ON ATTENTION HEADS AND BILINEAR FORMS

ANDREW O'DESKY

Abstract. We study the symmetric and antisymmetric parts of bilinear forms in the attention heads of trained large language models. We introduce an orthogonally invariant profile map from real bilinear forms to a three-dimensional simplex and observe that profiles of trained bilinear forms accumulate near profiles of rank-one bilinear forms. We prove that the symmetric part of a bilinear form in an attention head is the sum of a hyperbolic form and a zero form for a Zariski-dense subset of query-key matrices.

1. Introduction

The stunning success of modern language models relies primarily on computational scale and the discovery of a new class of functions known as *transformers* [12]. Transformers are remarkably versatile at approximating complicated functions traditionally requiring sophisticated algorithms or human cognition, yet their versatility is not well-understood. Researchers from diverse areas have proposed different models for interpreting them but it remains unclear why transformers are so effective.

The novel ingredient in transformers is the attention mechanism. Language to be processed is modeled as a sequence of vectors in a real vector space, and the attention mechanism facilitates real-time interactions between words which encode context for later stages of the transformer. These interactions are computed using a pretrained *bilinear form* encoded in the attention weights. While real bilinear forms are well-understood mathematically, the relationship between properties of the bilinear forms in attention heads and aspects of language processing is poorly understood. This paper is an attempt to better understand that relationship. We also give a formulation of transformers equivalent to the usual construction which we hope will be inviting to mathematicians who are interested in transformers. For readers from the machine learning community, we have attempted to provide enough background so that the paper is self-contained.

1.1. Symmetric bilinear forms in attention heads are balanced

Any bilinear form B can be uniquely decomposed as $B = S + T$ where S is symmetric and T is antisymmetric. In recent work [5] on the `gpt2` language model, Migliarini observed that the symmetric part S of one of its head's bilinear forms had 33 negative eigenvalues among the top 64 eigenvalues when sorted by absolute magnitude, and hypothesized that the presence of negative eigenvalues was responsible for the "self-suppressing" behavior of this head. Migliarini's observation raises many questions about the distributions of these eigenvalues in general, foremost being whether this distribution is atypical. We inspected all 9,512 heads in the fourteen models listed in §3. Their eigenvalues give a striking observation.

Observation 1. Every attention head A has a bilinear form whose symmetric part S has precisely $2n$ nonzero eigenvalues, precisely half of which are negative, where $n = n_A$ is the head dimension.

In the language of bilinear forms, the symmetric bilinear form in every attention head A has signature $(n_A, n_A, N_A - 2n_A)$ where N_A is the embedding dimension. This observation is explained by the following theorem which proves this is a linear-algebraic consequence of the standard query–key construction for attention heads. Let C be a real vector space of dimension N and let ℓ be a positive integer. Let H be a real vector space of dimension n. For linear maps $Q, K\colon C \to H$, let $A\colon C^\ell \to C^\ell$ denote the length ℓ attention head on C constructed from Q and K as query and key transformations.

Theorem 1 (Theorem 3)**.** *Let S denote the symmetric component of the bilinear form of A. If the stacked map*

$$\begin{pmatrix} Q \\ K \end{pmatrix}\colon C \to H \oplus H, \qquad x \mapsto (Qx, Kx)$$

Date: September 17, 2026.

is surjective, then the signature of S is $(n, n, N-2n)$. In particular, if the head dimension is no more than half the embedding dimension then there is a Zariski-dense subset of query and key transformations whose attention head has this property.

The theorem can also be applied to RoPE and ALiBi attention to obtain the same conclusion.

1.2. **Profiles of bilinear forms in language models**

We next turn to properties which are learned rather than structural. To measure these we introduce a map on real bilinear forms valued in the 3-simplex which is invariant under positive scaling and orthogonal changes of coordinates (Definition 6):

$$\pi\colon M_N(\mathbb{R}) \smallsetminus \{0\} \longrightarrow \Delta_3, \qquad L = S + T \longmapsto (a, b, c, d), \qquad a, b, c \geq 0,\ \ a + b + c \leq 1.$$

We call $\pi(L)$ the *profile* of L. The purpose of π is to give a qualitative description of bilinear forms. The fibers of π over the faces of Δ_3 can be explicitly described (Theorem 4), and the profiles of rank-one forms fill out a one-parameter family $\Theta \subset \Delta_3$.

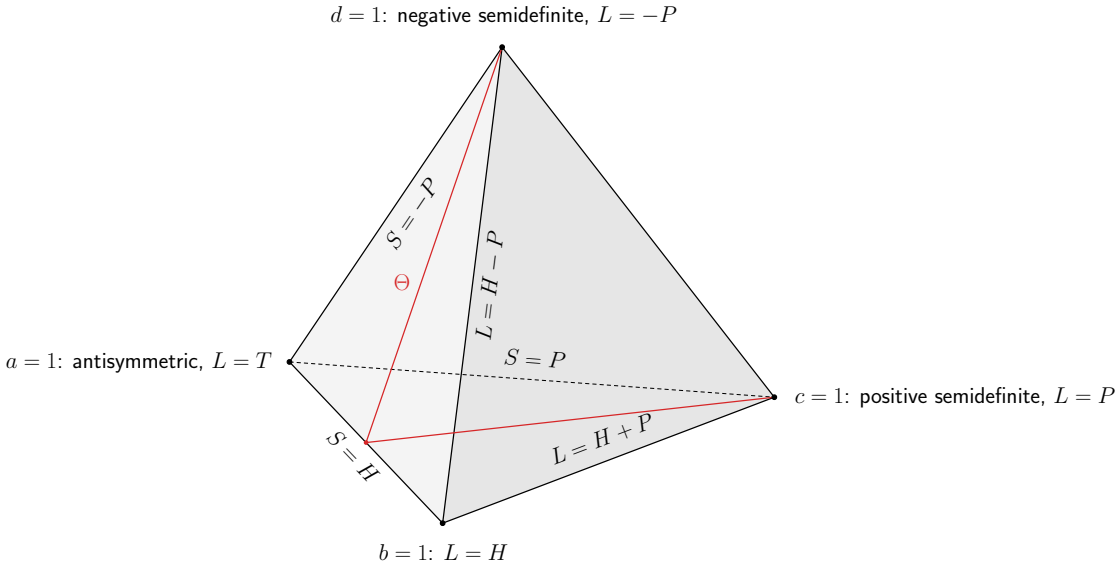

FIGURE 1. The simplex Δ_3 of profiles of nonzero bilinear forms. Red segments comprise $\Theta = \{a = b,\ cd = 0\}$. H denotes a symmetric matrix whose spectrum is symmetric about zero and P a positive semidefinite matrix. The front-right facet $a = 0$ corresponds to symmetric forms, the bottom facet $d = 0$ to $S = H + P$, and the left facet to $S = H - P$. Fibers of π over the relative interiors of the back facet $\{b = 0\}$ and the top-right edge are empty.

We examined the profiles of bilinear forms from fourteen language models. For large rank, the image of the profile map fills up most of the simplex. Nonetheless, the profiles of trained bilinear forms from language models tend to accumulate near profiles of rank-one forms.

Observation 2. In every one of the fourteen models except `opt-350m`, the profiles of their bilinear forms accumulate near the one-parameter family $\Theta = \{a = b,\ cd = 0\} \subset \Delta_3$ of profiles of rank-one forms.

Note that profiles of *randomly* initialized bilinear forms accumulate tightly at the mid-point $(\frac{1}{2}, \frac{1}{2}, 0, 0) \in \Theta$. Trained bilinear forms generally have large effective rank (see §5.2.2) so accumulation near Θ is not explained by proximity to low-rank forms. In short we do not have a complete explanation of Observation 2. The face $\Delta = \{a = 0\}$ contains the profiles of symmetric forms. Let $p\colon \Delta_3 \to \Delta$ denote the projection $p(a, b, c, d) = (0, a + b, c, d)$. As a partial explanation, we prove that $p(\Theta)$ is the limiting set of profiles in the large-dimensional limit for any random ensemble of symmetric matrices with proportional positive and negative eigenvalue distributions.

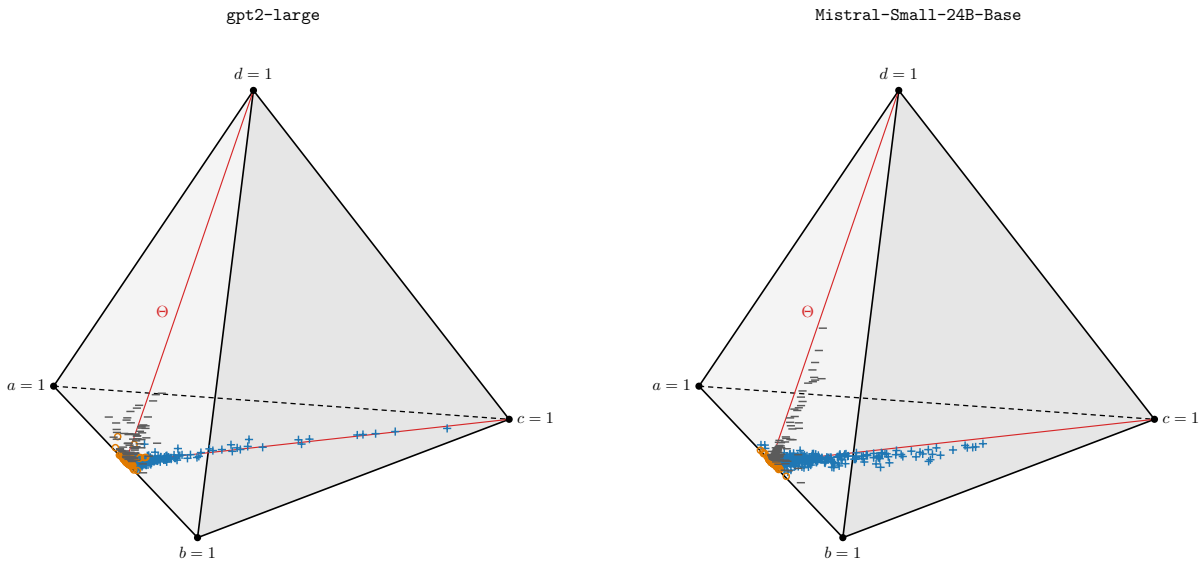

FIGURE 2. Profiles of the bilinear forms of `gpt2-large` and of `Mistral-Small-24B-Base`. The left model uses QKV attention while the right uses RoPE, showing that accumulation occurs for both types of positional encoding. Markers (+ Type I_+, ∘ Type II, − Type I_-) indicate parity.

Theorem 2 (Theorem 5)**.** *Let X be a positive random variable with finite, nonzero second moment. Fix $u, v > 0$ and set $\rho = u/v$. For each n let $N_n \geq 2n$ and let $S_n \in M_{N_n}(\mathbb{R})$ be a symmetric matrix with n positive eigenvalues sampled from vX, n negative eigenvalues sampled from $-uX$, and any remaining eigenvalues equal to zero. Then the profile $\pi(S_n)$ converges almost surely as $n \to \infty$ to*

$$\frac{1}{1+\rho^2}\big(0,\, 2\rho,\, \max(1-\rho, 0)^2,\, \max(\rho-1, 0)^2\big) \in p(\Theta).$$

The hypothesis about proportional positive and negative eigenvalue distributions is approximately satisfied by trained bilinear forms (Observation 5).

1.3. Eigenvalues of trained bilinear forms

Eigenvalue distributions of the bilinear forms in the `Qwen3` models were investigated by Darveshi [3]; they found that the absolute values of the eigenvalues are well-modelled by a gamma distribution. We performed the same experiment for the symmetric parts. Theorem 1 suggests the distribution should be multimodal: nonzero eigenvalues occur with equal frequency on either side of zero so the distribution should have at least two peaks. For $\nu \geq 0$, let γ_ν^+ denote a gamma-like probability distribution on $[0, \infty)$ with mode ν, and for $\mu \leq 0$ let γ_μ^- denote the reflection across 0 of a gamma-like probability distribution, supported on $(-\infty, 0]$ with mode μ.

Observation 3 (Figure 3)**.** Eigenvalues of the symmetric part S of trained bilinear forms have a bimodal distribution well-modelled by a sum of two gamma-like distributions with positive and negative modes:

$$\lambda(S) \sim \tfrac{1}{2}(\gamma_\mu^- + \gamma_\nu^+), \qquad \mu \leq 0 \leq \nu.$$

For 550 of the 9,512 heads (5.8%), one or both fitted modes were zero.

We observed that the modes μ, ν of the bimodal eigenvalue distributions predict which branch of Θ is nearest to the profile. The midpoint $(\frac{1}{2}, \frac{1}{2}, 0, 0)$ is denoted $\Theta_\circ$, the line segment from $\Theta_\circ$ to $c = 1$ is denoted Θ_+, and the line segment from $\Theta_\circ$ to $d = 1$ is denoted Θ_-. We call a head Type I_+ if its profile is near Θ_+, Type I_- if the profile is near Θ_-, and otherwise call it Type II (see §3.4 for details). We looked at the histogram of the mean sorted spectrum of the ten most extreme heads in each cluster (Figure 4) and observed the following relationships between modes and the three profile types:

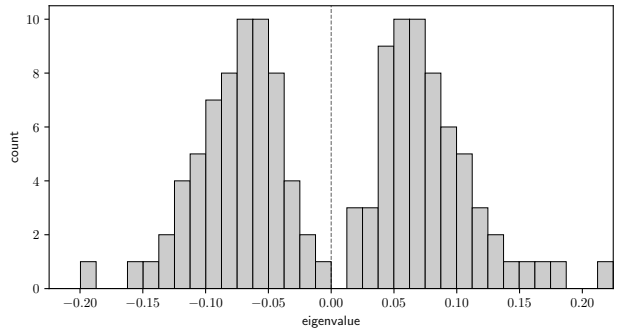

FIGURE 3. The mean eigenvalue distribution of symmetric parts for `gpt2`.

(1) Type I_+ (positive-definite-like): $|\mu| \ll \nu$;
(2) Type I_- (negative-definite-like): $\nu \ll |\mu|$;
(3) Type II (symmetric spectrum): $|\mu + \nu| \ll \min(|\mu|, \nu)$.

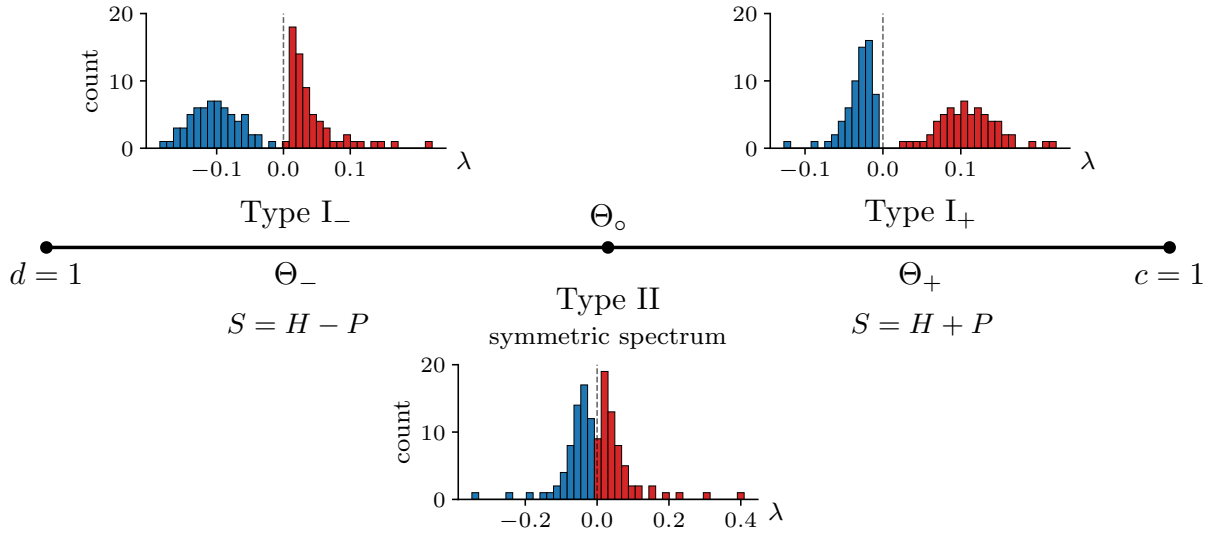

FIGURE 4. Measured eigenvalue distributions in `gpt2`, attached to their branches or midpoint of Θ. Each histogram shows the mean of the sorted, unit-length spectra of the ten most extreme heads in one cluster. Here H has spectrum symmetric about zero and P is positive semidefinite.

