeddings of graphs in Euclidean spaces. *Discrete & Computational Geometry*, 4(4):349–364, 1989.

Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1): 145–147, 1972.

Saharon Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2018.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2508.21038.

R. S. Wenocur and R. M. Dudley. Some special Vapnik–Chervonenkis classes. *Discrete Mathematics*, 33(3):313–318, 1981.

A Deferred Proofs

This appendix collects the proofs of every statement in the body, in the order the statements appear.

Proof of Theorem 2.4. First, $|V| \geq k$: otherwise the pair with $A = V$ and $B = \emptyset$ satisfies $|A \cup B| \leq k$ and admits no witness, since $V \setminus A$ is empty. Fix a set $S \subseteq V$ with $|S| = k$. For each $A \subseteq S$ the pair $(A, S \setminus A)$ has $|A \cup (S \setminus A)| = k$, so there is a witness $z_A \in V \setminus S$ adjacent to every vertex of A and to no vertex of $S \setminus A$; that is, the neighborhood of z_A inside S is exactly A. Distinct subsets A therefore give distinct witnesses, so $|V \setminus S| \geq 2^k$ and $|V| \geq 2^k + k$. In a tournament the same argument applies with the out-neighborhood of z_A inside S in place of the neighborhood, which is exactly A because $b \to z_A$ excludes $z_A \to b$ for every $b \in S \setminus A$. □

Proof of Theorem 2.5. For the lower bound, $e_1, \ldots, e_d$ is shattered: given $S \subseteq \{1, \ldots, d\}$, take w with $w_i = \tau$ for $i \in S$ and $w_i = \tau - 1$ otherwise, so that $\langle e_i, w \rangle \geq \tau$ exactly for $i \in S$.

For the upper bound, suppose distinct $p_1, \ldots, p_{d+1} \in \mathbb{R}^d$ were shattered. Being $d + 1$ vectors in $\mathbb{R}^d$ they are linearly dependent, so $\sum_i a_i p_i = 0$ for some $a \neq 0$; put $P = \{i : a_i > 0\}$ and $N = \{i : a_i < 0\}$, and replace a by $-a$ if necessary so that $P \neq \emptyset$. Write $\alpha = \sum_{i \in P} a_i > 0$ and $\beta = \sum_{i \in N} (-a_i) \geq 0$. Splitting $\sum_i a_i \langle p_i, w \rangle = 0$ gives

$$\sum_{i \in P} a_i \langle p_i, w \rangle = \sum_{i \in N} (-a_i) \langle p_i, w \rangle \qquad \text{for every } w \in \mathbb{R}^d.$$

Suppose first that $N = \emptyset$, so both sides vanish. A w realizing the trace $\{p_i : i \in P\}$ has $\langle p_i, w \rangle \geq \tau$ on P, whence $0 \geq \alpha\tau$; a w' realizing the empty trace has $\langle p_i, w' \rangle < \tau$ on P, whence $0 < \alpha\tau$. Suppose instead that $N \neq \emptyset$, so $\beta > 0$. A w realizing the trace $\{p_i : i \in P\}$ makes the common value c of the two sides satisfy $\alpha\tau \leq c < \beta\tau$, and a w' realizing the trace $\{p_i : i \in N\}$ makes the common value c' satisfy $\beta\tau \leq c' < \alpha\tau$; adding the two gives $\alpha\tau + \beta\tau \leq c + c' < \alpha\tau + \beta\tau$. Either way we have a contradiction, and note that neither branch divides by τ or uses its sign. □

Proof of Theorem 3.1. Write $G = G(\varphi, \tau)$ and $H_w = \{x \in \mathbb{R}^d : \langle x, w \rangle \geq \tau\}$, and suppose toward a contradiction that G has the m-extension property for $m = d + 1$. By Theorem 2.4, V contains at least $2^m + m$ vertices, so it contains m distinct ones; and $m = d + 1 \geq 2$, which is what the next step needs.

We first check that φ is injective. Suppose $\varphi(u) = \varphi(v)$ for distinct $u, v \in V$. Then every $z \in V \setminus \{u, v\}$ satisfies $\langle \varphi(z), \varphi(u) \rangle = \langle \varphi(z), \varphi(v) \rangle$ and so is adjacent to u if and only if it is adjacent to v. The pair $A = \{u\}$, $B = \{v\}$ has $|A \cup B| = 2 \leq m$ and therefore admits a witness $z \in V \setminus \{u, v\}$ adjacent to u and not to v, which is the contradiction.

Pick distinct vertices $v_1, \ldots, v_m$ and put $p_i = \varphi(v_i)$, so that $p_1, \ldots, p_m$ are m distinct points of $\mathbb{R}^d$. For a vertex $z \notin \{v_1, \ldots, v_m\}$ the set of indices i with $z \sim v_i$ is

$$\{i : \langle p_i, \varphi(z) \rangle \geq \tau\} = \{i : p_i \in H_{\varphi(z)}\}.$$

Let $A \subseteq \{v_1, \ldots, v_m\}$ be arbitrary and let B be its complement in that set, so that $|A \cup B| = m$. The m-extension property supplies a witness $z \in V \setminus (A \cup B)$ adjacent to every vertex of A and to no vertex of B, and for $w = \varphi(z)$ this reads

$$H_w \cap \{p_1, \ldots, p_m\} = \{p_i : v_i \in A\}.$$

Ranging over all 2^m subsets A exhibits every subset of $\{p_1, \ldots, p_m\}$ as the trace of a member of $\mathcal{H}_\tau$, so these $m = d + 1$ distinct points are shattered. This contradicts Theorem 2.5.

Finally, R has the k-extension property for every k by Theorem 2.2, in particular for $k = d+1$, so by the above it is not d-realizable for any finite d. No finite-dimensional head is therefore relationally universal. □

Proof of Theorem 3.3. Write $H_w = \{y \in \mathbb{R}^d : \langle w, y \rangle \geq \tau\}$ for the halfspace determined by a query w, where τ is the threshold of the head and x_v, y_v are its query and key at v. Assume the induced digraph is a tournament and suppose it has the directed m-extension property for $m = d + 1 \geq 2$. By Theorem 2.4 there are m distinct vertices.

The keys are distinct. If $y_u = y_v$ for distinct u, v then every $z \notin \{u, v\}$ has $z \to u$ if and only if $z \to v$. The pair $A = \{u\}$, $B = \{v\}$ has $|A \cup B| = 2 \le m$ and so admits a witness z with $z \to u$ and $v \to z$. In a tournament $v \to z$ excludes $z \to v$, so $z \to u$ and not $z \to v$, which is the contradiction.