1.3.1. *Alignment with experimental classifiers.* To see whether there were other discerning features, we performed principal component analysis to obtain low-dimensional projections of the eigenvalue distributions of the symmetric parts of the trained attention heads: sorting each distribution and rescaling it to unit length gives a point of $\mathbb{R}^{2n}$, and the leading coordinates of the singular value decomposition of the resulting matrix project these to three dimensions. The projection is plotted in Figure 5 for the 144 heads of the `gpt2` model. The results show that parity is well-aligned with the experimentally obtained classifiers, so it may be helpful for classifying head behavior. A second set of head classifiers following a different approach with moments is described in Appendix C.

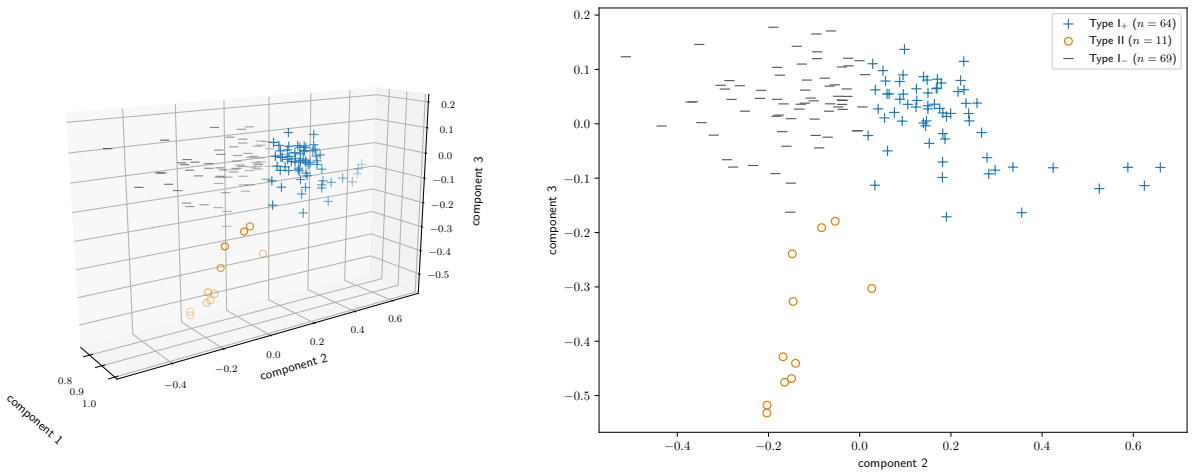

FIGURE 5. Projection of `gpt2`'s 144 symmetric part eigenvalue distributions to the first three principal components (left), and to the second and third components (right).

Notation

Let C be a finite-dimensional real vector space of dimension N. Let k be a nonnegative integer. Let

$$\Delta_k = \left\{ (p_0, \ldots, p_k) \in \mathbb{R}^{k+1}_{\geq 0} : \sum_{j=0}^{k} p_j = 1 \right\}$$

be the standard k-simplex. The identity map will be denoted id. The set of linear operators on C will be denoted $\mathrm{End}(C)$. The dual space of C, i.e. the vector space of linear maps $C \to \mathbb{R}$, will be denoted $C^\vee$. In what follows we make use of the canonical isomorphism of vector spaces $C^\ell \cong C \otimes \mathbb{R}^\ell$, in particular writing $U \otimes V \in \mathrm{End}(C^\ell)$ for $U \in \mathrm{End}(C)$, $V \in \mathrm{End}(\mathbb{R}^\ell)$. The set of $N \times N$ real matrices is denoted by $M_N(\mathbb{R})$. For a real matrix L let $\|L\| = \sqrt{\mathrm{tr}(L^T L)}$ denote the Frobenius norm.

2. Bilinear attention

In this section we give an equivalent definition for attention different from what is found in the literature. We restrict to decoder-only transformers and ignore layer normalization.

2.1. Bilinear forms

We recall some basic facts about real bilinear forms. A bilinear form on a real vector space C is a function

$$B \colon C \times C \to \mathbb{R}$$

which is linear in each argument separately, i.e. for all $x, y, z \in C$ and $\lambda \in \mathbb{R}$ we have

$$B(x+y,z) = B(x,z) + B(y,z), \qquad B(\lambda x, y) = \lambda B(x,y),$$
$$B(x,y+z) = B(x,y) + B(x,z), \qquad B(x,\lambda y) = \lambda B(x,y).$$

If $B(x,y) = B(y,x)$ (resp. $-B(y,x)$) for all $x, y \in C$ then B is called *symmetric* (resp. *antisymmetric*). Every bilinear form B can be uniquely expressed as a sum $B = S + T$ of a symmetric bilinear form S and an antisymmetric bilinear form T, where

$$S(x,y) = \tfrac{1}{2}\big(B(x,y) + B(y,x)\big), \qquad T(x,y) = \tfrac{1}{2}\big(B(x,y) - B(y,x)\big).$$

Given a basis $e_1, \ldots, e_N$ of C, the *Gram matrix* of B is the $N \times N$ matrix L with entries $L_{ij} = B(e_i, e_j)$. The *radical* of B is the subspace $\{x \in C : B(x,y) = 0 \text{ for all } y \in C\}$. A subspace $W \subset C$ is *isotropic* for B if $B(x,y) = 0$ for all $x, y \in W$. A nondegenerate antisymmetric bilinear form is called *symplectic*.

2.1.1. *Sylvester's Law of Inertia.* If B and B' are bilinear forms and there is an invertible linear map $f\colon C \to C$ such that $B'(x,y) = B(f(x), f(y))$ for all $x, y \in C$ then f is called an *isomorphism* between B and B'. Sylvester's Law of Inertia says that a symmetric bilinear form B on C is determined up to isomorphism by three non-negative integers (n_+, n_-, n_0). These are computed from the Gram matrix L of B, which is symmetric since B is: n_+ (resp. n_-) is the number of positive (resp. negative) eigenvalues of L counted with multiplicity, and n_0 is the dimension of the kernel of L. The tuple (n_+, n_-, n_0) is called the *signature* of B; by Sylvester's law it is independent of the chosen basis. The form B is *positive semidefinite* if $B(x,x) \geq 0$ for every x and *negative semidefinite* if $-B$ is positive semidefinite; it is *semidefinite* if it is one or the other. If $n_+ = n_-$ then we say B is *balanced*. A nondegenerate balanced form is called *hyperbolic*. Hyperbolicity requires equal numbers of positive and negative eigenvalues, but does not require their magnitudes to agree. Symmetry of the spectrum about zero is stronger than balance and depends on the chosen inner product. By Sylvester's law, there is exactly one hyperbolic form on $\mathbb{R}^{2n}$ up to isomorphism, namely the form with Gram matrix

$$\begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix} \in M_{2n}(\mathbb{R})$$

where I denotes the $n \times n$ identity matrix.

2.2. **Attention functions**

First we define a very general class of attention functions.

Definition 1. A smooth function $a\colon C^k \to C$ is an *attention function* if there are a smooth function

$$\omega = (\omega_1, \ldots, \omega_k)\colon C^k \to \Delta_{k-1},$$

a smooth convex function $\phi\colon \mathbb{R}^k \to \mathbb{R}$ and a polynomial map $\sigma\colon C^k \to \mathbb{R}^k$ such that

$$\omega = (\nabla\phi) \circ \sigma \qquad \text{and} \qquad a(x_1, \ldots, x_k) = \sum_{j=1}^{k} \omega_j(x_1, \ldots, x_k) x_j.$$

The components of ω are called the *attention weights*, σ is the *scoring function*, and ϕ we call the *potential function*; with an abuse of terminology, we regard ω, σ, ϕ as part of the data of the attention function.

For the next three sections, let H be a real inner product space of dimension n called the *head space*, and let $Q, K\colon C \to H$ be independent linear maps called the *query and key maps*.

2.3. **QKV attention**

We explain how to recover the definition of attention from [12]. First we construct the scoring function. To encode positional information, vectors $p_j \in C$ are fixed for each $j = 1, \ldots, k$. The score of x_k on x_j is the j-th component of σ, given by

$$\sigma_j(x_1, \ldots, x_k) = \langle Q(x_k + p_k),\, K(x_j + p_j)\rangle / \sqrt{n}.$$

Expanding, we get

$$\sigma_j(x_1, \ldots, x_k) = \frac{1}{\sqrt{n}}\Big(\langle Qx_k, Kx_j\rangle + \langle Qx_k, Kp_j\rangle + \langle Qp_k, Kx_j\rangle + \langle Qp_k, Kp_j\rangle\Big), \tag{1}$$

a polynomial function on C^k. The potential is given by[1]

$$\phi(s_1, \ldots, s_k) = \log(e^{s_1} + \cdots + e^{s_k}).$$

[1]To see that ϕ is convex, first set $\omega_j = e^{s_j} / \sum_i e^{s_i}$ and $\omega = (\omega_1, \ldots, \omega_k)^T$. Then

$$\frac{\partial \phi}{\partial s_j} = \omega_j, \qquad \frac{\partial^2 \phi}{\partial s_i \partial s_j} = \omega_i \delta_{ij} - \omega_i \omega_j,$$

so $\operatorname{Hess}\phi = \operatorname{diag}(\omega) - \omega\omega^T$. For any $v \in \mathbb{R}^k$, Cauchy–Schwarz applied to $(\sqrt{\omega_i})_i$ and $(\sqrt{\omega_i}\, v_i)_i$ gives

$$\Big(\sum_i \omega_i v_i\Big)^2 = \Big(\sum_i \sqrt{\omega_i} \cdot \sqrt{\omega_i}\, v_i\Big)^2 \leq \Big(\sum_i \omega_i\Big)\Big(\sum_i \omega_i v_i^2\Big) = \sum_i \omega_i v_i^2.$$

Therefore

$$v^T \operatorname{Hess}\phi\, v = \sum_i \omega_i v_i^2 - \Big(\sum_i \omega_i v_i\Big)^2 \geq 0,$$

so ϕ is convex by the Hessian criterion for convexity.

The attention weights ω formed from ϕ and σ recover QKV attention.[2]

2.4. **RoPE attention**

QKV attention encodes positional information in the linear and constant terms of the scores (1) and the bilinear term is *independent* of position. Rotary position embedding (RoPE) [11] is a variant of QKV attention with no linear or constant terms with all positional information in the bilinear terms. Assume $n = 2m$ is even and let $R\colon H \to H$ be an orthogonal transformation whose eigenvalues are

$$e^{\pm i\omega q^{r-1}}, \qquad r = 1, \dots, m,$$

for some fixed $\omega \in (0, \pi)$ and $0 < q < 1$ (in [11] one takes $\omega = 1$ and $q = 10000^{-1/m}$). In RoPE attention, the score of x_k on x_j is defined by

$$\sigma_j(x_1, \dots, x_k) = \langle Kx_j,\ R^{k-j}Qx_k\rangle/\sqrt{n}. \tag{2}$$

2.5. **ALiBi attention**

Attention with linear biases (ALiBi) [7] encodes positional information in the constant terms of the scores. Fix a positive scalar μ for each head. In ALiBi attention, the score of x_k on x_j is defined by

$$\sigma_j(x_1, \dots, x_k) = \langle Kx_j,\ Qx_k\rangle/\sqrt{n} - \mu(k-j).$$

The scalar μ is fixed before training, and the term $-\mu(k-j)$ penalizes attention to more distant inputs. There are no linear terms in the input vectors, and the bilinear term is independent of position.

2.6. **Biaffine attention**

We have seen that QKV attention encodes position in the linear and constant terms of the score and uses the same bilinear form at different distances; RoPE attention has no linear nor constant terms and encodes positional information by using distinct bilinear forms at different distances; ALiBi uses the same bilinear form at every distance and encodes position entirely in the constant terms. The bilinear, linear, and constant parts of a score are separated in Lemma 1 below. By lifting the constraints on the bilinear terms in RoPE and allowing for lower-order terms we obtain the following more general class of attention functions.

Definition 2. Let $a\colon C^k \to C$ be an attention function with scoring function $\sigma\colon C^k \to \mathbb{R}^k$ of the form

$$\sigma(x_1, \dots, x_k) = \big(P_{k-1}(x_1, x_k), P_{k-2}(x_2, x_k), \dots, P_1(x_{k-1}, x_k), P_0(x_k, x_k)\big)$$

for polynomials $P_j\colon C^2 \to \mathbb{R}$. By an abuse of terminology, if the polynomials P_j all have bidegree at most $(1, 1)$ then a is called *biaffine*. If the biaffine polynomials are bilinear forms then a is *bilinear*.

The following elementary lemma shows that the bilinear, linear, and constant components of a biaffine score are well-defined; we use it to attach a bilinear form to each attention head in §3.

Lemma 1. *Every polynomial $P : C \times C \to \mathbb{R}$ of bidegree at most $(1, 1)$ admits a unique decomposition*

$$P(u, v) = B(u, v) + L(u) + R(v) + D,$$

where B is bilinear, L and R are linear, and $D \in \mathbb{R}$. The components are given by

$$D = P(0, 0),\ \ L(u) = P(u, 0) - P(0, 0),\ \ R(v) = P(0, v) - P(0, 0),$$

and $B(u, v) = P(u, v) - P(u, 0) - P(0, v) + P(0, 0)$.

[2]Strictly speaking we have recovered "QK attention" with trivial value map $V = \mathrm{id}$. In our formulation of attention, value maps come at a later stage; see §2.10.

2.6.1. *RoPE-like attention functions.* In [11, §3.4.1] there is an argument given to justify their choice of bilinear form (cf. (2) with $d = k - j$)

$$P_d(x, y) = \langle Kx, R^d Qy\rangle/\sqrt{n}.$$

We give another derivation of their formula and generalize it to a larger class of bilinear attention functions. Let $a\colon H^k \to H$ be a bilinear attention function with scoring function $\sigma\colon H^k \to \mathbb{R}^k$ given by

$$\sigma(u_1, \dots, u_k) = \big(B_{k-1}(u_1, u_k),\ B_{k-2}(u_2, u_k),\ \dots,\ B_0(u_k, u_k)\big)$$

for bilinear forms $B_{k-1}, \dots, B_0$ on $H \times H$. Each bilinear form B_j determines a map $L_j\colon H \to H^\vee$ defined by $L_j(v)(u) = B_j(u, v)$. Assume that B_0 is nondegenerate. Define the linear operator

$$T_d = L_0^{-1} L_d \in \mathrm{End}(H).$$

Definition 3. If $T_{d+e} = T_d T_e$ for all integers $d, e \geq 0$ with $d + e < k$ then a is called *RoPE-like*.

The next proposition classifies RoPE-like bilinear attention functions.