Fix distinct $v_1, \dots, v_m$ with keys $y_1, \dots, y_m$, which are m distinct points of $\mathbb{R}^d$. A vertex z outside this set satisfies $z \to v_i$ exactly when $y_i \in H_{x_z}$. Given $A \subseteq \{v_1, \dots, v_m\}$ with complement B in that set, the directed m-extension property supplies z with $z \to a$ for every $a \in A$ and $b \to z$ for every $b \in B$. Since the digraph is a tournament, $b \to z$ excludes $z \to b$, so the out-neighborhood of z inside $\{v_1, \dots, v_m\}$ is exactly A and

$$H_{x_z} \cap \{y_1, \dots, y_m\} = \{y_i : v_i \in A\}.$$

Ranging over A shatters the $m = d + 1$ distinct key points, contradicting Theorem 2.5. Since the countable random tournament has the directed k-extension property for every k by Theorem 2.3, it is not realized by any finite-dimensional directed head. □

Proof of Theorem 4.1. Let φ and τ realize G, write $X = [\varphi(v_1) \cdots \varphi(v_n)] \in \mathbb{R}^{d \times n}$ and $J = \mathbf{1}\mathbf{1}^\top$, and for $\varepsilon > 0$ set

$$N(\varepsilon) = X^\top X - (\tau - \varepsilon) J.$$

Every entry of $N(\varepsilon)$ is an affine function of ε with slope 1, so each vanishes at exactly one value of ε, and there are at most n^2 such values in all. Put

$$\varepsilon_0 = \min \{\tau - \langle \varphi(v_i), \varphi(v_j) \rangle : i \ne j \text{ and } v_i \not\sim v_j\} > 0,$$

with $\varepsilon_0 = 1$ if G is complete, and fix $\varepsilon \in (0, \varepsilon_0)$ avoiding those values. Write $N = N(\varepsilon)$. Then no entry of N is zero, and off the diagonal $\operatorname{sign}(N_{ij}) = H_{ij}$: an adjacent pair has $\langle \varphi(v_i), \varphi(v_j) \rangle - \tau \ge 0$ and hence $N_{ij} \ge \varepsilon > 0$, while a nonadjacent pair has $\langle \varphi(v_i), \varphi(v_j) \rangle - \tau \le -\varepsilon_0 < -\varepsilon$ and hence $N_{ij} < 0$.

Let M be the sign matrix with $M_{ij} = H_{ij}$ for $i \ne j$ and $M_{ii} = \operatorname{sign}(N_{ii})$. Then $\operatorname{sign}(N) = M$ entrywise, and $X^\top X$ has rank at most d while J has rank 1, so

$$\operatorname{sr}(M) \le \operatorname{rank}(N) \le d + 1.$$

The matrix $M - H$ is diagonal with entries in $\{-2, 0, 2\}$, so $\|M - H\| \le 2$ and $\|M\| \le \|H\| + 2$. By Theorem 2.6, $\operatorname{sr}(M) \ge n / \|M\| \ge n/(\|H\| + 2)$, which gives the bound.

For the two consequences, take H to be the sign matrix in question. A symmetric Hadamard matrix H of order n, for instance the Sylvester matrix $H_2^{\otimes k}$ with $H_2 = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$, satisfies $HH^\top = nI$ and hence $\|H\| = \sqrt{n}$; the Hadamard graph is the graph it defines off the diagonal, and

$$d \ge \frac{n}{\sqrt{n} + 2} - 1 = \sqrt{n} - \frac{2\sqrt{n}}{\sqrt{n} + 2} - 1 \ge \sqrt{n} - 3.$$

For the second, let H be symmetric with independent uniform signs above the diagonal and $H_{ii} = +1$, so that the graph it defines is a uniform random graph. Write $H = W + I$, where W has zero diagonal and the off-diagonal entries of H, and realize the matrices for all n as the leading blocks of one infinite symmetric array. The Bai–Yin theorem (1988) says that the largest eigenvalue of W is $(2 + o(1))\sqrt{n}$ almost surely. Its hypotheses hold for W: off the diagonal the entries have zero mean, unit variance and a finite fourth moment, and on the diagonal the positive part has a finite second moment. The same holds for $-W$, which has the same law, so $\|W\| = (2 + o(1))\sqrt{n}$ and $\|H\| + 2 \le \|W\| + 3 = (2 + o(1))\sqrt{n}$ almost surely. Hence $d + 1 \ge n/(\|H\| + 2) \ge (\frac{1}{2} - o(1))\sqrt{n}$, and $d \ge (\frac{1}{2} - o(1))\sqrt{n}$ since $1 = o(\sqrt{n})$. □

The proof of Theorem 5.7 below uses one elementary bound on the measure of a spherical cap.

Lemma A.1. *For every $d \ge 2$ and $\theta \in (0, \pi/2]$,*

$$p(\theta) \ge \frac{1}{\pi} \left(\frac{\sin \theta}{2} \right)^{d-1}.$$

Proof of Theorem A.1. In spherical coordinates $p(\theta) = \int_0^\theta \sin^{d-2}\phi \mathrm{d}\phi / \int_0^\pi \sin^{d-2}\phi \mathrm{d}\phi$, and the denominator is at most π. Since $\sin$ is increasing on $[0, \pi/2]$, the numerator is at least $\int_{\theta/2}^\theta \sin^{d-2}\phi \mathrm{d}\phi \geq \frac{\theta}{2}\sin^{d-2}\frac{\theta}{2}$, and $\sin\frac{\theta}{2} \geq \frac{\sin\theta}{2}$ together with $\theta \geq \sin\theta$ gives $\frac{\theta}{2}\sin^{d-2}\frac{\theta}{2} \geq \frac{\sin\theta}{2}\left(\frac{\sin\theta}{2}\right)^{d-2}$. □

Proof of Theorem 5.2. Let $W = \operatorname{span}\{\hat{x}_i : i \in S\}$ and let P be the orthogonal projection onto W. Each constraint reads $\langle \hat{x}_i, g\rangle = \langle \hat{x}_i, Pg\rangle$, and Pg is a standard Gaussian on W. Choosing an orthonormal basis of W replaces the $\hat{x}_i$ by vectors of $\mathbb{R}^{\dim W}$ with the same pairwise inner products, so $q_s(G)$ is a function of G alone. For the second claim, replace $\hat{x}_i$ by $s_i\hat{x}_i$: the event $\operatorname{sign}\langle \hat{x}_i, g\rangle = s_i$ becomes $\langle s_i\hat{x}_i, g\rangle \geq 0$ up to the null event $\langle \hat{x}_i, g\rangle = 0$, and the Gram matrix of the new vectors is D_sGD_s. Conjugation by the orthogonal matrix D_s preserves the spectrum and the diagonal. □