Proposition 1. *A bilinear attention function a as above with nondegenerate B_0 is RoPE-like if and only if there exists a linear map $S \in \mathrm{End}(H)$ such that $T_d = S^d$ for all $0 \leq d < k$.*

Note: when S exists it is unique and given by $S = T_1 = L_0^{-1} L_1$.

Proof. Let a be a RoPE-like bilinear attention function. Then $L_{d+e} = L_d L_0^{-1} L_e$ for all $d, e \geq 0$ such that $d + e < k$; in particular, this shows that $L_{d+1} = L_d S$. By induction, we conclude that $L_d = L_0 S^d$ and therefore $T_d = S^d$. The other direction is trivial. □

For RoPE itself, the head-space forms are $B_d(u, v) = \langle u, R^d v\rangle/\sqrt{n}$. The form B_0 is a scaled inner product, so it is nondegenerate, and $L_d = L_0 R^d$. Thus $T_d = R^d$, and these forms define a RoPE-like attention function on H with $S = R$. The scores on C in §2.4 are recovered by setting $P_d(x, y) = B_d(Kx, Qy)$.

2.7. **Rank reduction**

It is crucial to parametrize through families of low-rank attention functions. Let E be a real finite-dimensional vector space. Let $F\colon C^k \to E^k$ be a polynomial mapping. We have a family of attention functions on C^k given by "pulling back along F"

$$\big\{\text{attention functions } E^k \to E\big\} \to \big\{\text{attention functions } C^k \to C\big\}$$
$$a = (\sigma, \phi) \mapsto F^* a = (\sigma \circ F, \phi).$$

2.7.1. *The standard construction.* Set $E = H \oplus H$, let $\mu \geq 0$ and let $R \in \mathrm{End}(H)$ be an orthogonal transformation. Set

$$\sigma((q_1, k_1), \dots, (q_k, k_k)) = (\langle R^{k-1} q_k, k_1\rangle, \langle R^{k-2} q_k, k_2\rangle, \dots, \langle q_k, k_k\rangle)/\sqrt{n} + c, \qquad c = -\mu(k-1, k-2, \dots, 0).$$

This defines a polynomial map $\sigma\colon E^k \to \mathbb{R}^k$. Let $\binom{Q}{K}\colon C \to E : x \mapsto (Qx, Kx)$ be a linear map. Let $p \in C^k$ and let $F\colon C^k \to E^k$ be given by

$$F(x_1, \dots, x_k) = ((Qy_1, Ky_1), (Qy_2, Ky_2), \dots, (Qy_k, Ky_k)), \qquad y = x + p.$$

The pullback of σ by F gives a family of attention functions on C^k including QKV, RoPE, and ALiBi attention functions as special cases.

2.8. **Potentials**

To orient the reader we give some examples of potential functions.

Softmax potential. Let $\phi(s_1, \dots, s_k) = \log(e^{s_1} + \cdots + e^{s_k})$. The attention function is given by

$$a(x_1, \dots, x_k) = \omega_1 x_1 + \cdots + \omega_k x_k \quad \text{where} \quad \omega_j = \frac{e^{\sigma_j(x)}}{\sum_{i=1}^k e^{\sigma_i(x)}}.$$

Linear potential. As a degenerate case we take a linear potential function ϕ. Then $\omega = \nabla\phi \circ \sigma = \nabla\phi$ is independent of the scoring function. In order for the weights to be simplex-valued, the potential must be given by $\phi(s) = p_0 \cdot s$ for some $p_0 \in \Delta_{k-1}$. Upon identifying $C \otimes \mathbb{R}^k = C^k$ the attention function is given by

$$a(x_1, \dots, x_k) = (\mathrm{id} \otimes \phi)(x_1, \dots, x_k).$$

We give two examples of linear potentials. Let $\phi = \frac{1}{k}(s_1 + \cdots + s_k)$. The associated attention function is

$$a(x_1, \ldots, x_k) = \tfrac{1}{k}(x_1 + \cdots + x_k).$$

Next fix $m \in \{1, \ldots, k\}$ and let $\phi(s_1, \ldots, s_k) = s_m$. Then $\nabla\phi = e_m$, the m-th standard basis vector, so

$$a(x_1, \ldots, x_k) = x_m.$$

Remark 1 (Max potential). Softmax potential is a smoothed form of the max function: for $\tau > 0$ let

$$\phi_\tau(s_1, \ldots, s_k) = \tau \log\Big(e^{s_1/\tau} + \cdots + e^{s_k/\tau}\Big).$$

This potential is smooth and convex with gradient

$$(\nabla\phi_\tau(s))_j = \frac{e^{s_j/\tau}}{\sum_i e^{s_i/\tau}}.$$

At $\tau = 1$ this is softmax, and as $\tau \to 0^+$, ϕ_τ converges pointwise to the max potential

$$\phi_0(s_1, \ldots, s_k) = \max(s_1, \ldots, s_k).$$

On the open subset where ϕ_0 is differentiable, the attention function outputs the token at the position with the largest score. While we have not attempted to do so, it would be natural and straightforward to relax the definition of attention functions to include such mildly nondifferentiable potentials.

2.9. **Attention heads**

Definition 4. For each integer $k = 2, \ldots, \ell$ let $a_k \colon C^k \to C$ be an attention function. The *(causal) attention head* $A \colon C^\ell \to C^\ell$ *associated to* $a_2, \ldots, a_\ell$ is

$$A(x_1, \ldots, x_\ell) = \big(x_1, a_2(x_1, x_2), \ldots, a_\ell(x_1, \ldots, x_\ell)\big).$$

(Causality is enforced by the independence of a_k from x_j for $j > k$.)

Example 1. We use the softmax potential $\phi(s_1, \ldots, s_k) = \log(e^{s_1} + \cdots + e^{s_k})$. The length two ($k = 2$) bilinear attention head is

$$A(x, y) = \big(x, y + (x - y)\sigma\big) \quad \text{where} \quad 0 < \sigma = \frac{e^b}{1 + e^b} < 1, \quad b = B_1(x, y).$$

The length three ($k = 3$) bilinear attention head is

$$A(x, y, z) = \big(x,\ y + (x - y)\sigma,\ z + (x - z)\sigma_1 + (y - z)\sigma_2\big)$$

where $\sigma = e^b/(1 + e^b)$, $b = B_1(x, y)$ as above, and

$$\sigma_1 = \frac{e^{b_2}}{1 + e^{b_1} + e^{b_2}}, \qquad \sigma_2 = \frac{e^{b_1}}{1 + e^{b_1} + e^{b_2}}, \qquad b_1 = B_1(y, z), \quad b_2 = B_2(x, z).$$

2.10. **Transformers**

Transformers are a type of piecewise-smooth function $T \colon C^\ell \to C^\ell$ operating on sequences of elements in C. Let $A_1, \ldots, A_h \colon C^\ell \to C^\ell$ be attention heads and let $V_1, \ldots, V_h \colon C \to C$ be linear maps, called *value maps*. The *attention block associated to* $\{(A_i, V_i)\}_{i=1}^h$ is

$$A = \sum_{i=1}^{h} (V_i \otimes \mathrm{id}) \circ A_i.$$

Let $A \colon C^\ell \to C^\ell$ be an attention block and let $M \colon C \to C$ be a piecewise smooth function. The *attention component* $T_A \colon C^\ell \to C^\ell$ is defined as

$$T_A = \mathrm{id} \otimes \mathrm{id} + A,$$

or explicitly, if $a_1, \ldots, a_\ell$ are the components of A, then

$$T_A(x_1, \ldots, x_\ell) = (x_1 + a_1(x_1), x_2 + a_2(x_1, x_2), \ldots, x_\ell + a_\ell(x_1, \ldots, x_\ell)).$$

The *neural network component* $T_M \colon C^\ell \to C^\ell$ is defined as

$$T_M = \mathrm{id} \otimes \mathrm{id} + M \otimes \mathrm{id},$$

or explicitly,

$$T_M(y_1, \ldots, y_\ell) = (y_1 + M(y_1), y_2 + M(y_2), \ldots, y_\ell + M(y_\ell)).$$

Definition 5. The *transformer* $T\colon C^\ell \to C^\ell$ *associated with* A, M is the function given by

$$T = T_M \circ T_A = (\mathrm{id}\otimes \mathrm{id} + M \otimes \mathrm{id}) \circ (\mathrm{id}\otimes \mathrm{id} + A).$$

Note that the neural network M acts in parallel (component-wise) while cross-term interactions only arise from the attention block.

2.10.1. *Transformers are unipotent.* The presence of the so-called "residual connection" is a significant design decision which has the effect of making transformers near-identity or "unipotent". (In our formulation, the residual connection corresponds to the $\mathrm{id}\otimes\mathrm{id}$ summand in T_A and T_M.) The next proposition makes this precise with a basic estimate.

Proposition 2. *Fix a norm* $\|\cdot\|$ *on* C*, and for* $x = (x_1, \dots, x_\ell) \in C^\ell$ *set* $\|x\|_\infty = \max_k \|x_k\|$*. Let* T *be the transformer associated to an attention block* $A = \sum_{i=1}^h (V_i \otimes \mathrm{id}) \circ A_i$ *and a neural network component* M*, and suppose that* $\|M(y)\| \le L\,\|y\|$ *for all* $y \in C$*. Set* $V = \sum_{i=1}^h \|V_i\|$*, where* $\|V_i\|$ *denotes the operator norm. Then for all* $x \in C^\ell$*,*

$$\|T(x) - x\|_\infty \le \big(L + V + LV\big)\,\|x\|_\infty.$$

Proof. Since the attention weights take values in the simplex Δ_{k-1}, every attention function satisfies

$$\|a(x_1, \dots, x_k)\| = \Big\|\sum_{j=1}^k \omega_j x_j\Big\| \le \sum_{j=1}^k \omega_j \|x_j\| \le \max_j \|x_j\|,$$

so every component of an attention head A_i has norm at most $\|x\|_\infty$, the first component being x_1 itself. The components of the attention block therefore satisfy

$$\|A(x)_k\| = \Big\|\sum_{i=1}^h V_i\big(A_i(x)_k\big)\Big\| \le \sum_{i=1}^h \|V_i\|\,\|x\|_\infty = V\|x\|_\infty.$$

Writing $T = T_M \circ T_A$, the k-th component of $T(x) - x$ is

$$A(x)_k + M\big(x_k + A(x)_k\big),$$

whose norm is at most

$$V\|x\|_\infty + L\big(\|x\|_\infty + V\|x\|_\infty\big) = \big((1+L)(1+V) - 1\big)\|x\|_\infty. \qquad \square$$

Example 2. For a fully connected neural network $M(y) = W_2\,\rho(W_1 y)$ whose activation ρ is 1-Lipschitz and fixes 0 (e.g. RELU), the hypothesis holds with $L = \|W_2\|\,\|W_1\|$.

2.10.2. *Standard construction (multiple heads).* The standard construction (§2.7.1) extends naturally to multiple heads. For each $i = 1, \dots, h$ let $A_i\colon C^\ell \to C^\ell$ be an attention head constructed from attention functions using the standard construction, and let $V_i\colon C \to C$ be a linear map factoring through H.[3] The *multi-head attention block associated to* $\{(A_i, V_i)\}_{i=1}^h$ is the attention block $A = \sum_{i=1}^h (V_i \otimes \mathrm{id}) \circ A_i$.

2.11. **Parameter counts**

We compare the counts of trainable parameters for the variants of attention introduced above. The potential function ϕ is regarded as fixed. The positional vectors p_j in QKV attention are also regarded as fixed, as are the orthogonal transformation in RoPE attention and the μ parameter in ALiBi attention. For biaffine and bilinear attention we assume the scoring polynomials depend only on the distance between inputs: the attention functions $a_2, \dots, a_\ell$ of the head share a single list of polynomials $P_0, \dots, P_{\ell-1}$, with a_k using $P_{k-1}, \dots, P_0$. A biaffine attention head on C^ℓ is then determined by ℓ biaffine maps. The value maps enter the attention block in the same way for every variant so we have excluded their parameters.

[3] This map is typically written $V_i = W_{O,i}W_{V_i}$ where $W_{V_i}\colon C \to H$ and $W_{O,i}\colon H \to C$ is the "output map".

Model	Scoring parameters ($V = \mathrm{id}$)
QKV/RoPE/ALiBi attention	$2N^2$
Biaffine attention	$\ell(N+1)^2$
Bilinear attention	ℓN^2
Multi-head QKV/RoPE/ALiBi attention	$2N^2$
Multi-head bilinear	$\dfrac{\ell N^2}{h}$

TABLE 1. Trainable scoring parameter counts in the attention blocks; value map and positional encoding parameters are excluded. Here $N = \dim C$, ℓ is the sequence length, and h is the number of heads. The last row uses fixed coordinate projections as described above.

2.12. **Bilinear attention recovers QKV attention**

Although QKV attention introduces linear and constant terms due to the positional embedding vectors, it can be realized as bilinear attention using a basic construction in projective geometry called homogenization.

Proposition 3. *Let $\widehat{C} = C \oplus \mathbb{R}e$ and let $\iota\colon C \to \widehat{C}$, $x \mapsto x + e$. For every single-head QKV transformer $T\colon C^\ell \to C^\ell$ there is a single-head bilinear transformer $\widehat{T}\colon \widehat{C}^\ell \to \widehat{C}^\ell$ with the same potential such that $\widehat{T}(\iota x_1, \dots, \iota x_\ell) = \iota^\ell\big(T(x_1, \dots, x_\ell)\big)$ for all $(x_1, \dots, x_\ell) \in C^\ell$.*

Proof. We will show that every QKV transformer on C is realized on the affine chart $\iota(C) = C + e \subset \widehat{C}$ by a bilinear transformer on $\widehat{C}$. For a QKV head with query and key maps $Q, K\colon C \to H$ and fixed positional encoding vectors $p_1, \dots, p_\ell \in C$, the score of the k-th input x_k on the j-th input x_j is

$$\begin{aligned} s_{kj}(x_j, x_k) &= \frac{\langle Q(x_k + p_k), K(x_j + p_j)\rangle}{\sqrt{n}} \\ &= \frac{1}{\sqrt{n}}\Big(\langle Qx_k, Kx_j\rangle + \langle Qx_k, Kp_j\rangle + \langle Qp_k, Kx_j\rangle + \langle Qp_k, Kp_j\rangle\Big). \end{aligned}$$

Define a bilinear form $\widehat{B}_{k-j,j}$ on $\widehat{C}$ by

$$\widehat{B}_{k-j,j}(u + \alpha e, v + \beta e) \coloneqq \frac{1}{\sqrt{n}}\Big(\langle Qv, Ku\rangle + \alpha\langle Qv, Kp_j\rangle + \beta\langle Qp_k, Ku\rangle + \alpha\beta\langle Qp_k, Kp_j\rangle\Big).$$

Then $\widehat{B}_{k-j,j}(\iota x_j, \iota x_k) = s_{kj}(x_j, x_k)$ so the bilinear scores of the homogenized attention head agree with the original QKV scores on the affine chart $C + e$. Since the score vectors are equal, the attention weights are equal. If the original attention function is $a_k(x_1, \dots, x_k) = \sum_{j=1}^k \omega_j x_j$ then the homogenized bilinear attention function satisfies

$$\begin{aligned} \widehat{a}_k(\iota x_1, \dots, \iota x_k) &= \sum_{j=1}^k \omega_j \iota(x_j) \\ &= \sum_{j=1}^k \omega_j (x_j + e) \\ &= \left(\sum_{j=1}^k \omega_j x_j\right) + \left(\sum_{j=1}^k \omega_j\right) e \\ &= a_k(x_1, \dots, x_k) + e \\ &= \iota(a_k(x_1, \dots, x_k)). \end{aligned}$$

Extend the value map and the neural network by ignoring the new coordinate: $\widehat{V}(u+\alpha e) \coloneqq V(u)$, $\widehat{M}(u+\alpha e) \coloneqq M(u)$. For any $x, z \in C$ we have $\iota(x) + \widehat{V}(\iota z) = \iota(x + Vz)$ and $\iota(x) + \widehat{M}(\iota x) = \iota(x + M(x))$ so the transformer satisfies $\widehat{T}(\iota x_1, \dots, \iota x_\ell) = \iota^\ell\big(T(x_1, \dots, x_\ell)\big)$. □

3. Spectral experiments on bilinear forms

In this section we inspect the bilinear forms in the attention heads of open-source large language models.