Proof of Theorem 5.3. By Theorem 5.2 we may replace G by D_sGD_s, which has the same unit diagonal and the same smallest eigenvalue, and assume $s \equiv +1$. Let $u = (\langle \hat{x}_i, g\rangle)_i$, so $u \sim \mathcal{N}(0, G)$, and lower bound $\mathbf{Pr}[u \geq 0]$. Since $\lambda_{\min}(G) \geq \frac{1}{2}$ the matrix $H := G - \frac{1}{2}I$ is positive semidefinite, so we may write $u = \frac{1}{\sqrt{2}}z + h$ with $z \sim \mathcal{N}(0, I_m)$ and $h \sim \mathcal{N}(0, H)$ independent. Conditioning on h, the coordinates of z are independent, so

$$\mathbf{Pr}[u \geq 0] = \mathbb{E}_h\left[\prod_{i=1}^m \mathbf{Pr}_z\left[z_i \geq -\sqrt{2}h_i | h\right]\right] = \mathbb{E}_h\left[\exp\sum_{i=1}^m \log\Phi\left(\sqrt{2}h_i\right)\right],$$

where Φ is the standard normal distribution function. By Jensen's inequality applied to the convex function $\exp$,

$$\mathbf{Pr}[u \geq 0] \geq \exp\mathbb{E}_h\left[\sum_{i=1}^m \log\Phi\left(\sqrt{2}h_i\right)\right].$$

Each h_i is centered Gaussian with variance $H_{ii} = G_{ii} - \frac{1}{2} = \frac{1}{2}$, so $\sqrt{2}h_i$ is a standard Gaussian Z. Since Φ is continuous and strictly increasing, $\Phi(Z)$ is uniform on $(0, 1)$, whence

$$\mathbb{E}[\log\Phi(Z)] = \int_0^1 \log t \mathrm{d}t = -1.$$

Therefore $\mathbf{Pr}[u \geq 0] \geq e^{-m}$, and $e^{-1} > \frac{1}{4}$ gives $e^{-m} \geq 2^{-2m}$. □

Proof of Theorem 5.4, lower bound. The bound on n is Theorem 2.4. For the dimension, C_k is the k-extension property, and by Theorem 3.1 a d-dimensional head fails the $(d+1)$-extension property at every threshold, hence fails C_j for every $j \geq d+1$. Hence C_k forces $k \leq d$. □

Proof of Theorem 5.4, upper bound. Fix disjoint A, B with $m := |A \cup B| \leq k$ and write X_{AB} for the embeddings of the active tokens. A token $z \notin A \cup B$ witnesses (A, B) exactly when its embedding realizes the sign pattern s that is $+1$ on A and -1 on B; by Theorem 5.2 this happens with probability $q_s(G_{AB})$, where G_{AB} is the normalized Gram matrix of X_{AB}. The $n-m$ candidate witnesses are independent of X_{AB} and of one another, so

$$\mathbf{Pr}[\text{no witness for } (A, B)] = \mathbb{E}_{X_{AB}}\left[(1 - q_s(G_{AB}))^{n-m}\right].$$

By Theorem 5.3, $q_s(G_{AB}) \geq e^{-m} \geq e^{-k}$ whenever $\lambda_{\min}(G_{AB}) \geq \frac{1}{2}$, which by the second half of Theorem 5.1 fails with probability at most $e^{-c_0 d}$ provided $d \geq C_1 k$. Splitting on that event,

$$\mathbf{Pr}[\text{no witness for } (A, B)] \leq \left(1 - e^{-k}\right)^{n-k} + e^{-c_0 d} \leq e^{-e^{-k}(n-k)} + e^{-c_0 d}.$$

The number of ordered pairs (A, B) of disjoint sets with $|A \cup B| \leq k$ is $\sum_{m \leq k}\binom{n}{m}2^m \leq \sum_{m \leq k}(2n)^m \leq 2(2n)^k \leq (3n)^k$, the last step using $k \geq 2$. A union bound gives

$$\mathbf{Pr}[C_k \text{ fails}] \leq (3n)^k\left(e^{-e^{-k}(n-k)} + e^{-c_0 d}\right),$$

which is at most $1/n$ as soon as each of $e^{-k}(n-k)$ and $c_0 d$ is at least $(k+1)\log(3n)$, since then each term is at most $\frac{1}{2n}$. The first holds for $n \geq 2^{Ck}$ with C large, because the left side grows linearly in n while the right side grows logarithmically, and at $n = 2^{Ck}$ it reads $e^{-k}(2^{Ck} - k) \geq (k+1)\log(3 \cdot 2^{Ck}) = O(Ck^2)$. The second holds for $d \geq Ck \log n$ with $C \geq 4/c_0$, since $(k+1)\log(3n) \leq 4k \log n$ for $n \geq 3$ and $k \geq 2$. Finally $d \geq Ck \log n \geq C_1 k$, as Theorem 5.1 requires, once $C \geq C_1$. □

Proof of Theorem 5.5. Let $W = \operatorname{span}\{\hat{x}_i : i \in B\}$ and write $\hat{x}_w = y + \zeta$ with y the orthogonal projection of $\hat{x}_w$ onto W and $\zeta \perp W$. Since $v_B \in W$ we have $\langle y, v_B \rangle = \langle \hat{x}_w, v_B \rangle = \cos\theta$, so $\|y\| \geq \cos\theta$ by Cauchy–Schwarz, and since $\cos\theta \geq 0$ for $\theta \leq \frac{\pi}{2}$, squaring and $\|y\|^2 + \|\zeta\|^2 = 1$ give $\|\zeta\| \leq \sin\theta$. Let $g \sim \mathcal{N}(0, I_d)$ and put $u_i = \langle \hat{x}_i, g \rangle$ for $i \in B$ and $u_w = \langle \hat{x}_w, g \rangle$. Then $u_B \sim \mathcal{N}(0, G_B)$, and, assuming first that $\zeta \neq 0$ (the degenerate case $\zeta = 0$ is settled at the end), since $\zeta \perp W$ the variable $g_0 := \langle \zeta, g \rangle / \|\zeta\|$ is a standard normal independent of u_B, so

$$u_w = \langle y, g \rangle + \|\zeta\| g_0 = \langle \alpha, u_B \rangle + \|\zeta\| g_0, \qquad \alpha = G_B^{-1} G_{B,w},$$

where $G_{B,w} = (\langle \hat{x}_i, \hat{x}_w \rangle)_{i \in B} = (\langle \hat{x}_i, y \rangle)_{i \in B}$ and α is the coordinate vector of y in the basis $(\hat{x}_i)_{i \in B}$ of W.

We show every coordinate of $-\alpha$ is at least $1/(8\sqrt{m})$. Decompose $y = (\cos\theta)\, v_B + y_\perp$ with $y_\perp \in W$ orthogonal to v_B, so $\|y_\perp\|^2 = \|y\|^2 - \cos^2\theta \leq \sin^2\theta$. The coordinate vector of v_B in the basis $(\hat{x}_i)$ is $-\left(\|x_i\| / \|S_B\|\right)_{i \in B}$, directly from $v_B = -\sum_i \|x_i\| \hat{x}_i / \|S_B\|$, and each of its coordinates is at most $-\frac{1}{2}\sqrt{d} / \left(2\sqrt{md}\right) = -1/(4\sqrt{m})$. Let ξ be the coordinate vector of $y_\perp$ in the basis $(\hat{x}_i)_{i \in B}$, so that $y_\perp = \sum_{i \in B} \xi_i \hat{x}_i$ and $\|y_\perp\|^2 = \xi^\top G_B \xi \geq \kappa \|\xi\|^2$; hence $\|\xi\| \leq \|y_\perp\| / \sqrt{\kappa} \leq \sin\theta / \sqrt{\kappa}$, and no coordinate of ξ exceeds this in absolute value. Hence every coordinate of $-\alpha$ is at least

$$\frac{\cos\theta}{4\sqrt{m}} - \frac{\sin\theta}{\sqrt{\kappa}} \geq \frac{1}{8\sqrt{m}},$$

where we used the hypothesis $\sin\theta \leq \sqrt{\kappa/m}/16$ twice: it gives $\sin\theta/\sqrt{\kappa} \leq 1/(16\sqrt{m})$, and, since $\kappa \leq 1 \leq m$, it gives $\sin\theta \leq \frac{1}{16}$, which together with $\theta \leq \frac{\pi}{2}$ gives $\cos\theta = \sqrt{1 - \sin^2\theta} \geq \frac{3}{4}$.

Consequently, on the event $\{u_i > 0 \text{ for } i \in B\}$ we have $\langle -\alpha, u_B \rangle \geq T/(8\sqrt{m}) > 0$ with $T = \sum_{i \in B} u_i$, so $u_w > 0$ forces $\|\zeta\| g_0 > T/(8\sqrt{m}) > 0$, whence $g_0 > 0$ and, since $\|\zeta\| \leq \sin\theta$, $T < 8\sqrt{m}\, \|\zeta\| g_0 \leq 8\sqrt{m}\, (\sin\theta)\, g_0$. The density of $\mathcal{N}(0, G_B)$ is at most $(2\pi)^{-m/2} (\det G_B)^{-1/2} \leq (2\pi\kappa)^{-m/2}$, and $\{u > 0, \sum_i u_i < t\}$ is a simplex of volume $t^m/m!$, so, since g_0 is independent of u_B, conditioning on g_0 gives

$$q_+(G_S) \leq \frac{(8\sqrt{m}\sin\theta)^m}{(2\pi\kappa)^{m/2} m!} \mathbb{E}[|g_0|^m] \leq \frac{8^m (\sin\theta)^m (m(m+1))^{m/2}}{(2\pi\kappa)^{m/2} m!} \leq \frac{(8\sin\theta)^m (m+1)^m}{(2\pi\kappa)^{m/2} m!}.$$

The second inequality uses $\mathbb{E}\left[|g_0|^m\right] \leq (m+1)^{m/2}$: for even m this is $\mathbb{E}[g_0^m] = (m-1)!!$, a product of $m/2$ factors each at most m, and for odd m Lyapunov's inequality gives $\mathbb{E}\left[|g_0|^m\right] \leq \left(\mathbb{E}\left[|g_0|^{m+1}\right]\right)^{m/(m+1)}$, where $\mathbb{E}\left[|g_0|^{m+1}\right] = m!!$ is a product of $(m+1)/2$ factors each at most $m+1$. And $(m+1)^m/m! = (m+1)^{m+1}/(m+1)! \leq e^{m+1}$, since $e^x = \sum_{j \geq 0} x^j/j! \geq x^n/n!$ at $x = n = m+1$, giving $q_+(G_S) \leq e\left(8e\sin\theta/\sqrt{2\pi\kappa}\right)^m$. Finally, if $\zeta = 0$ then $u_w = \langle \alpha, u_B \rangle \leq -T/(8\sqrt{m}) < 0$ on the event $\{u_i > 0 \text{ for } i \in B\}$, so $q_+(G_S) = 0$ and the bound holds because its right-hand side is nonnegative. □

The existence of the density below is standard. The pair (u, u') has a nondegenerate Lebesgue density on $\mathbb{R}^d \times \mathbb{R}^d$, so polar factorization gives the pair of directions a density against the product of the uniform measures. The joint rotation invariance of the law of (u, u') makes that density a function of $\langle \omega, \omega' \rangle$ alone. What the lemma supplies is the value at coincidence. The standard two-point alternative, Gaussian hypercontractivity, does not reach it; the paragraph after the proof of Theorem 5.7 says why.

Proof of Theorem 5.6. Write $\omega = u/\|u\|$ and $\omega' = u'/\|u'\|$, let σ be the uniform probability measure on S^{d-1}, and let $S_{d-1} = 2\pi^{d/2}/\Gamma(d/2)$ be the surface area of S^{d-1}. The joint density of (u, u') is

$$\alpha(u, u') = (2\pi)^{-d} \left(1 - \rho^2\right)^{-d/2} \exp\left(-\frac{\|u\|^2 - 2\rho \langle u, u' \rangle + \|u'\|^2}{2\left(1 - \rho^2\right)}\right),$$

and passing to polar coordinates $u = r\omega$ and $u' = r'\omega'$ shows that the joint density of (ω, ω') with respect to $\sigma \times \sigma$ is

$$f(\omega, \omega') = S_{d-1}^2 \int_0^\infty \int_0^\infty r^{d-1} r'^{d-1} \alpha(r\omega, r'\omega') \mathrm{d}r \mathrm{d}r',$$

which depends on (ω, ω') only through $t = \langle \omega, \omega' \rangle$, through the term $\exp\left(\rho r r' t/(1 - \rho^2)\right)$. That term is nondecreasing in t because $\rho \geq 0$ and $r, r' > 0$, so f is maximized at $t = 1$. Writing K for the cap of angular radius θ about x, whose measure is $\sigma(K) = p(\theta)$,

$$\mathbf{Pr}[\omega \in K, \omega' \in K] = \int_K \int_K f \mathrm{d}\sigma \mathrm{d}\sigma \leq f(1) p(\theta)^2,$$

where $f(1)$ denotes the common value of f at coincident arguments. It remains to show $f(1) \leq ((1 + \rho)/(1 - \rho))^{d/2}$.