3.1. Models

We analyze fourteen publicly available pretrained language models, comprising 9,512 attention heads. The models have head dimensions 64, 80, or 128 and use learned absolute positional embeddings, ALiBi, or rotary position embeddings (RoPE):

- `distilgpt2` [9]: distilled `gpt2`, 6 layers, 12 heads per layer (72 heads), $N = 768$, and $n = 64$.
- `gpt2` [8]: GPT-2 small, 12 layers, 12 heads per layer (144 heads), $N = 768$, and $n = 64$.
- `gpt2-medium` [8]: 24 layers, 16 heads per layer (384 heads), $N = 1024$, and $n = 64$.
- `gpt2-large` [8]: 36 layers, 20 heads per layer (720 heads), $N = 1280$, and $n = 64$.
- `opt-125m` [13]: OPT decoder-only, 12 layers, 12 heads per layer (144 heads), $N = 768$, and $n = 64$.
- `opt-350m` [13]: 24 layers, 16 heads per layer (384 heads), $N = 1024$, and $n = 64$.
- `opt-1.3b` [13]: 24 layers, 32 heads per layer (768 heads), $N = 2048$, and $n = 64$.
- `opt-2.7b` [13]: 32 layers, 32 heads per layer (1,024 heads), $N = 2560$, and $n = 80$.
- `opt-6.7b` [13]: 32 layers, 32 heads per layer (1,024 heads), $N = 4096$, and $n = 128$.
- `opt-13b` [13]: 40 layers, 40 heads per layer (1,600 heads), $N = 5120$, and $n = 128$.
- `bloom-3b` [2]: 30 layers, 32 heads per layer (960 heads), $N = 2560$, $n = 80$, and ALiBi positional encoding.
- `bloom-7b1` [2]: 30 layers, 32 heads per layer (960 heads), $N = 4096$, $n = 128$, and ALiBi positional encoding.
- `pythia-70m-deduped` [1]: GPT-NeoX family, 6 layers, 8 heads per layer (48 heads), $N = 512$, $n = 64$, and RoPE on the first 25% of each head's dimensions.
- `Mistral-Small-24B-Base-2501` [6]: 40 layers, 32 query heads per layer (1,280 heads), 8 key–value heads, $N = 5120$, $n = 128$, and RoPE.

As we saw in §2.3, QKV attention is the special case of biaffine attention where the bilinear components (in the sense of Lemma 1) of the scoring function all equal the same bilinear form. In particular, a QKV attention head A is associated with a single bilinear form B_A on C^2. In the notation of [12] where $C = \mathbb{R}^N$ and $H = \mathbb{R}^n$ and $Q, K \in M_{n\times N}(\mathbb{R})$, this bilinear form is given by $B_A(x, y) = x^T K^T Q y$. We write L_A for the Gram matrix $K^T Q$. For `pythia-70m-deduped` and `Mistral-Small-24B-Base-2501`, which use RoPE attention, the form $L_A = K^T Q$ reported in the cross-model comparisons is the base form at relative position $d = 0$.

3.2. Symmetry

For each head A we asked whether the matrix $L = L_A$ was biased towards either symmetry or skew-symmetry. Let $\|L\|$ denote the Frobenius norm. Define the *symmetric energy*

$$\mathscr{S}(A) := \frac{\|\frac{1}{2}(L + L^T)\|^2}{\|L\|^2}.$$

The expected value of the symmetric energy for a product $L = K^T Q$ where K, Q are $n \times N$ random matrices with iid standard normal entries is $1/2 + 1/(2N)$ (Proposition 5).

Observation 4. Attention heads show a bias for symmetric forms: their symmetric energies $\mathscr{S}$ lie predominantly above the random baseline (Figure 13); the per-model share of heads above it ranges from 76% to 99%.

This computation corroborates the findings of [10]. Some heads are almost entirely symmetric.

3.3. **Pairing score**

Write $S = \frac{1}{2}(L + L^T)$. Let λ_+ (resp. λ_-) denote the vector of positive (resp. absolute values of negative) eigenvalues of S sorted in decreasing order. By appending zeros as needed, we will assume λ_+ and λ_- have the same length. Define the *pairing score*

$$\mathscr{P}(A) := 1 - \frac{\|\lambda_+ - \lambda_-\|}{\|S\|}.$$

The pairing score is one if and only if the set of eigenvalues is symmetric about zero. It is determined by the inner product of the two lists, and is bounded in terms of the ratio of their norms.

Lemma 2. *Let S be a symmetric bilinear form with $\lambda_+ \neq 0$, and let $\rho = \|\lambda_-\|/\|\lambda_+\|$. Then*

$$\mathscr{P} = 1 - \sqrt{1 - \frac{2\langle \lambda_+, \lambda_- \rangle}{\|S\|^2}} \;\leq\; 1 - \frac{|1-\rho|}{\sqrt{1+\rho^2}},$$

with equality if and only if $\lambda_- = \rho\,\lambda_+$.

Proof. Since $\|S\|^2 = \|\lambda_+\|^2 + \|\lambda_-\|^2$, we have $\|\lambda_+ - \lambda_-\|^2 = \|S\|^2 - 2\langle \lambda_+, \lambda_- \rangle$, which gives the equality. By the triangle inequality, $\|\lambda_+ - \lambda_-\| \geq \big|\|\lambda_+\| - \|\lambda_-\|\big| = |1-\rho|\,\|\lambda_+\|$, with equality if and only if one of $\lambda_\pm$ is a nonnegative multiple of the other, that is, $\lambda_- = \rho\,\lambda_+$. Dividing by $\|S\|^2 = (1+\rho^2)\,\|\lambda_+\|^2$ gives the inequality. □

Observation 5. The positive eigenvalues and the absolute values of the negative eigenvalues of each symmetric part, listed in decreasing order, are approximately proportional. Set $\rho = \|\lambda_-\|/\|\lambda_+\|$. The relative error

$$\frac{\|\lambda_- - \rho\,\lambda_+\|}{\|\lambda_-\|}$$

has median 0.13 over the heads, with 80% between 0.05 and 0.33; the per-model median lies between 0.08 and 0.18. Consistently, the pairing score is close to the bound of Lemma 2, whose equality case is proportionality: the per-model median gap between that bound and $\mathscr{P}$ lies between 0.017 and 0.041.

Figure 14 compares the pairing scores of trained heads with a Gaussian QK baseline, and Figure 11 shows their distributions by model. The per-model median pairing score lies between 0.77 and 0.87.

3.4. **The parity of trained heads**

The midpoint $(\frac{1}{2}, \frac{1}{2}, 0, 0)$ is denoted $\Theta_\circ$, the line segment from $\Theta_\circ$ to $c = 1$ is denoted Θ_+, and the line segment from $\Theta_\circ$ to $d = 1$ is denoted Θ_-. For a bilinear form L with nonzero symmetric part S, write $\pi(S) = (0, b_S, c_S, d_S)$. We use the neighborhood

$$\mathscr{N} = \{(0, b, c, d) \in \Delta : c + d \leq \tfrac{1}{50}\}$$

of the vertex $(0,1,0,0)$ in the face $\Delta = \{a = 0\}$. We call L *Type* II if $\pi(S) \in \mathscr{N}$. Outside $\mathscr{N}$, we call it *Type* I_+ if $c_S > d_S$, and *Type* I_- otherwise. Figure 12 shows how these proportions vary across layers in each model, using the same partition as Table 2.

3.5. **Eigenvalue distributions of trained heads**

Whereas earlier experiments used coarse measures on eigenvalues, in this section we attempt to model the distribution itself. Darveshi [3] observed that the absolute values of the eigenvalue distributions of the bilinear forms in the `Qwen3` family are well-modeled by a gamma distribution. We corroborate their findings for the fourteen models in our survey. We also examined the symmetric and antisymmetric parts separately. For the symmetric parts, the distribution is a bimodal sum of gamma-like distributions (Observation 3). For the absolute-values and the antisymmetric parts, we made the following observation.

Observation 6. The absolute values of the nonzero eigenvalues of the bilinear forms, and the positive imaginary parts $\sigma(T)$ of the eigenvalues of their antisymmetric parts, are unimodal and well-modeled by a gamma distribution.

model	heads	Type I_+	Type II	Type I_-
`distilgpt2`	72	34.7	26.4	38.9
`gpt2`	144	36.8	34.0	29.2
`gpt2-medium`	384	54.2	28.6	17.2
`gpt2-large`	720	64.3	22.6	13.1
`opt-125m`	144	79.2	16.7	4.2
`opt-350m`	384	78.9	20.8	0.3
`opt-1.3b`	768	72.3	17.6	10.2
`opt-2.7b`	1,024	71.4	14.8	13.8
`opt-6.7b`	1,024	78.7	15.9	5.4
`opt-13b`	1,600	75.7	6.0	18.3
`bloom-3b`	960	29.7	49.9	20.4
`bloom-7b1`	960	23.2	55.2	21.6
`pythia-70m-deduped`	48	18.8	50.0	31.2
`Mistral-Small-24B-Base`	1,280	36.8	53.8	9.5
all	9,512	57.4	28.5	14.1

TABLE 2. Percentage of the heads by parity, classified using $\pi(S)$, with tolerance $c_S + d_S \leq \frac{1}{50}$ for Type II.

3.5.1. *Gamma vs. lognormal vs. Weibull.* We compared the gamma distribution's fit with two other families. For each head we fitted a gamma, a lognormal and a Weibull distribution to each of the following quantities: the magnitudes $|\lambda(L)|$ of the nonzero eigenvalues of the full form, the two lobes $\lambda_+(S)$ and $\lambda_-(S)$ of the spectrum of the symmetric part, and the positive imaginary parts $\sigma(T)$ of the eigenvalues of the antisymmetric part. Figure 6 reports, for each fit, the average error between the observed histogram and the fitted density.

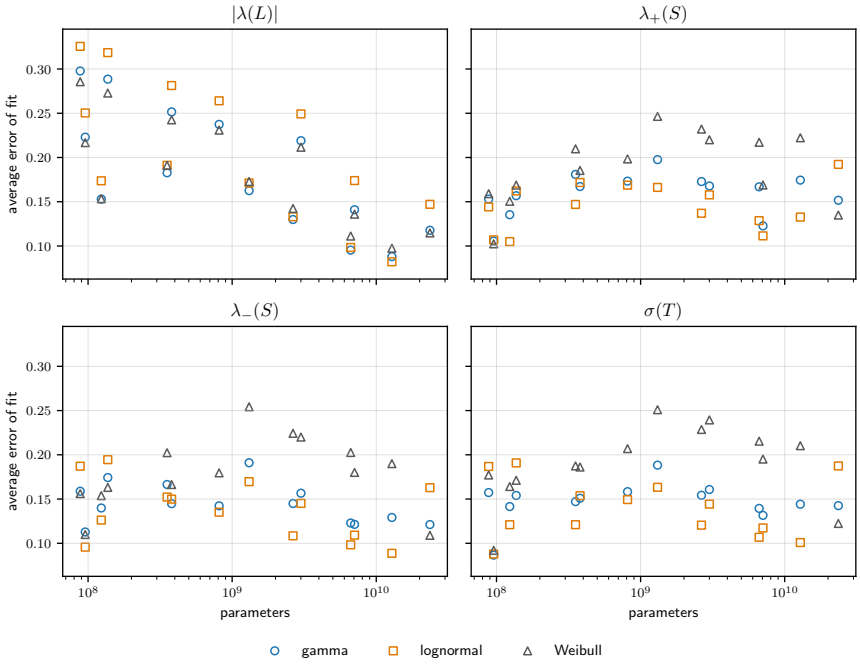

FIGURE 6. Average error of each fitted family over the heads of a model, against the size of the model; lower is a closer fit. The error is the root mean square difference between the heights of a 20-bin histogram of the sample and the fitted density, after rescaling the sample to unit mean.

The three families fit about equally well. Over all heads, the median error is between 0.12 and 0.21 for every family and every quantity. The lognormal is closest on the two lobes and on $\sigma(T)$ in a majority of models, while on $|\lambda(L)|$ the gamma and the Weibull are closest about equally often.

3.5.2. *Relations between $\lambda_\pm$ and σ.* We observed that the eigenvalue distributions of S and T are closely related. To explain this we use the gamma fit in the observations below because these relations can be naturally described using its shape and mode parameters. Let k_T and τ denote the shape and mode of $\sigma(T)$, and $k_\pm$ and μ, ν the shapes and the negative and positive modes of the two lobes of the eigenvalue distribution of S. For the comparisons of modes and shapes, we retain only heads for which all three fitted distributions have positive modes. (This excludes 550 of the 9,512 bilinear forms because at least one fitted distribution has mode zero, leaving 8,962 forms.)

Observation 7. The fitted mode of $\sigma(T)$ is close to the geometric mean of the two lobe modes of S: the median of $\tau/\sqrt{|\mu|\nu}$ over the retained heads of each model lies between 1.01 and 1.05.

It is interesting to note this relation holds exactly for rank-one forms. If $L = h\,uv^T$ for $h > 0$ with unit vectors u, v and $t = u^T v$, then S has the nonzero eigenvalues $h(1 \pm t)/2$ and T has the single repeated singular value $h\sqrt{1-t^2}/2$, which is their geometric mean. Nonetheless trained bilinear forms are far from rank one (cf. §5.2.2).

Observation 8. The fitted shape of $\sigma(T)$ is close to the geometric mean of the two lobe shapes of S, with a modest upward bias: the median of $k_T/\sqrt{k_+k_-}$ over the retained heads of each model lies between 1.02 and 1.15.

Observation 9. The singular values of T and the absolute values of the eigenvalues of S are strongly correlated within heads. Sorting both lists in decreasing order, the median of their correlation over all heads of each model lies between 0.986 and 0.997.

4. Why heads are balanced

The most striking observation from the eigenvalue experiments is that *the symmetric part of every attention head has equally many positive and negative eigenvalues* (Observation 1). This is explained by the following theorem. Let H be a real vector space of dimension n equipped with an inner product $\langle\cdot,\cdot\rangle$. For linear maps $Q, K\colon C \to H$, let $A\colon C^\ell \to C^\ell$ denote the length ℓ attention head on C constructed from Q and K as query and key maps. Call the head A *nondegenerate* if the stacked map

$$\begin{pmatrix} Q \\ K \end{pmatrix} \colon C \to H \oplus H, \qquad x \mapsto (Qx, Kx)$$

is surjective.

Theorem 3. *Let A be a head in an attention block with query and key maps Q, K. If A is nondegenerate, then the symmetric bilinear form*

$$S(x,y) = \tfrac{1}{2}\big(\langle Kx, Qy\rangle + \langle Qx, Ky\rangle\big)$$

has signature

$$(n_+, n_-, n_0) = (n,\ n,\ N-2n).$$

To prove the theorem we start with an observation about symmetric parts of low-rank matrices.