Put $A = S_{d-1}^2 (2\pi)^{-d}$. At $\rho = 0$ the two directions are independent and uniform, so $f \equiv 1$ and $A \left[\int_0^\infty r^{d-1} e^{-r^2/2} \mathrm{d}r\right]^2 = 1$, that is $A = 1/\left(2^{d-2}\Gamma(d/2)^2\right)$. For general ρ, substitute $r = R\cos\alpha$ and $r' = R\sin\alpha$ with $\alpha \in (0, \pi/2)$, so that $r^2 - 2\rho r r' + r'^2 = R^2\left(1 - \rho \sin 2\alpha\right)$ and $\mathrm{d}r\mathrm{d}r' = R\mathrm{d}R\mathrm{d}\alpha$. Using $\int_0^\infty R^{2d-1} e^{-aR^2} \mathrm{d}R = \Gamma(d)/\left(2a^d\right)$ with $a = \left(1 - \rho \sin 2\alpha\right)/\left(2(1 - \rho^2)\right)$, and $\cos\alpha \sin\alpha = \frac{1}{2}\sin 2\alpha$, and then $\psi = 2\alpha$ followed by the symmetry $\psi \mapsto \pi - \psi$, which folds the resulting integral over $(0, \pi)$ into twice the integral over $(0, \frac{\pi}{2})$,

$$f(1) = A\left(1 - \rho^2\right)^{d/2} \Gamma(d) \int_0^{\pi/2} \frac{\sin^{d-1}\psi}{\left(1 - \rho \sin\psi\right)^d} \mathrm{d}\psi.$$

Bound $1 - \rho\sin\psi \geq 1 - \rho$ and use the Wallis integral

$$\int_0^{\pi/2} \sin^{d-1}\psi \mathrm{d}\psi = \frac{\sqrt{\pi}\Gamma(d/2)}{2\Gamma\left(\frac{d+1}{2}\right)},$$

which give

$$f(1) \leq \frac{\left(1 - \rho^2\right)^{d/2}}{(1 - \rho)^d} \cdot \frac{\Gamma(d)\sqrt{\pi}\Gamma(d/2)}{2^{d-1}\Gamma(d/2)^2\Gamma\left(\frac{d+1}{2}\right)}.$$

By the Legendre duplication formula $\Gamma(d) = 2^{d-1}\Gamma(d/2)\Gamma\left(\frac{d+1}{2}\right)/\sqrt{\pi}$, the second factor equals 1, and the first is $((1 + \rho)/(1 - \rho))^{d/2}$. □

Proof of Theorem 5.7. Put $\gamma = (\log_2 n)/k \geq \gamma_0$. Each condition of the form "for k large" below is weakest at $\gamma = \gamma_0$, so one absolute k_0 covers them all. Write $m = k - 1$ and split the tokens into two halves, a *base pool* P_1 and an *apex pool* P_2, each of size $N = n/2$. Let $\mathcal{B}$ be the family of all m-subsets of P_1, so $|\mathcal{B}| = \binom{N}{m}$. For $B \in \mathcal{B}$ put $S_B = \sum_{i \in B} x_i$ and $v_B = -S_B/\|S_B\|$, which is defined almost surely. Call B *good* if it satisfies the three conclusions of Theorem 5.1; since $d \geq C_0 m$, a fixed B is good with probability at least $1 - e^{-c_0 d} = 1 - o(1)$.

The counting variable. For $\theta \in (0, \pi/2)$ let

$$Z = \#\left\{(B, w) \in \mathcal{B} \times P_2 : B \text{ good and } \angle(x_w, v_B) \leq \theta\right\}.$$

For a fixed pair, x_w is independent of $(x_i)_{i \in B}$ and its direction is uniform, so conditionally on v_B the angle condition has probability exactly $p(\theta)$; hence

$$\mathbb{E}[Z] = (1 - o(1)) \binom{N}{m} N p(\theta).$$

Choose θ so that $\left(\frac{1}{2}\sin\theta\right)^{d-1} = \frac{\pi k}{\binom{N}{m}N}$, which is possible for k large since the right side is less than $2^{-(d-1)}$. By Theorem A.1, $p(\theta) \geq k/\left(\binom{N}{m}N\right)$, so $\mathbb{E}[Z] \geq (1-o(1))\,k \to \infty$. Moreover, taking logarithms in the defining equation for θ gives

$$\log_2 \frac{1}{\sin\theta} \geq \frac{1}{d-1}\log_2 \frac{\binom{N}{m}N}{\pi k} - 1 \geq \frac{\gamma k^2}{2d} - 2, \tag{1}$$

where the last step uses $\binom{N}{m} \geq (N/m)^m$ together with $N = 2^{\gamma k - 1}$, so that

$$\log_2 \binom{N}{m} \geq (k-1)\,(\gamma k - 1 - \log_2 k) \geq \tfrac{3}{4}\gamma k^2$$

for $\gamma \geq \gamma_0$ and k large.

The second moment. Consider ordered pairs $((B,w),(B',w'))$ contributing to $\mathbb{E}[Z^2]$. Because bases lie in P_1 and apexes in P_2, an apex never belongs to a base, which is what makes the following case analysis close.

- $B = B'$ and $w = w'$: total contribution $\mathbb{E}[Z]$.
- $B = B'$ with $w \neq w'$: conditionally on $(x_i)_{i\in B}$ the two angle events are independent, each of probability $p(\theta)$, so each such pair contributes at most $p(\theta)^2$. There are at most $\binom{N}{m}N^2$ of them, and the total is at most $\mathbb{E}[Z]^2/\left((1-o(1))^2\binom{N}{m}\right) = o\big(\mathbb{E}[Z]^2\big)$.
- $B \neq B'$ with $w \neq w'$: conditionally on all of P_1, the two angle events depend on the independent vectors x_w and $x_{w'}$ and have conditional probabilities $p(\theta)$ each, so each pair contributes at most $p(\theta)^2$; the total is at most $\binom{N}{m}^2 N^2 p(\theta)^2 = (1+o(1))\,\mathbb{E}[Z]^2$.
- $B \neq B'$ with $w = w'$: condition on x_w. The pair $(-S_B, -S_{B'})/\sqrt{m}$ is jointly Gaussian, each marginal $\mathcal{N}(0, I_d)$, with covariance ρI_d for $\rho = a/m$ and $a = |B \cap B'| < m$. By Theorem 5.6 the two angle events hold together with probability at most $p(\theta)^2 F(a)$, where $F(a) = \left(\frac{m+a}{m-a}\right)^{d/2}$.

Only the last case needs work. The number of ordered pairs (B,B') in $\mathcal{B}^2$ with $|B\cap B'| = a$ is $\binom{N}{m}\binom{m}{a}\binom{N-m}{m-a}$, and

$$\begin{aligned}\beta_a := \frac{\binom{m}{a}\binom{N-m}{m-a}}{\binom{N}{m}} &= \frac{1}{a!}\left(\frac{m!}{(m-a)!}\right)^2 \frac{\prod_{j=0}^{m-a-1}(N-m-j)}{\prod_{j=0}^{m-1}(N-j)} \\ &\leq \frac{1}{a!}\left(\frac{m!}{(m-a)!}\right)^2 \frac{1}{\prod_{i=0}^{a-1}(N-i)} \leq \frac{1}{a!}\left(\frac{m^2}{N}\right)^a.\end{aligned}$$

Here the first step is the definition of the binomial coefficients. The second pairs the numerator factor $N-m-j$ with the denominator factor $N-a-j$ for $0 \leq j \leq m-a-1$, each such ratio being at most 1 because $a \leq m$, and leaves the factors $N, N-1, \ldots, N-a+1$ in the denominator. The third writes $m!/(m-a)! = m^a \prod_{j=0}^{a-1}\left(1-\frac{j}{m}\right)$ and $\prod_{i=0}^{a-1}(N-i) = N^a \prod_{j=0}^{a-1}\left(1-\frac{j}{N}\right)$ and uses $\left(1-\frac{j}{m}\right)^2 \leq 1-\frac{j}{m} \leq 1-\frac{j}{N}$, which holds because $0 \leq j < m \leq N$. Therefore the contribution of the last case, divided by $\mathbb{E}[Z]^2$, is at most

$$\frac{(1+o(1))}{N}\sum_{a=0}^{m-1}\beta_a F(a) \leq \frac{2}{N}\left(1 + \sum_{a=1}^{m-1}\frac{1}{a!}\left(\frac{m^2}{N}\right)^a F(a)\right).$$

We claim each summand is at most $2^{-a\gamma k/2}$. Taking logarithms, this asks

$$2a\log_2 m - a\,(\gamma k - 1) + \frac{d}{2}\log_2\frac{m+a}{m-a} \leq -\frac{a\gamma k}{2},$$

and since $2a\log_2 m \le a\gamma k/4$ for $\gamma \ge \gamma_0$ and k large, it suffices that $\frac{d}{2}\log_2\frac{m+a}{m-a} \le \frac{a\gamma k}{4} - a$, and hence, using $\gamma k \ge 8$ for k large, that $\frac{d}{2}\log_2\frac{m+a}{m-a} \le \frac{a\gamma k}{8}$ for every $1 \le a \le m-1$. For $a \le m/2$ use $\log_2\frac{m+a}{m-a} = \log_2\left(1+\frac{2a}{m-a}\right) \le \frac{2a}{(m-a)\ln 2} \le \frac{6a}{m}$, which reduces the requirement to $d \le \gamma km/24$, implied by $d \le c_2 k^2/\log_2 k$. For $a > m/2$ use $\log_2\frac{m+a}{m-a} \le \log_2(2m)$ and $a > m/2$, which reduces it to $d \le \gamma km/\left(8\log_2(2m)\right)$, again implied by $d \le c_2 k^2/\log_2 k$ once c_2 is small enough. Hence the sum is at most $2^{-\gamma k/2+1}$ and the whole case contributes $O(1/N)\cdot\mathbb{E}[Z]^2 = o\big(\mathbb{E}[Z]^2\big)$.

Collecting the four cases, $\mathrm{Var}(Z) \le \mathbb{E}[Z] + o\big(\mathbb{E}[Z]^2\big)$, and since $\mathbb{E}[Z] \to \infty$, Chebyshev's inequality gives $Z \ge \frac{1}{2}\mathbb{E}[Z]$ with probability $1 - o(1)$.

The constants. Every constraint the argument places on c_2 has the form $c_2 \le \eta(\gamma)$ with η nondecreasing in γ, so a value admissible at γ_0 is admissible at every $\gamma \ge \gamma_0$ and c_2 may be taken absolute. There are two such constraints, one from the case $a > m/2$ just closed and one from the witnessing bound below: $d \le \gamma km/\left(8\log_2(2m)\right)$, which for $m = k-1$ reduces to $c_2 \le (1-o(1))\,\gamma/8$, and $\gamma\log_2 k/(2c_2) \ge \log_2(16k/\kappa) + 2$, which reduces to $c_2 \le (1-o(1))\,\gamma/2$. We fix $c_2 = \gamma_0/16$. Nothing constrains γ_0 beyond $\gamma_0 > 0$, since every condition below of the form "for k large" holds for any fixed $\gamma_0 > 0$ once k is large enough in terms of it; we take $\gamma_0 = 1$, so that $c_2 = 1/16$. Every remaining use of γ is a condition of the form "for k large": the existence of θ, which needs $\binom{N}{m}N > \pi k 2^{d-1}$; the bound $N \ge \pi k$ inside Eq. (1); $\log_2\binom{N}{m} \ge \frac{3}{4}\gamma k^2$; $2a\log_2 m \le a\gamma k/4$; $\gamma k \ge 8$; $d \le \gamma km/24$; the two $(1-o(1))$ factors in the constraints on c_2; and $\mathbb{E}[Z'] = o(\mathbb{E}[Z])$. Each is weaker for larger γ, so the single k_0 chosen at γ_0 suffices. Moreover every error term in the argument is bounded by a function of k and the absolute constants alone, uniformly over $\gamma \ge \gamma_0$, because each is smaller for larger γ: the failure probability $e^{-c_0 d} \le e^{-2c_0 C_0 k}$ of a fixed B being good; the ratio $1/\binom{N}{m} \le 2^{-3\gamma_0 k^2/4}$ in the case $B = B'$, $w \ne w'$; the factor $\frac{2}{N}\left(1 + 2^{1-\gamma k/2}\right) \le 2^{3-\gamma_0 k}$ in the case $B \ne B'$, $w = w'$; the Chebyshev bound $4/\mathbb{E}[Z]$; and the ratio $\mathbb{E}[Z']/\mathbb{E}[Z]$, whose logarithm is bounded in the next paragraph by an expression decreasing in γ. This is what makes the failure probability in the statement a function $\varepsilon(k)$ of k alone.

Degenerate pairs are not witnessed. Let Z' count the pairs contributing to Z for which some token $z \notin B \cup \{w\}$ satisfies $\langle x_z, x_i\rangle \ge 0$ for every $i \in B \cup \{w\}$. Fix a pair (B, w) and condition on $(x_i)_{i\in B\cup\{w\}}$. The remaining $n-k$ tokens are independent of that conditioning, and each lies in the offending cone with probability $q_+\big(G_{B\cup\{w\}}\big)$, so by a union bound the conditional probability that some token witnesses the pattern is at most $nq_+\big(G_{B\cup\{w\}}\big)$. On the event that B is good and $\angle(x_w, v_B) \le \theta$, Theorem 5.5 applies to B and w with the actual angle $\angle(x_w, v_B)$. That angle is at most $\theta < \frac{\pi}{2}$, so its sine is at most $\sin\theta$. The lemma's remaining hypothesis is $\sin\theta \le \sqrt{\kappa/m}/16$, with κ the constant of Theorem 5.1. That constant is at most 1, as the lemma requires, since a Gram matrix of unit vectors has unit diagonal and hence $\lambda_{\min} \le 1$. We verify the stronger $\sin\theta \le \kappa/(16m)$, which holds by Eq. (1) because $\gamma k^2/(2d) \ge \gamma\log_2 k/(2c_2)$ exceeds $\log_2(16k/\kappa)+2$ for c_2 small and k large. Therefore

$$\mathbb{E}[Z'] \le \mathbb{E}[Z]\cdot ne\left(C_\kappa\sin\theta\right)^m.$$

By Eq. (1), $\log_2\left(ne\left(C_\kappa\sin\theta\right)^m\right) \le \gamma k + 2 + m\log_2 C_\kappa - m\left(\frac{\gamma k^2}{2d} - 2\right)$, and $\frac{\gamma k^2}{2d} \ge \frac{\gamma\log_2 k}{2c_2}$, so the right-hand side is at most $\gamma\left(k - \frac{m\log_2 k}{2c_2}\right) + 2 + m\left(\log_2 C_\kappa + 2\right)$. With c_2 absolute the bracket is negative and of order $-k\log_2 k$ for k large, so this tends to $-\infty$, and does so faster for larger γ, whence $\mathbb{E}[Z'] = o(\mathbb{E}[Z])$. By Markov's inequality $Z' < \frac{1}{2}\mathbb{E}[Z]$ with probability $1-o(1)$.