Lemma 3. *Let $L \in M_N(\mathbb{R})$ with $\operatorname{rank} L \le n$, and let (n_+, n_-, n_0) be the signature of its symmetric part $S = \frac{1}{2}(L + L^T)$. Then $n_+ \le n$ and $n_- \le n$.*

Proof. Let $K = \ker L$, of dimension at least $N - n$. For $x \in K$, $2\,x^T S x = x^T L x + x^T L^T x = 0$, so the quadratic form of S vanishes on K. Let V_+ be the span of the eigenvectors of S with positive eigenvalues, of dimension n_+. Then $V_+ \cap K = 0$: a nonzero vector x of the intersection would satisfy both $x^T S x > 0$ and $x^T S x = 0$. Hence $n_+ + (N - n) \le N$, that is, $n_+ \le n$; the same argument with the negative eigenvectors gives $n_- \le n$. □

Remark 2. Low rank alone does not force balance: a positive semidefinite symmetric L of rank n has $S = L$ of signature $(n, 0, N-n)$. The nondegeneracy hypothesis of Theorem 3 is what forces S to have rank $2n$.

Proof of Theorem 3. The Gram matrix of $2S$ is $K^TQ + Q^TK$, the symmetric part of $2\,K^TQ$; since K^TQ factors through H, it has rank at most n. By Lemma 3 it therefore suffices to show that S has rank $2n$. Let $E = H \oplus H$ with the symmetric bilinear form $B\big((a,b),(a',b')\big) = \langle a, b'\rangle + \langle b, a'\rangle$, which is nondegenerate, and let $J\colon C \to E$ be the stacked map, so that $B(Jx, Jy) = 2S(x,y)$. If $Jx = 0$ then $S(x,\cdot) = 0$; conversely, if $S(x,\cdot) = 0$ then $B(Jx,\cdot)$ vanishes on $JC = E$, so $Jx = 0$ by nondegeneracy of B. The radical of S is therefore $\ker J$, of dimension $N - 2n$ by surjectivity, and S has rank $2n$. □

4.0.1. *Reconciling Migliarini's count with Theorem 3.* Migliarini [5] studies one of the heads in `gpt2`, reaching a conclusion which is apparently in conflict with our Theorem 3. The matrix Migliarini calls `W_QK` is the operator

$$W_{QK} := Q^T\,K \in \mathrm{End}(C).$$

Its symmetrization, which Migliarini calls `W_sym`, is the matrix

$$W_{\mathrm{sym}} := \tfrac{1}{2}\big(W_{QK} + W_{QK}^T\big).$$

Migliarini reports that "33 of 64 eigenvalues" of W_{sym} are negative. In fact Theorem 3 implies that W_{sym} has precisely 64 positive and 64 negative eigenvalues. Migliarini's count is actually the negative count among the subset of the largest 64 eigenvalues ordered by magnitude.

5. Trained bilinear forms in attention heads

In this section we attempt to understand what types of bilinear forms arise in trained language models. A different interpretation based on moments is described in Appendix C. We introduce a map on real bilinear forms valued in the 3-simplex:

$$\pi\colon M_N(\mathbb{R}) \smallsetminus \{0\} \longrightarrow \Delta_3, \qquad L \longmapsto (a,b,c,d).$$

The map π is invariant under positive scaling and orthogonal congruence $L \mapsto Q^TLQ$.

5.1. The profile of a real bilinear form

Let $L \in M_N(\mathbb{R})$ with $\|L\| = 1$, and write $L = S+T$ with S symmetric and T antisymmetric. Let λ_+ (resp. λ_-) list the positive eigenvalues (resp. the absolute values of the negative eigenvalues) of S in decreasing order, appending zeros so that the two lists have equal length. For a real vector $x \in \mathbb{R}^m$ let $x_+ = \max(x,0)$ and $x_- = \max(-x,0)$ applied entry-wise.

Definition 6. The *profile* of L is $\pi(L) = (a,b,c,d)$, where

$$a = \|T\|^2, \qquad b = 2\,\langle \lambda_+, \lambda_-\rangle, \qquad c = \|(\lambda_+ - \lambda_-)_+\|^2, \qquad d = \|(\lambda_+ - \lambda_-)_-\|^2.$$

For nonzero $L \in M_N(\mathbb{R})$ we set $\pi(L) = \pi(L/\|L\|)$.

The components of the profile are nonnegative and satisfy $a + b + c + d = 1$. The facet $\{a = 0\}$ is the triangle $\Delta = \{(b,c,d) \in \mathbb{R}^2 : b,c,d \geq 0,\ b + c + d = 1\}$ containing the profiles of symmetric forms.

Theorem 4.

(1) $a = 0$ if and only if L is symmetric, and $a = 1$ if and only if L is antisymmetric;
(2) $b = 0$ if and only if S is semidefinite, and $b = 1$ if and only if L is symmetric with $\lambda_+ = \lambda_-$;
(3) $d = 0$ if and only if $S = H + P$ for symmetric matrices H and P such that the spectrum of H is symmetric about zero and P is positive semidefinite; equivalently, $(\lambda_+)_j \geq (\lambda_-)_j$ for every j. In that case H and P may be chosen to commute with S and with each other, with $\|P\|^2 = c\,\|L\|^2$. Moreover, $d = 1$ if and only if L is symmetric negative semidefinite;
(4) $c = 0$ if and only if $S = H - P$ with H as in (3) and P positive semidefinite; equivalently, $(\lambda_-)_j \geq (\lambda_+)_j$ for every j; and then $\|P\|^2 = d\,\|L\|^2$. Moreover, $c = 1$ if and only if L is symmetric positive semidefinite.

Proof. See Appendix B. □

Lemma 4. *Let $L \in M_N(\mathbb{R})$ with symmetric part $S \neq 0$, and write $\pi(L) = (a,b,c,d)$ and $\pi(S) = (0, b_S, c_S, d_S)$. Then $(b,c,d) = (1-a)\,(b_S, c_S, d_S)$.*

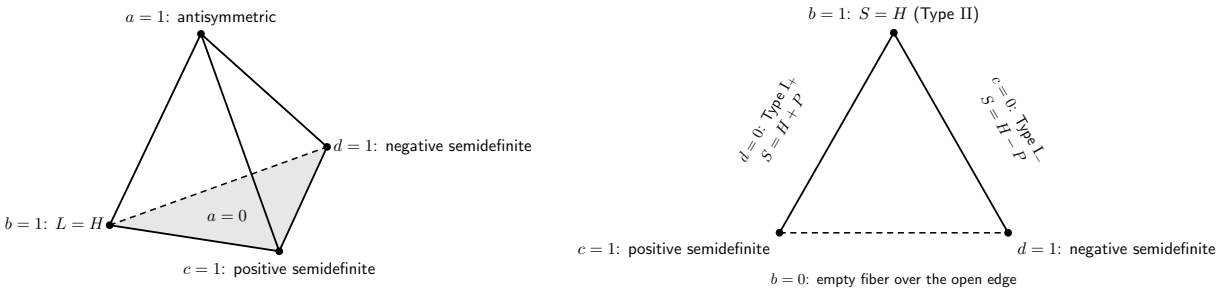

FIGURE 7. The simplex Δ_3 of profiles of bilinear forms (left) and its face Δ of symmetric forms (right). Here H has spectrum symmetric about zero and P is positive semidefinite. The fiber over the dashed edge $\{b = 0\}$ is empty away from its endpoints.

	face	fiber of π
vertex	$a = 1$	antisymmetric forms
vertex	$b = 1$	symmetric forms with $\lambda_+ = \lambda_-$
vertex	$c = 1$	positive semidefinite forms
vertex	$d = 1$	negative semidefinite forms
edge	$\{c = d = 0\}$	forms with $\lambda_+ = \lambda_-$
edge	$\{b = d = 0\}$	forms with S positive semidefinite
edge	$\{b = c = 0\}$	forms with S negative semidefinite
edge	$\{a = d = 0\}$	symmetric forms $H + P$
edge	$\{a = c = 0\}$	symmetric forms $H - P$
edge	$\{a = b = 0\}$	empty away from the endpoints
facet	$\Delta = \{a = 0\}$	symmetric forms
facet	$\{b = 0\}$	forms with S semidefinite; relative interior empty
facet	$\{d = 0\}$	forms with $S = H + P$
facet	$\{c = 0\}$	forms with $S = H - P$
interior		$T \neq 0$ and S of neither form $H + P$ nor $H - P$

TABLE 3. The fibers of π over the faces of Δ_3; H denotes a symmetric matrix with spectrum symmetric about zero and P a positive semidefinite matrix.

Proof. The three quantities b, c, d are computed from the eigenvalues of S and divided by $\|L\|^2$, whereas b_S, c_S, d_S are the same quantities divided by $\|S\|^2$. The claim follows from $\|S\|^2/\|L\|^2 = 1 - \|T\|^2/\|L\|^2 = 1 - a$. □

Proposition 4. *Let L be a rank-one matrix of unit norm. Then*

$$\pi(L) = \left(\frac{1-t^2}{2}, \frac{1-t^2}{2}, t_+^2, t_-^2\right), \qquad \textit{where } t = \operatorname{tr}(L).$$

The set of profiles of rank-one matrices is equal to $\Theta = \{a = b,\, cd = 0\} \subset \Delta_3$.

Proof. Write $L = uv^T$ for unit vectors u, v. If u and v are orthogonal then L is nilpotent and has no nonzero eigenvalues, and otherwise the eigenspace of the nonzero eigenvalue is spanned by u and the nonzero eigenvalue is $\langle u, v\rangle$. Thus $t = \langle u, v\rangle$.

We bring u to the first standard basis vector by an orthogonal change of coordinates. If $u = \pm v$ then the top-left entry of L is ± 1 with all other entries zero. If $u = v$ then $\lambda_+ = (1, 0, \ldots, 0)$, $\lambda_- = 0$, and $\pi(L) = (0, 0, 1, 0)$. If $u = -v$ then $\lambda_+ = 0$, $\lambda_- = (1, 0, \ldots, 0)$, and $\pi(L) = (0, 0, 0, 1)$. This verifies the formula when $t = \pm 1$.

Suppose $|t| < 1$. Define the vectors

$$e = \frac{u+v}{\sqrt{2+2t}}, \qquad f = \frac{v-u}{\sqrt{2-2t}}.$$

These form an orthonormal set. We may extend this to an orthonormal basis in which L has top-left block equal to

$$\frac{1}{2}\begin{pmatrix} 1+t & \sqrt{1-t^2} \\ -\sqrt{1-t^2} & t-1 \end{pmatrix}$$

and elsewhere is zero. This block is equal to

$$S+T = \frac{1}{2}\begin{pmatrix} 1+t & 0 \\ 0 & t-1 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 0 & \sqrt{1-t^2} \\ -\sqrt{1-t^2} & 0 \end{pmatrix}.$$

From this we find $\operatorname{tr}(T^T T) = \frac{1}{2}(1-t^2)$, $\lambda_+ = \frac{1}{2}(1+t, 0)$, $\lambda_- = \frac{1}{2}(1-t, 0)$, and the rest follows easily. □

5.2. **Profiles of trained attention heads**

In this section we study the attention heads of the fourteen models with the profile map. To attempt a partial explanation for Observation 2, we consider an ensemble of large symmetric matrices whose positive and negative eigenvalue distributions have the same shape up to rescaling.

Theorem 5. *Let X be a positive random variable with finite, nonzero second moment. Fix $u, v > 0$ and set $\rho = u/v$. For each n let $N_n \geq 2n$ and let $S_n \in M_{N_n}(\mathbb{R})$ be a symmetric matrix with positive eigenvalues $vX_1, \ldots, vX_n$ and negative eigenvalues $-uX_1', \ldots, -uX_n'$, where the X_i and X_i' are independent copies of X. Then the profile $\pi(S_n)$ converges almost surely to*

$$\frac{1}{1+\rho^2}\big(0,\ 2\rho,\ (1-\rho)_+^2,\ (\rho-1)_+^2\big) \in p(\Theta). \tag{3}$$

As ρ ranges over $(0,1]$ the limit traces the edge $\{d=0\}$ of Δ from $(0,1,0)$ to $(1,0,0)$, and as ρ ranges over $[1,\infty)$ it traces the edge $\{c=0\}$ from $(1,0,0)$ to $(0,0,1)$.

For exactly proportional lobes, $\lambda_- = \rho\,\lambda_+$, direct substitution in the profile definition gives (3). The proof consists of showing that for randomly sampled matrices, this formula still holds in the high-dimensional limit.

Proof. Let α_n and β_n denote the positive eigenvalues of S_n and the absolute values of its negative eigenvalues, each listed in decreasing order, so that $\lambda_+ = \alpha_n$ and $\lambda_- = \beta_n$. By Lemma 6 the sorted lobes become proportional in the limit: $\|\beta_n - \rho\,\alpha_n\|/\|\alpha_n\| \to 0$ almost surely. Write $\beta_n = \rho\,\alpha_n + \varepsilon_n$ with $\delta_n = \|\varepsilon_n\|/\|\alpha_n\| \to 0$. By the Cauchy–Schwarz and triangle inequalities,

$$\left|\frac{\langle \alpha_n, \beta_n\rangle}{\|\alpha_n\|^2} - \rho\right| \leq \delta_n, \qquad \left|\frac{\|\beta_n\|^2}{\|\alpha_n\|^2} - \rho^2\right| \leq \delta_n^2 + 2\rho\,\delta_n,$$

and, since $x \mapsto x_\pm$ is 1-Lipschitz in each entry and $\big((1-\rho)\,\alpha_n\big)_\pm = (1-\rho)_\pm\,\alpha_n$,

$$\left|\frac{\|(\alpha_n - \beta_n)_\pm\|}{\|\alpha_n\|} - (1-\rho)_\pm\right| \leq \frac{\|(\alpha_n - \beta_n) - (1-\rho)\,\alpha_n\|}{\|\alpha_n\|} = \delta_n.$$

Since $\|S_n\|^2 = \|\alpha_n\|^2 + \|\beta_n\|^2$, dividing each of $\|S_n\|^2$, $2\langle \alpha_n, \beta_n\rangle$ and $\|(\alpha_n - \beta_n)_\pm\|^2$ by $\|\alpha_n\|^2$ gives quantities that converge almost surely to $1+\rho^2$, 2ρ and $(1-\rho)_\pm^2$ respectively. The profile map is homogeneous of degree zero in S_n, so

$$b = \frac{2\langle \alpha_n, \beta_n\rangle}{\|S_n\|^2} = \frac{2\langle \alpha_n, \beta_n\rangle/\|\alpha_n\|^2}{\|S_n\|^2/\|\alpha_n\|^2},$$

and likewise for c and d; the factor $\|\alpha_n\|^2$ cancels, and each ratio converges almost surely to the corresponding entry of (3). □

5.2.1. *The hypothesis of Theorem 5.* Observation 5 quantifies the approximate proportionality of the two sorted spectral lobes at trained heads. Figure 10 illustrates the theorem at $\rho = \frac{1}{2}$, where the limiting profiles of S_n are $(b, c, d) = (0.8, 0.2, 0)$.