Combining, with probability $1-o(1)$ we have $Z \ge \frac{1}{2}\mathbb{E}[Z] > Z'$, so some pair (B, w) counted by Z is not counted by Z'. For that pair, $S = B\cup\{w\}$ is a set of k tokens with no token adjacent to all of them, which is a failure of C_k for the pair $(A, \emptyset)$ with $A = S$. □

Gaussian hypercontractivity, the standard two-point alternative to Theorem 5.6, would bound the probability in Theorem 5.6 by $p(\theta)^{2/(1+\rho)}$. The angle θ chosen above satisfies $\log_2(1/p(\theta)) = \Theta(\gamma k^2)$, as the rest of this paragraph shows. So for $\rho \le \frac{1}{2}$ the hypercontractive bound exceeds $p(\theta)^2$ by a factor $2^{\Omega(\rho\gamma k^2)}$, where Theorem 5.6 loses only $\left((1+\rho)/(1-\rho)\right)^{d/2} \le e^{2\rho d} = 2^{o(\rho k^2)}$. The upper bound $\log_2(1/p(\theta)) \le k\log_2 N < \gamma k^2$ follows from $p(\theta) \ge k/\left(\binom{N}{m}N\right)$ and $\binom{N}{m}N \le N^k$.

The lower bound comes from the formula in the proof of Theorem A.1: bounding $\sin^{d-2}\phi$ by $\sin^{d-2}\phi\cos\phi/\cos\theta$ in the numerator, and the denominator below by $2\int_0^{\pi/2}\sin^{d-2}\phi\cos\phi\mathrm{d}\phi = 2/(d-1)$, gives $p(\theta) \le \sin^{d-1}\theta/(2\cos\theta) \le \sin^{d-1}\theta$ once $\cos\theta \ge \frac{1}{2}$, which holds since Eq. (1) gives $\log_2(1/\sin\theta) \ge \gamma\log_2 k/(2c_2) - 2 \ge 1$ for k large. The defining equation for θ then gives $\log_2(1/p(\theta)) \ge \log_2\left(\binom{N}{m}N/(\pi k)\right) - (d-1) \ge \frac{3}{4}\gamma k^2 - d - \log_2(\pi k)$, which is $\Omega(\gamma k^2)$ since $d = o(k^2)$.

Proof of Theorem 5.8. For $d \ge 2C_0k$ this is Theorem 5.7. Suppose $k \le d < 2C_0k$ and set $j = \lfloor d/(2C_0)\rfloor$, so that $2C_0j \le d$ and $j \le k$. The head has $n = 2^{\gamma' j}$ tokens for $\gamma' = (\log_2 n)/j \ge \gamma_0 k/j \ge \gamma_0$. Since n is even, Theorem 5.7 applies at budget j provided $j \ge k_0$ and $d \le c_2 j^2/\log_2 j$. The first condition holds for k large: once $d \ge 4C_0$, we have $j \ge d/(4C_0) \ge k/(4C_0) \to \infty$. It remains to check the dimension condition $d \le c_2 j^2/\log_2 j$, and since $j \ge d/(4C_0)$ it suffices that

$$d \le \frac{c_2}{16C_0^2}\cdot\frac{d^2}{\log_2 d}, \qquad \text{that is} \qquad d \ge \frac{16C_0^2}{c_2}\log_2 d,$$

which holds for all d beyond a threshold depending only on c_2 and C_0, and $d \ge k$ exceeds it once k is large. Theorem 5.7 at budget j and token exponent γ' therefore gives a failure of C_j with probability at least $1-\varepsilon(j)$, where ε is the function of that theorem. That function does not depend on γ', which is what the uniformity in the statement is for, since γ' varies with k and d here. Moreover $\varepsilon(j) \to 0$, because $j \ge k/(4C_0) \to \infty$. A head with C_k has C_j for every $j \le k$, since a pair (A,B) with $|A\cup B| \le j$ also has $|A\cup B| \le k$, so the failure of C_j is a failure of C_k. □

Proof of Theorem 5.12. For the upper bound, fix disjoint A, B with $|A\cup B| \le k$. A token z witnesses (A,B) when $\langle x_z, y_a\rangle \ge 0$ for every $a \in A$ and $\langle x_b, y_z\rangle \ge 0$ for every $b \in B$. The first family of constraints involves only the query x_z and the active keys, and the second only the key y_z and the active queries. Since x_z and y_z are independent, the witness probability factors as a product of two sign-cone measures, over $|A|$ and $|B|$ vectors. By Theorem 5.3 the two factors are at least $e^{-|A|}$ and $e^{-|B|}$ when the corresponding Gram matrices have $\lambda_{\min} \ge \frac{1}{2}$. Theorem 5.1 supplies this for each of the two Gram matrices with failure probability at most $e^{-c_0 d}$ apiece. So the witness probability is at least $e^{-|A|-|B|} \ge e^{-k}$ off an event of probability at most $2e^{-c_0 d}$. The union bound of Theorem 5.4, whose count of at most $(3n)^k$ pairs holds also at $k=1$, where it is $1+2n$, then reads

$$\mathbf{Pr}[\text{the directed } C_k \text{ fails}] \le (3n)^k\left(e^{-e^{-k}(n-k)} + 2e^{-c_0 d}\right),$$

which is at most $1/n$ as soon as $e^{-k}(n-k)$ and $c_0 d$ are each at least $(k+2)\log(3n)$. The first holds at $n = 2^{C'k}$ for every $k \ge 1$ once $C' \ge 12$: there the left side is at least $\frac{1}{2}\cdot 2^{(C'-\log_2 e)k}$, since $k \le 2^{C'k}/2$, and the right side is at most $3k(2+C'k) \le 9C'k^2$, since $k+2 \le 3k$ and $\log\left(3\cdot 2^{C'k}\right) \le 2 + C'k$. The first bound exceeds the second: at $k=1$ and $C'=12$ they are about 753 and 108, and their ratio increases in both k and C'. It persists for every larger n, since the derivative in n of $e^{-k}(n-k) - (k+2)\log(3n)$ is $e^{-k} - (k+2)/n \ge 0$ once $n \ge (k+2)e^k$, which $n \ge 2^{C'k}$ ensures for $C' \ge 12$. The second holds for $d \ge C'k\log n$ with $C' \ge 6/c_0$, since $(k+2)\log(3n) \le 6k\log n$ for $n \ge 3$ and $k \ge 1$. Finally, $d \ge C'k\log n \ge C_1 k$ once $C' \ge C_1$.