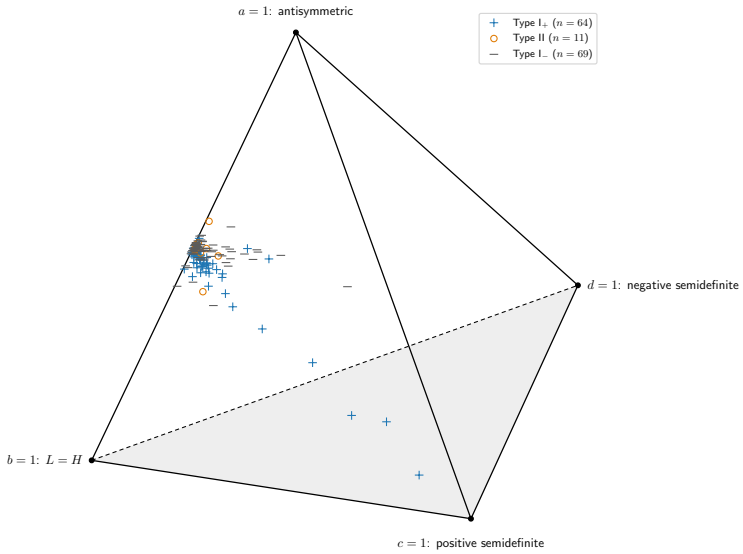

FIGURE 8. Profiles of the heads of `gpt2` on the simplex Δ_3.

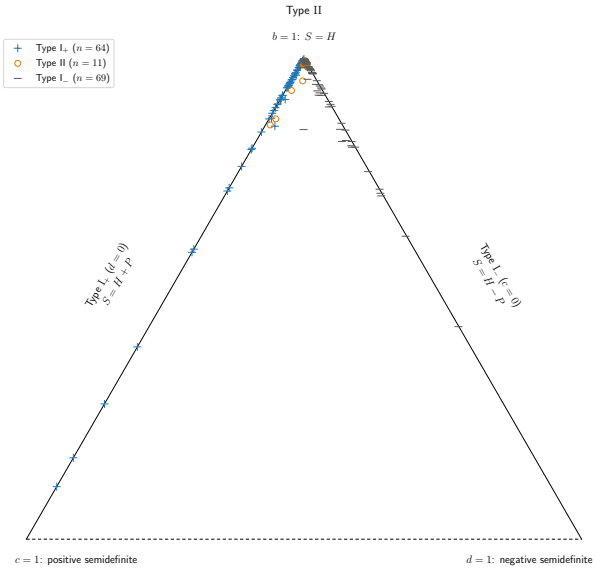

FIGURE 9. Profiles of the symmetric parts of the heads of `gpt2` on Δ. Here H has spectrum symmetric about zero and P is positive semidefinite.

5.2.2. *Effective rank of trained bilinear forms.* Let $\sigma_1 \geq \sigma_2 \geq \cdots$ be the singular values of L. There are many measures of effective rank used in the literature. One of them is the *participation ratio* defined by $(\sum_i \sigma_i^2)^2 / \sum_i \sigma_i^4$. This equals 1 for a rank-one form and n for a flat spectrum. For the heads of the fourteen models, the participation ratio has median 48.5 and lies below 2 for fewer than 1% of them.

6. DISCUSSION

We have presented several observations on trained attention heads showing that their bilinear forms have distinguished properties. It is important to discern which properties are learned and which are structural. Equal counts for positive and negative eigenvalues is a structural consequence of the low-rank query-key attention head construction rather than a pattern learned from language. Training shapes the distributions of these eigenvalues: across the models studied, the symmetric part is stronger than a random baseline, and

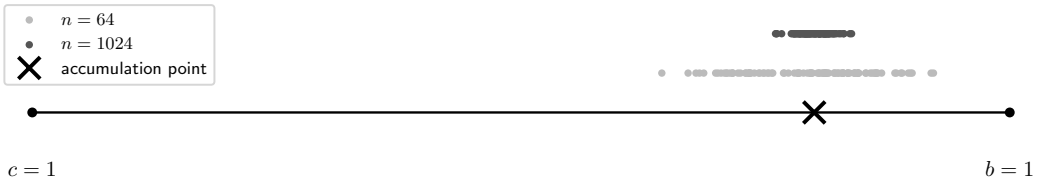

FIGURE 10. Profiles of 100 spectra sampled from $\frac{1}{2}(\gamma_{\mu,k} + \gamma_{\nu,k})$ with $\mu = -\frac{1}{2}$, $\nu = 1$, $k = 2$, for $n = 64$ and $n = 1024$ eigenvalues per side, plotted by their coordinate b along the edge $\{d = 0\}$, drawn horizontally from the vertex $c = 1$ to the vertex $b = 1$, together with the limit point $b = 0.8$ of Theorem 5 ($\times$).

its positive and negative eigenvalues follow closely related distributions. We have proposed metrics using the classification of real bilinear forms which describe variation between trained heads. They are useful metrics for head classification, although they do not by themselves identify what a head computes in an explicit mechanistic sense.

Several observations remain unexplained. In particular, we do not know why the eigenvalues of the symmetric part and the singular values of the antisymmetric part have gamma-like profiles, or why their fitted parameters and magnitudes are so closely related. These patterns occur across all fourteen models studied, but the present results do not determine whether they arise from the parametrization, optimization, training data, or an interaction among them. The curious observation that profiles of trained bilinear forms accumulate near the family of profiles of rank-one forms also remains only partially explained.

The empirical scope is also limited. We examine the bilinear forms in the heads of fourteen decoder-only models. However an attention head acts within a larger system: its output is combined with other heads and then processed by the neural network in each transformer block. Future work should examine how the symmetric and antisymmetric parts and their spectral distributions interact with these components.

Observation	Status	Source
1: the symmetric part is balanced	Proved: Theorem 3	
2: head profiles accumulate near rank-one profiles $\Theta \subset \Delta_3$	Partially explained: Theorem 5; the proportionality of the lobes in its proof holds approximately at the heads (§5.2)	
3: $\lambda(S)$ has two gamma-like lobes	Unexplained	
4: symmetric bias of attention heads	Partially explained: [10, Theorem 2.4] for the bidirectional case with assumptions, unexplained for decoder-only	[10]
5: the sorted spectral lobes are approximately proportional	Linked to profile accumulation by Theorem 5; see §5.2.1	
6: $\lvert\lambda(L)\rvert$ and $\sigma(T)$ have gamma-like profiles	Unexplained	[3]
7: $\tau \approx \sqrt{\lvert\mu\rvert\nu}$	Unexplained	
8: $k_T \approx \sqrt{k_+ k_-}$	Unexplained	
9: sorted $\sigma(T)$ and $\lvert\lambda(S)\rvert$ are strongly correlated	Unexplained	

ACKNOWLEDGEMENTS

The ideas underlying Theorem 3 and Proposition 3 arose out of extended conversations with ChatGPT 5.5. The observation that Θ is the set of profiles of rank-one forms was made by ChatGPT 6. Theorem 5 and its proof, as well as the constructions in Appendix C, arose out of extended conversations with Claude Fable 5. The appendices (with the exception of Appendix C), figures, and experimental code, as well as the statements

and proofs of Proposition 2 and Lemma 2 were generated by Claude Fable 5. The mathematical assertions in the appendices were formalized in Lean with the help of Claude Fable 5. The author is responsible for the final text and released code.

Supplementary information

All experimental code, Lean formalizations, as well as additional data visualizations are publicly available.[4]

Appendix A. Random matrices

This appendix collects the symmetric/skew energy identities and the expected energy splits for random matrices used in §3 to justify the random-baseline value $\mathscr{S} \approx 1/2$ in Figure 13.

Lemma 5 (Symmetric and skew energy identities; cf. [10, Lemma S1.14 and Eq. (S118)])**.** *Let $L \in M_N(\mathbb{R})$, and write $S = \frac{1}{2}(L + L^T)$, $T = \frac{1}{2}(L - L^T)$. Then $\|L\|^2 = \|S\|^2 + \|T\|^2$ and $\|S\|^2 = \frac{1}{2}\|L\|^2 + \frac{1}{2}\operatorname{tr}(L^2)$, $\|T\|^2 = \frac{1}{2}\|L\|^2 - \frac{1}{2}\operatorname{tr}(L^2)$. Consequently,*

$$\frac{\|S\|^2}{\|L\|^2} = \frac{1}{2} + \frac{\operatorname{tr}(L^2)}{2\|L\|^2}, \qquad \frac{\|T\|^2}{\|L\|^2} = \frac{1}{2} - \frac{\operatorname{tr}(L^2)}{2\|L\|^2}.$$

Proof. The subspaces of symmetric and antisymmetric matrices are orthogonal with respect to the Frobenius inner product $\langle X, Y\rangle = \operatorname{tr}(XY^T)$: for S symmetric and T antisymmetric,

$$\langle S, T\rangle = \operatorname{tr}(ST^T) = -\operatorname{tr}(ST) = -\operatorname{tr}\big((ST)^T\big) = -\operatorname{tr}(T^T S^T) = \operatorname{tr}(TS) = \operatorname{tr}(ST),$$

so $\operatorname{tr}(ST) = 0$ and $\langle S, T\rangle = 0$. Since $L = S + T$, it follows that $\|L\|^2 = \|S\|^2 + \|T\|^2$. Expanding,

$$\|S\|^2 = \tfrac{1}{4}\big(\langle L, L\rangle + 2\langle L, L^T\rangle + \langle L^T, L^T\rangle\big).$$

Since $\langle L, L\rangle = \langle L^T, L^T\rangle = \|L\|^2$ and $\langle L, L^T\rangle = \operatorname{tr}(L^2)$, we obtain $\|S\|^2 = \frac{1}{2}\|L\|^2 + \frac{1}{2}\operatorname{tr}(L^2)$. The formula for $\|T\|^2$ follows by subtraction from $\|L\|^2$. □

Corollary 1 (Iid square random matrix baseline)**.** *Let $L \in M_N(\mathbb{R})$ be a random matrix whose entries are independent, mean-zero, with common variance σ^2. Then*

$$\mathbb{E}\|S\|^2 = \frac{N(N+1)}{2}\sigma^2, \qquad \mathbb{E}\|T\|^2 = \frac{N(N-1)}{2}\sigma^2,$$

so

$$\frac{\mathbb{E}\|S\|^2}{\mathbb{E}\|L\|^2} = \frac{N+1}{2N}, \qquad \frac{\mathbb{E}\|T\|^2}{\mathbb{E}\|L\|^2} = \frac{N-1}{2N}.$$

If the entries of L are iid Gaussian, then also $\mathbb{E}\big[\|S\|^2/\|L\|^2\big] = (N+1)/(2N)$ and $\mathbb{E}\big[\|T\|^2/\|L\|^2\big] = (N-1)/(2N)$.

Proof. For $i < j$, $S_{ij} = S_{ji} = \frac{1}{2}(L_{ij} + L_{ji})$ with L_{ij}, L_{ji} independent of variance σ^2, so $\mathbb{E}S_{ij}^2 = \frac{1}{2}\sigma^2$. Each off-diagonal pair contributes σ^2 in expectation; there are $N(N-1)/2$ such pairs. The diagonal contribution is $\sum_i \mathbb{E}S_{ii}^2 = \sum_i \mathbb{E}L_{ii}^2 = N\sigma^2$. Hence $\mathbb{E}\|S\|^2 = N\sigma^2 + N(N-1)/2 \cdot \sigma^2 = N(N+1)\sigma^2/2$. The skew case is similar with no diagonal contribution.

For the iid Gaussian case, L is an isotropic Gaussian vector in the N^2-dimensional inner-product space $M_N(\mathbb{R})$. The symmetric and antisymmetric subspaces are orthogonal of dimensions $N(N+1)/2$ and $N(N-1)/2$ respectively, so the expected fraction of squared norm in each subspace equals its dimension divided by N^2. □

Corollary 2 (Random QK-product baseline)**.** *Let $K, Q \in M_{n\times N}(\mathbb{R})$ be independent random matrices with independent, mean-zero entries of variances σ_K^2 and σ_Q^2. Set $L = K^T Q$ and S, T as before. Then*

$$\mathbb{E}\|L\|^2 = nN^2\sigma_K^2\sigma_Q^2, \qquad \mathbb{E}\|S\|^2 = \frac{nN(N+1)}{2}\sigma_K^2\sigma_Q^2,$$

$$\mathbb{E}\|T\|^2 = \frac{nN(N-1)}{2}\sigma_K^2\sigma_Q^2,$$

[4] https://aodesky.github.io/attention-bilinear-forms/

so

$$\frac{\mathbb{E}\|S\|^2}{\mathbb{E}\|L\|^2} = \frac{N+1}{2N}, \qquad \frac{\mathbb{E}\|T\|^2}{\mathbb{E}\|L\|^2} = \frac{N-1}{2N}.$$

Proof. Writing $L_{ab} = \sum_{r=1}^{n} (K)_{ra}(Q)_{rb}$ and using independence, $\mathbb{E}L_{ab} = 0$ and $\mathbb{E}L_{ab}^2 = n\sigma_K^2\sigma_Q^2$. Summing over $a, b \in \{1, \dots, N\}$ gives $\mathbb{E}\|L\|^2 = nN^2\sigma_K^2\sigma_Q^2$.

By Lemma 5, $\mathbb{E}\|S\|^2 = \frac{1}{2}\mathbb{E}\|L\|^2 + \frac{1}{2}\mathbb{E}\,\mathrm{tr}(L^2)$. Expanding $\mathrm{tr}(L^2) = \sum_{a,b} L_{ab}L_{ba}$: for $a \neq b$, $\mathbb{E}(L_{ab}L_{ba}) = 0$ by independence and centering; for $a = b$, $\mathbb{E}L_{aa}^2 = n\sigma_K^2\sigma_Q^2$. Hence $\mathbb{E}\,\mathrm{tr}(L^2) = nN\sigma_K^2\sigma_Q^2$, and $\mathbb{E}\|S\|^2 = nN(N+1)\sigma_K^2\sigma_Q^2/2$. The skew formula follows by subtraction. □

The preceding corollary computes the ratio of expectations $\mathbb{E}\|S\|^2/\mathbb{E}\|L\|^2$. Under the additional hypothesis that the entries are Gaussian, the same value is also the expectation of the normalized symmetric energy — and the low-rank head-projection constraint $\mathrm{rank}(L) \leq n$ does not change this expectation.

Proposition 5 (Expected symmetric energy for random low-rank QK products)**.** *Let $K, Q \in M_{n\times N}(\mathbb{R})$ be independent random matrices with iid centered Gaussian entries, where $1 \leq n \leq N$. Set $L = K^TQ \in M_N(\mathbb{R})$, $S = \frac{1}{2}(L + L^T)$, $T = \frac{1}{2}(L - L^T)$, and define $\mathscr{S} = \frac{\|S\|^2}{\|L\|^2}, \mathscr{A} = \frac{\|T\|^2}{\|L\|^2}$. Then*

$$\mathbb{E}[\mathscr{S}] = \frac{N+1}{2N}, \qquad \mathbb{E}[\mathscr{A}] = \frac{N-1}{2N},$$

independently of the head dimension n.

Proof. By Lemma 5, $\mathscr{S} = \frac{1}{2} + \frac{\mathrm{tr}(L^2)}{2\|L\|^2}$, so it suffices to show that $\mathbb{E}[\mathrm{tr}(L^2)/\|L\|^2] = 1/N$. Since K and Q have iid Gaussian entries, their joint law is invariant under right multiplication of either factor by any $V \in O_N(\mathbb{R})$. Therefore the law of $L = K^TQ$ is invariant under $L \longmapsto U^TLV$ for every $U, V \in O_N(\mathbb{R})$. Set $X = L/\|L\|$, so $\|X\| = 1$ almost surely. By permutation invariance of the row and column indices, the expectations $\mathbb{E}[X_{ij}^2]$ are all equal, and since $\sum_{i,j} X_{ij}^2 = 1$, $\mathbb{E}[X_{ij}^2] = \frac{1}{N^2}$ for every i, j. For $i \neq j$, changing the sign of the i-th column of K flips the i-th row of $L = K^TQ$, hence flips X_{ij} while leaving X_{ji} unchanged; this preserves the law of X, so $\mathbb{E}[X_{ij}X_{ji}] = 0$. Hence

$$\mathbb{E}\left[\frac{\mathrm{tr}(L^2)}{\|L\|^2}\right] = \mathbb{E}[\mathrm{tr}(X^2)] = \sum_{i,j} \mathbb{E}[X_{ij}X_{ji}] = \sum_{i=1}^{N} \mathbb{E}[X_{ii}^2] = \frac{1}{N},$$

and therefore $\mathbb{E}[\mathscr{S}] = \frac{1}{2} + \frac{1}{2N} = (N+1)/(2N)$. Since $\mathscr{A} = 1 - \mathscr{S}$, we get $\mathbb{E}[\mathscr{A}] = (N-1)/(2N)$. □

Lemma 6. *Let X be a positive random variable with finite, nonzero second moment. Fix scales $u, v > 0$ and set $\rho = u/v$. For each n, let $\alpha_n \in \mathbb{R}^n$ be the decreasing sorting of $vX_1, \dots, vX_n$ and let $\beta_n \in \mathbb{R}^n$ be the decreasing sorting of $uX_1', \dots, uX_n'$, where the X_i and X_i' are independent copies of X. Then the sorted lobes become proportional with ratio ρ:*

$$\frac{\|\beta_n - \rho\,\alpha_n\|}{\|\alpha_n\|} \longrightarrow 0 \qquad \textit{almost surely as } n \to \infty.$$

Proof of Lemma 6. We may write $\beta_n = \rho\,\alpha_n'$, where α_n' is the decreasing sorting of $vX_1', \dots, vX_n'$. Since the claim is unchanged when both lists are divided by v, we may assume $v = 1$. Let $m_2 = \mathbf{E}[X^2]$, which is finite and nonzero by assumption.

Let $\overline{F}(x) = \mathbf{P}(X > x)$, and let $G\colon (0,1) \to (0,\infty)$ be the decreasing generalized inverse $G(t) = \inf\{x > 0 : \overline{F}(x) \leq t\}$, so that $G(t) > x$ if and only if $t < \overline{F}(x)$. Let U be a random variable with $\mathbf{P}(U < t) = t$ for $t \in [0,1]$. Then $\mathbf{P}(G(U) > x) = \mathbf{P}(U < \overline{F}(x)) = \overline{F}(x)$, so $G(U)$ has the law of X and

$$\int_0^1 G(t)^2\,dt = m_2.$$

Given n independent samples $X_1, \dots, X_n$ of X with decreasing sorting $x_1 \geq \cdots \geq x_n$, define the step function $h_n(t) = x_{\lceil nt \rceil}$ on $(0,1)$, so that $\int_0^1 h_n^2 = \frac{1}{n}\sum_i x_i^2$. We claim that $h_n \to G$ in $L^2(0,1)$ almost surely, and prove it in three steps.

First, $h_n(t) \to G(t)$ almost surely for each fixed t at which G is continuous, hence for all but countably many $t \in (0,1)$. Let $\overline{F}_n(x) = n^{-1}\#\{i : X_i > x\}$, so that $h_n(t) > x$ exactly when $\overline{F}_n(x) \geq \lceil nt \rceil/n$. Fix $\varepsilon > 0$. Since G is continuous at t we have $\overline{F}(G(t) + \varepsilon) < t$, so by the strong law of large numbers $\overline{F}_n(G(t) + \varepsilon) \to$

$\overline{F}(G(t)+\varepsilon) < t$ almost surely; eventually $\overline{F}_n(G(t)+\varepsilon) < t \le \lceil nt \rceil/n$ and hence $h_n(t) \le G(t)+\varepsilon$. The lower bound is symmetric, using $\overline{F}(G(t)-\varepsilon) > t$.

Second, by the strong law of large numbers,

$$\int_0^1 h_n(t)^2\,dt = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \longrightarrow m_2 = \int_0^1 G(t)^2\,dt \qquad \text{almost surely.}$$

Third, expanding $\int (h_n - G)^2 = \int h_n^2 + \int G^2 - 2\int h_n G$ and applying Fatou's lemma to the nonnegative functions $h_n G$, which converge to G^2 almost everywhere by the first step and so give $\liminf \int h_n G \ge \int G^2$, we conclude

$$\limsup_n \int_0^1 (h_n - G)^2\,dt \;\le\; m_2 + m_2 - 2m_2 = 0,$$

which proves the claim.

Now let g_n and g_n' be the step functions of the two samples defining α_n and α_n'. Then

$$\frac{\|\alpha_n - \alpha_n'\|^2}{n} = \int_0^1 (g_n - g_n')^2 \le 2\int_0^1 (g_n - G)^2 + 2\int_0^1 (g_n' - G)^2 \longrightarrow 0, \qquad \frac{\|\alpha_n\|^2}{n} = \int_0^1 g_n^2 \longrightarrow m_2 > 0,$$

almost surely, and therefore $\|\beta_n - \rho\,\alpha_n\|/\|\alpha_n\| = \rho\,\|\alpha_n' - \alpha_n\|/\|\alpha_n\| \to 0$. □

Appendix B. Balanced bimodal spectra

This appendix collects identities for the normalized spectra studied in §3 and §5. All matrix norms below are Frobenius norms, $\|L\| = \sqrt{\operatorname{tr}(L^T L)}$.

Fix $m \ge 1$ and set $n = 2m$. Let Λ_m denote the set of unit vectors $\lambda = (\lambda_1 \le \cdots \le \lambda_n) \in \mathbb{R}^n$ with $\lambda_m < 0 < \lambda_{m+1}$. For $\lambda \in \Lambda_m$ define $\lambda_+, \lambda_- \in \mathbb{R}^m$ by

$$(\lambda_+)_j = \lambda_{n+1-j}, \qquad (\lambda_-)_j = -\lambda_j, \qquad j = 1, \dots, m,$$

so that λ_+ and λ_- list the positive entries and the absolute values of the negative entries of λ, both in decreasing order. Define

$$\mathscr{I}(\lambda) = \|\lambda_+\|^2 - \|\lambda_-\|^2, \qquad C(\lambda) = \langle \lambda_+, \lambda_- \rangle, \qquad \mathscr{P}(\lambda) = 1 - \|\lambda_+ - \lambda_-\|.$$

Finally, define the linear map $\iota\colon \mathbb{R}^n \to \mathbb{R}^n$ by $(\iota\lambda)_i = -\lambda_{n+1-i}$.

Lemma 7. *For $\lambda \in \Lambda_m$, abbreviating $\mathscr{I} = \mathscr{I}(\lambda)$ and $C = C(\lambda)$:*

(1) $\|\lambda_+\|^2 = \frac{1}{2}(1+\mathscr{I})$ and $\|\lambda_-\|^2 = \frac{1}{2}(1-\mathscr{I})$;
(2) $2\,\|\lambda_+\|\,\|\lambda_-\| = \sqrt{1-\mathscr{I}^2}$;
(3) $\|\lambda_+ - \lambda_-\|^2 = 1 - 2C$, and hence $\mathscr{P}(\lambda) = 1 - \sqrt{1-2C}$;
(4) $\sum_{j=1}^{m} \big((\lambda_+)_j + i(\lambda_-)_j\big)^2 = \mathscr{I} + 2iC$.

Proof. Since the entries of λ_+ and λ_- are the absolute values of the entries of λ, $\|\lambda_+\|^2 + \|\lambda_-\|^2 = \|\lambda\|^2 = 1$; combining with the definition of $\mathscr{I}$ gives (1), and (2) follows by multiplying the two expressions in (1). For (3), expand $\|\lambda_+ - \lambda_-\|^2 = \|\lambda_+\|^2 + \|\lambda_-\|^2 - 2\langle \lambda_+, \lambda_- \rangle = 1 - 2C$. For (4), expand the square: the real part is $\|\lambda_+\|^2 - \|\lambda_-\|^2 = \mathscr{I}$ and the imaginary part is $2\langle \lambda_+, \lambda_- \rangle = 2C$. □

Note that $0 \le 2C \le 1$ ($C \ge 0$ since the entries of $\lambda_\pm$ are nonnegative, and $2C \le 1$ by part (3)), so $0 \le \mathscr{P} \le 1$.

Lemma 8. *The map ι is an orthogonal, symmetric involution of $\mathbb{R}^n$ and maps Λ_m to Λ_m, with $\lambda_+(\iota\lambda) = \lambda_-(\lambda)$ and $\lambda_-(\iota\lambda) = \lambda_+(\lambda)$. Consequently $\mathscr{I} \circ \iota = -\mathscr{I}$, $C \circ \iota = C$, and $\mathscr{P} \circ \iota = \mathscr{P}$. Moreover, for $\lambda \in \Lambda_m$,*

$$\|\lambda - \iota\lambda\|^2 = 2\,\|\lambda_+ - \lambda_-\|^2, \qquad \text{so} \qquad \mathscr{P}(\lambda) = 1 - \tfrac{1}{\sqrt{2}}\|\lambda - \iota\lambda\|;$$

in particular $\mathscr{P}(\lambda) = 1$ if and only if $\iota\lambda = \lambda$.

Proof. The matrix of ι has entries $-\delta_{j,\,n+1-i}$, which is symmetric and orthogonal, and $\iota^2 = \mathrm{id}$. If λ is increasing then so is $\iota\lambda$, and the entries of $\iota\lambda$ form the same multiset as the entries of $-\lambda$, so $\iota(\Lambda_m) = \Lambda_m$. The positive entries of $\iota\lambda$ are $-\lambda_1 \ge \cdots \ge -\lambda_m$, whence $\lambda_+(\iota\lambda) = \lambda_-(\lambda)$; the other identity is symmetric, and the transformation rules for $\mathscr{I}$, C, $\mathscr{P}$ follow. For the displacement identity, the j-th and $(n+1-j)$-th coordinates of $\lambda - \iota\lambda$ both equal $\lambda_j + \lambda_{n+1-j} = (\lambda_+)_j - (\lambda_-)_j$ up to sign, so $\|\lambda - \iota\lambda\|^2 = 2\sum_j \big((\lambda_+)_j - (\lambda_-)_j\big)^2$. The final statement follows from Lemma 7(3) since $\mathscr{P}(\lambda) = 1$ iff $\lambda_+ = \lambda_-$. □

We prove Theorem 4, using the notation of its statement. The argument is organized by the dimension of the face: the partition of the energy first, then the facets, then the faces of lower dimension, whose fibers are intersections of the facet fibers.

Proof of Theorem 4. Rescaling, we may assume $\|L\| = 1$.

The partition. The entries of λ_+ and λ_- are the absolute values of the eigenvalues of S together with the appended zeros, so $\|\lambda_+\|^2 + \|\lambda_-\|^2 = \|S\|^2$; expanding $\|\lambda_+ - \lambda_-\|^2 = \|\lambda_+\|^2 + \|\lambda_-\|^2 - 2\langle\lambda_+, \lambda_-\rangle$ and using $\|\lambda_+ - \lambda_-\|^2 = \|(\lambda_+ - \lambda_-)_+\|^2 + \|(\lambda_+ - \lambda_-)_-\|^2$ gives

$$2\langle\lambda_+, \lambda_-\rangle + \|(\lambda_+ - \lambda_-)_+\|^2 + \|(\lambda_+ - \lambda_-)_-\|^2 = \|S\|^2.$$

Adding $a = \|T\|^2$ and using $\|S\|^2 + \|T\|^2 = \|L\|^2 = 1$ (Lemma 5) yields $a + b + c + d = 1$. Nonnegativity is clear since the entries of λ_+ and λ_- are nonnegative.

The facets. We prove the vanishing conditions of (1)–(4); the vertex conditions are proved below with the other zero-dimensional faces. If $S = 0$ then $b = c = d = 0$, S is semidefinite, and the decompositions of (3) and (4) hold with $H = P = 0$; so assume $S \neq 0$.

(1): $a = \|T\|^2$ vanishes if and only if $T = 0$.

(2): If S is semidefinite then $\lambda_+ = 0$ or $\lambda_- = 0$, so $b = 2\langle\lambda_+, \lambda_-\rangle = 0$. Conversely, if S is indefinite then $(\lambda_+)_1 > 0$ and $(\lambda_-)_1 > 0$; since the lists λ_+ and λ_- are sorted in decreasing order, $\langle\lambda_+, \lambda_-\rangle \geq (\lambda_+)_1(\lambda_-)_1 > 0$, so $b > 0$.

(3): First, $d = \|(\lambda_+ - \lambda_-)_-\|^2$ vanishes iff $(\lambda_+)_j \geq (\lambda_-)_j$ for every j, which is the stated rank condition.

Assume the rank condition. Choose orthonormal eigenvectors of S: for each j with $(\lambda_+)_j > 0$ let v_j be an eigenvector for the eigenvalue $(\lambda_+)_j$, and for each j with $(\lambda_-)_j > 0$ let w_j be an eigenvector for $-(\lambda_-)_j$, taking each eigenvalue with its multiplicity. Since $(\lambda_-)_j > 0$ implies $(\lambda_+)_j > 0$, every w_j is matched with a v_j. Define

$$H = \sum_{j\,:\,(\lambda_-)_j > 0} (\lambda_-)_j \big(v_j v_j^T - w_j w_j^T\big), \qquad P = S - H = \sum_{j\,:\,(\lambda_+)_j > 0} \big((\lambda_+)_j - (\lambda_-)_j\big)\, v_j v_j^T.$$

The spectrum of H consists of the pairs $\pm(\lambda_-)_j$ together with zeros, so it is symmetric about zero; P is positive semidefinite since $(\lambda_+)_j \geq (\lambda_-)_j$; and $\|P\|^2 = \sum_j \big((\lambda_+)_j - (\lambda_-)_j\big)^2 = \|(\lambda_+ - \lambda_-)_+\|^2 = c$. All three matrices are diagonal in a common eigenbasis of S, so they commute.

Conversely, assume $S = H + P$. Let $\mu_1 \geq \cdots \geq \mu_N$ be the eigenvalues of H; symmetry of the spectrum gives $\mu_{N+1-k} = -\mu_k$. Writing $\lambda_k(S)$ for the k-th largest eigenvalue of S, the Courant–Fischer min–max theorem [4, Theorem 4.2.6] gives

$$\lambda_k(S) = \max_{\dim V = k} \min_{\substack{x \in V \\ \|x\| = 1}} x^T (H + P)\, x \;\geq\; \max_{\dim V = k} \min_{\substack{x \in V \\ \|x\| = 1}} x^T H x = \mu_k,$$

since $x^T P x \geq 0$ for every x. Fix $t > 0$. If $\lambda_k(S) < -t$ then $\mu_k < -t$, so $\mu_{N+1-k} = -\mu_k > t$ and hence $\lambda_{N+1-k}(S) \geq \mu_{N+1-k} > t$. As $k \mapsto N + 1 - k$ is injective,

$$\#\{k : \lambda_k(S) < -t\} \;\leq\; \#\{k : \lambda_k(S) > t\} \qquad \text{for every } t > 0. \tag{4}$$

Now suppose the rank condition fails, say $(\lambda_+)_j < (\lambda_-)_j$, and choose t with $(\lambda_+)_j < t < (\lambda_-)_j$. Since λ_+ is nonincreasing, the eigenvalues of S exceeding t are among $(\lambda_+)_i$ for $i < j$, so the right side of (4) is at most $j - 1$; since $(\lambda_-)_i \geq (\lambda_-)_j > t$ for $i \leq j$, the left side is at least j. This contradicts (4).

(4): Apply (3) to $-L$, which preserves a and b, exchanges c with d and λ_+ with λ_-, and negates H and P.