For the lower bound, note that the failing pattern produced by Theorem 5.7 is one-sided: it is the pair $(A,\emptyset)$ with A a k-set of tokens, and the requirement on a witness z is $\langle x_z, y_a\rangle \ge 0$ for every $a \in A$ alone. Apply the base-pool and apex-pool argument of Theorem 5.7 to the independent Gaussian keys $y_1,\dots,y_n$. Its second-moment bound produces a good k-set of keys with high probability, and Theorem 5.5 bounds its positive cone by $e(C_\kappa\sin\theta)^{k-1}$. Conditional on the keys, all n queries are independent Gaussian candidates, so the probability that any query lies in this cone is at most $ne(C_\kappa\sin\theta)^{k-1} = o(1)$ by the witnessing bound in the proof of Theorem 5.7. Thus the directed k-extension property fails with probability $1-o(1)$. □

The covering heuristic behind Theorem 5.11. We expect a negative answer, if there is one, to come from a covering phenomenon. The $\binom{n/2}{k-1} = 2^{\Theta(k^2)}$ subset-average directions used in Theorem 5.7 are as numerous as a $2^{-\Theta(k)}$-net of S^{d-1} at $d = \Theta(k)$. If they were spread out like

a net, every token would nearly oppose the average of some subset, the configuration Theorem 5.5 exploits. Extending Theorem 5.7 to every head would still need the other hypotheses of Theorem 5.5, and a way to keep other tokens out of the resulting small cone, which the proof of Theorem 5.7 gets from independence. The three constructions we examined each appear to produce a degenerate cone, by a different route: the Gaussian one by probabilistic conspiracy, which Theorem 5.7 proves; points on the moment curve $t \mapsto (t, t^2, \ldots, t^d)$ by pigeonhole clustering into an ill-conditioned Gram matrix; and embeddings built from algebraic codes by exact linear dependency. We have not proved the last two.

The construction behind Theorem 5.13. Take queries $x_{a_i} = (i^2, -2i, 1, 1)$ and $x_{b_i} = (0, 0, 0, 1)$, and keys $y_{a_j} = (0, 0, 0, -1)$ and $y_{b_j} = (1, j, j^2, -\varepsilon)$, with $0 < \varepsilon < 1$. Then $\langle x_{a_i}, y_{b_j} \rangle = (i-j)^2 - \varepsilon$, which is nonnegative exactly when $i \neq j$, and every other score $\langle x_u, y_v \rangle$ is negative. In a closely related model, some posets do need large dimension. The reachability embeddings of Gokhale et al. (2026) are, off the diagonal, directed heads at threshold 0 with a strict inequality. In that model every rooted directed tree is realized in dimension 3 (Gokhale et al., 2026, Theorem 9), while a counting argument shows that some bipartite directed acyclic graphs on n vertices, which are posets of height two, need dimension $\Omega(n)$ (Gokhale et al., 2026, Theorem 16).

Tournaments of large μ at dimension six. This paragraph shows that the flip-class measure μ of Section 1.5 does not bound the dimension of a skew-symmetric head from below. Let C be the tournament on four vertices formed by a directed 3-cycle $c_1 \to c_2 \to c_3 \to c_1$ and a fourth vertex p with $p \to c_i$ for each i. Index the standard basis of $\mathbb{R}^4$ by the vertices of C, as $g_p, g_{c_1}, g_{c_2}, g_{c_3}$, and let A_4 be the signed adjacency matrix of C, with $(A_4)_{ab} = 1$ when $a \to b$ and $(A_4)_{ab} = -1$ when $b \to a$, for $a \neq b$, and zero diagonal. Then A_4 is skew-symmetric and $g_a^\top A_4 g_b = (A_4)_{ab}$, which is positive exactly when $a \to b$ in C, for $a \neq b$.

Every tournament in the flip class of C contains a directed 3-cycle. A cut and its complement reverse the same arcs, so up to complement and up to rotating the 3-cycle there are four cuts to check, with one side $\emptyset$, $\{p\}$, $\{c_1\}$ or $\{p, c_1\}$. The first two reverse no arc of the 3-cycle, which survives. Reversing across $\{c_1\}$ turns $c_1 \to c_2$ and $p \to c_1$ around and keeps $p \to c_2$, which gives the 3-cycle $p \to c_2 \to c_1 \to p$. Across $\{p, c_1\}$ the reversal turns $c_3 \to c_1$ and $p \to c_3$ around and keeps $p \to c_1$, which gives the 3-cycle $p \to c_1 \to c_3 \to p$. Relabelling the vertices keeps a 3-cycle, so the claim holds for the whole flip class.

Now fix $m \geq 2$, put $\delta = \pi/(m+1)$ and $A_2 = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, and take the $n = 4m$ vertices (t, a) with $t \in \{1, \ldots, m\}$ and $a \in \{p, c_1, c_2, c_3\}$. Give (t, a) the query $x_{(t,a)} = (\cos t\delta, \sin t\delta, \varepsilon g_a) \in \mathbb{R}^6$ and the key $\Lambda x_{(t,a)}$, where Λ is the skew-symmetric block-diagonal matrix with blocks A_2 and A_4, and $\varepsilon > 0$ is fixed below. The score from (t, a) to (s, b) is

$$x_{(t,a)}^\top \Lambda x_{(s,b)} = \sin\left((s-t)\delta\right) + \varepsilon^2 g_a^\top A_4 g_b.$$

For $s \neq t$ we have $\delta \leq |s-t|\,\delta \leq (m-1)\delta = \pi - 2\delta$, so $|\sin((s-t)\delta)| \geq \sin\delta$, and choosing ε with $\varepsilon^2 \max_{a,b} \left|g_a^\top A_4 g_b\right| < \sin\delta$ makes every arc between two blocks point from the smaller block index to the larger. For $s = t$ the first term vanishes, so the arcs inside each block are those of C. No score between distinct vertices is zero, so the head, which is the skew-symmetric specialization of Theorem 2.8 at threshold 0, realizes a tournament T.

Let T' be obtained from T by reversing the arcs across a cut $(S, \bar{S})$. Its restriction to a block B_t is C with the arcs across $(S \cap B_t, B_t \setminus S)$ reversed, which contains a directed 3-cycle by the claim above. A 3-cycle has an arc pointing backward in every ordering, so every ordering of T', and of any relabelling of it, has an arc pointing backward inside each of the m blocks, with its tail in that block. Hence $\mu(T) \geq m = n/4$, at dimension 6.