The vertices. Each vertex is the face on which the other three coordinates vanish. If $a = 1$ then $c = d = 0$ gives $\lambda_+ = \lambda_-$ by (3) and (4), and $b = 0$ makes S semidefinite by (2), so $\lambda_+ = \lambda_- = 0$ and $S = 0$: L is antisymmetric. Conversely, an antisymmetric L has $S = 0$, so $a = \|T\|^2 = 1$. If $b = 1$ then L is symmetric with $\lambda_+ = \lambda_-$ by (1), (3), and (4); the converse is immediate from the definitions and the partition. If $c = 1$ then L is symmetric by (1) and S is semidefinite by (2); were S negative semidefinite, then $\lambda_+ = 0$ and $d = \|\lambda_-\|^2 = \|S\|^2 = 1$, contradicting $d = 0$; so L is positive semidefinite. Conversely, a symmetric positive semidefinite L has $T = 0$ and $\lambda_- = 0$, so $a = b = d = 0$ and $c = 1$. The case $d = 1$ follows by applying this to $-L$, which exchanges λ_+ with λ_- and hence c with d. □

Appendix C. Moments

Let $S \neq 0$ be the symmetric component of the bilinear form of a head and let $\widehat{S} = S/\|S\|$, where $\|S\| = \sqrt{\operatorname{tr}(S^2)}$ is the Frobenius norm. Every eigenvalue of $\widehat{S}$ has absolute value less than one, since $\widehat{S}$ has unit Frobenius norm and more than one nonzero eigenvalue; hence $I - \widehat{S}^2$ is invertible. Define

$$O = \operatorname{tr}\big(\widehat{S}\,(I - \widehat{S}^2)^{-1}\big), \qquad E = \operatorname{tr}\big(\widehat{S}^2(I - \widehat{S}^2)^{-1}\big).$$

In terms of the eigenvalues λ_i of $\widehat{S}$,

$$O = \sum_i \frac{\lambda_i}{1-\lambda_i^2} = \sum_{k \text{ odd}} \sum_i \lambda_i^k, \qquad E = \sum_i \frac{\lambda_i^2}{1-\lambda_i^2} = \sum_{\substack{k \geq 2 \\ k \text{ even}}} \sum_i \lambda_i^k :$$

O is the sum of all odd moments of the spectrum and E the sum of all even moments (Lemma 9; Lemma 10 expresses O and E directly in terms of the key and query matrices). Replacing S by $-S$ changes the sign of O and leaves E unchanged.

We keep the notation of Appendix B: Λ_m, the involution ι, and the sorted lists $\lambda_\pm$.

Lemma 9. *Every $\lambda \in \Lambda_m$ satisfies $|\lambda_i| < 1$ for all i. Define*

$$O(\lambda) = \sum_i \frac{\lambda_i}{1-\lambda_i^2}, \qquad E(\lambda) = \sum_i \frac{\lambda_i^2}{1-\lambda_i^2}, \qquad R(\lambda) = \sum_i \frac{1}{1-\lambda_i}.$$

Then:

1. *$O(\lambda) = \sum_{\substack{k \geq 1 \\ k \text{ odd}}} \sum_i \lambda_i^k$ and $E(\lambda) = \sum_{\substack{k \geq 2 \\ k \text{ even}}} \sum_i \lambda_i^k$, both series converging absolutely;*
2. *$O = \frac{1}{2}(R - R \circ \iota)$ and $n + E = \frac{1}{2}(R + R \circ \iota)$; in particular $O \circ \iota = -O$ and $E \circ \iota = E$;*
3. *if S is a symmetric matrix whose spectrum, with multiplicity, is the entries of λ, then*

$$O(\lambda) = \operatorname{tr}\big(S(I - S^2)^{-1}\big), \quad E(\lambda) = \operatorname{tr}\big(S^2(I - S^2)^{-1}\big), \quad R(\lambda) = \operatorname{tr}\big((I - S)^{-1}\big).$$

Proof. If $|\lambda_i| = 1$ for some i then, since $\|\lambda\| = 1$, all other entries vanish, contradicting the presence of both a negative and a positive entry; and $|\lambda_i| \leq \|\lambda\| = 1$. Hence $|\lambda_i| < 1$. For (1), apply $\sum_{k \text{ odd}} t^k = t/(1-t^2)$ and $\sum_{k \geq 2 \text{ even}} t^k = t^2/(1-t^2)$, valid for $|t| < 1$, to each entry and sum over i. For (2), the entries of $\iota\lambda$ form the same multiset as the entries of $-\lambda$ (Lemma 8), so $R(\iota\lambda) = \sum_i (1+\lambda_i)^{-1}$, and the claims follow from

$$\frac{1}{2}\left(\frac{1}{1-t} - \frac{1}{1+t}\right) = \frac{t}{1-t^2}, \qquad \frac{1}{2}\left(\frac{1}{1-t} + \frac{1}{1+t}\right) = 1 + \frac{t^2}{1-t^2}.$$

The parity statements follow from (2) since $\iota^2 = \mathrm{id}$. Part (3) is the spectral mapping $\operatorname{tr} f(S) = \sum_i f(\lambda_i)$ applied to $f(t) = t/(1-t^2)$, $t^2/(1-t^2)$, and $(1-t)^{-1}$. □

Lemma 10. *Let $K, Q \in M_{n \times N}(\mathbb{R})$, let $S = \frac{1}{2}(K^T Q + Q^T K)$, and let*

$$M = \begin{pmatrix} K \\ Q \end{pmatrix} : \mathbb{R}^N \to \mathbb{R}^{2n}, \qquad G = MM^T, \qquad J = \begin{pmatrix} 0 & I_n \\ I_n & 0 \end{pmatrix}.$$

Then $S = \frac{1}{2} M^T J M$ and

$$\operatorname{tr}(S^k) = \operatorname{tr}\big((\tfrac{1}{2}JG)^k\big) \qquad (k \geq 1);$$

in particular $\|S\|^2 = \frac{1}{4}\operatorname{tr}\big((JG)^2\big)$. If moreover M is surjective and $n \geq 1$, then the nonzero spectrum of S, with multiplicity, is the spectrum of $\frac{1}{2}JG$; the sorted normalized nonzero spectrum λ of S lies in Λ_n, and with $H = JG/(2\|S\|)$,

$$O(\lambda) = \operatorname{tr}\big(H(I - H^2)^{-1}\big), \qquad E(\lambda) = \operatorname{tr}\big(H^2(I - H^2)^{-1}\big), \qquad R(\lambda) = \operatorname{tr}\big((I - H)^{-1}\big).$$

Proof. Block multiplication gives $M^T J M = K^T Q + Q^T K = 2S$. By the cyclic property of the trace,

$$\operatorname{tr}\big((M^T J M)^k\big) = \operatorname{tr}\big(M^T (JG)^{k-1} J M\big) = \operatorname{tr}\big((JG)^{k-1} J M M^T\big) = \operatorname{tr}\big((JG)^k\big).$$

For the second claim, recall that for any $A \in M_{N \times 2n}(\mathbb{R})$ and $B \in M_{2n \times N}(\mathbb{R})$ the nonzero eigenvalues of AB and BA coincide with multiplicity; applying this to $A = M^T$ and $B = \frac{1}{2}JM$ identifies the nonzero spectrum of S with the nonzero spectrum of $\frac{1}{2}JMM^T = \frac{1}{2}JG$. If M is surjective then G is positive definite. The matrix $B = G^{1/2} J G^{1/2}$ is symmetric and congruent to J, so by Sylvester's law of inertia it has exactly

n positive and n negative eigenvalues; and $JG = G^{-1/2} B\, G^{1/2}$ is similar to B, hence diagonalizable with the same real spectrum. Consequently the nonzero spectrum of S consists of n positive and n negative eigenvalues, so its sorted normalization λ lies in Λ_n and, by Lemma 9, every entry of λ has absolute value less than one. The matrix H is diagonalizable with spectrum λ, and for a diagonalizable matrix the trace of a rational function with poles off the spectrum is the sum of its values on the spectrum, which gives the three displayed identities. □

Appendix D. Supplementary Figures

This appendix collects supplementary figures for the experiments in §3.

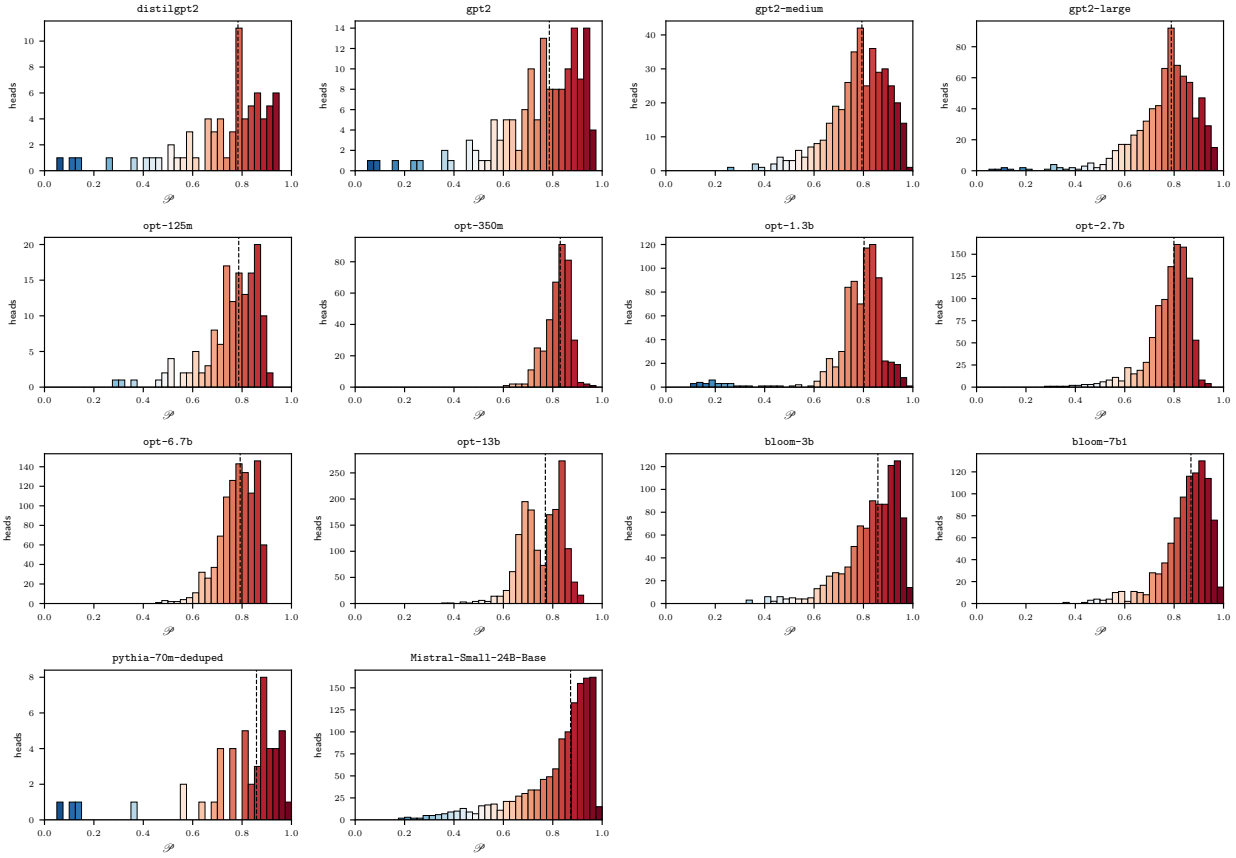

Figure 11. Distribution of the trained-head pairing scores $\mathscr{P}$ in Figure 14. Dashed lines mark model medians.

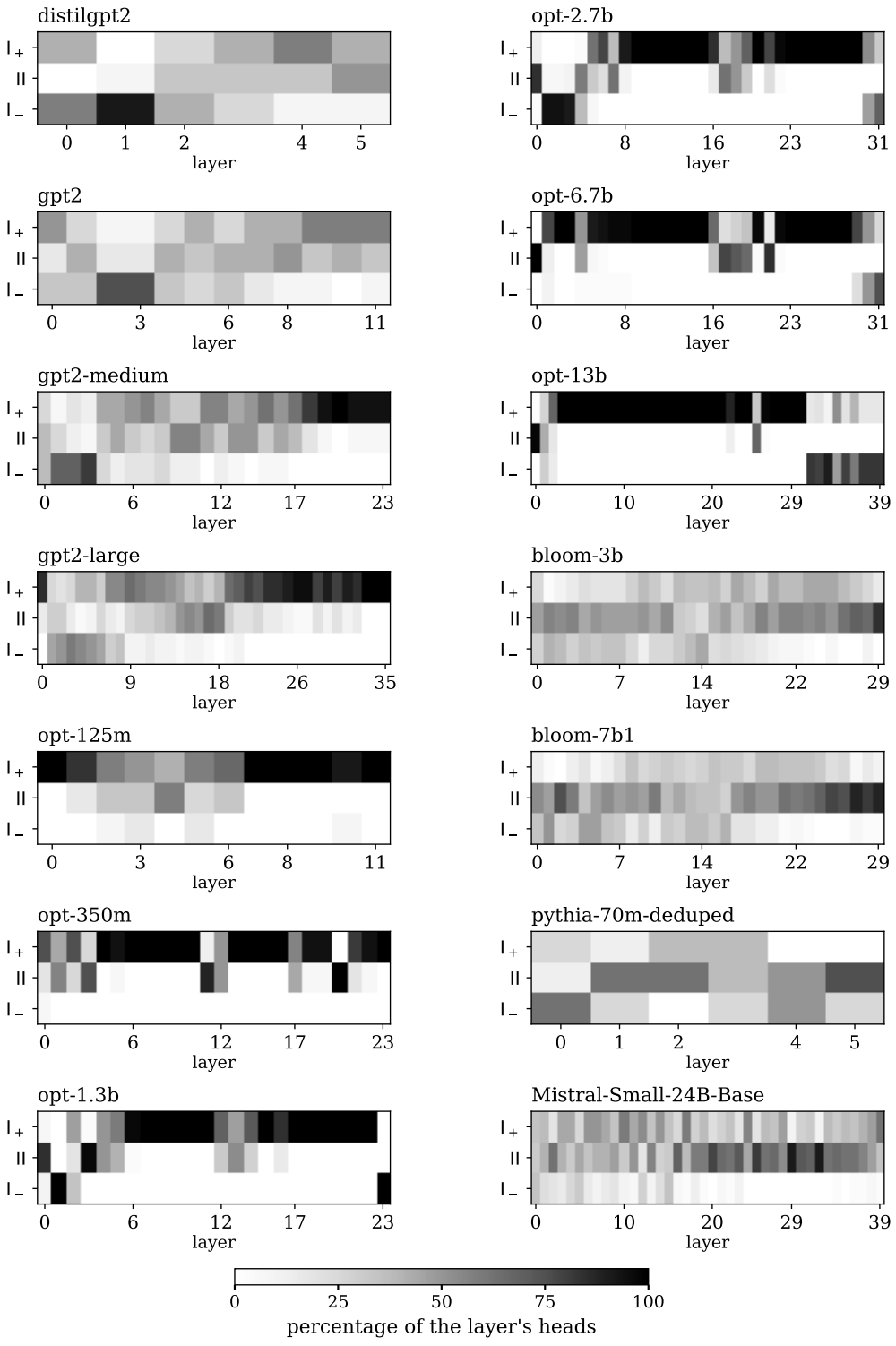

FIGURE 12. Parity by layer in all fourteen models. Each cell shows the percentage of a layer's heads of the indicated type, with a shared scale from white (0%) to black (100%). Types use the symmetric profile: Type II has $c_S + d_S \leq 1/50$; otherwise $c_S > d_S$ gives Type I_+ and the remaining heads are Type I_-. Layers are indexed from zero; RoPE models use relative position zero.

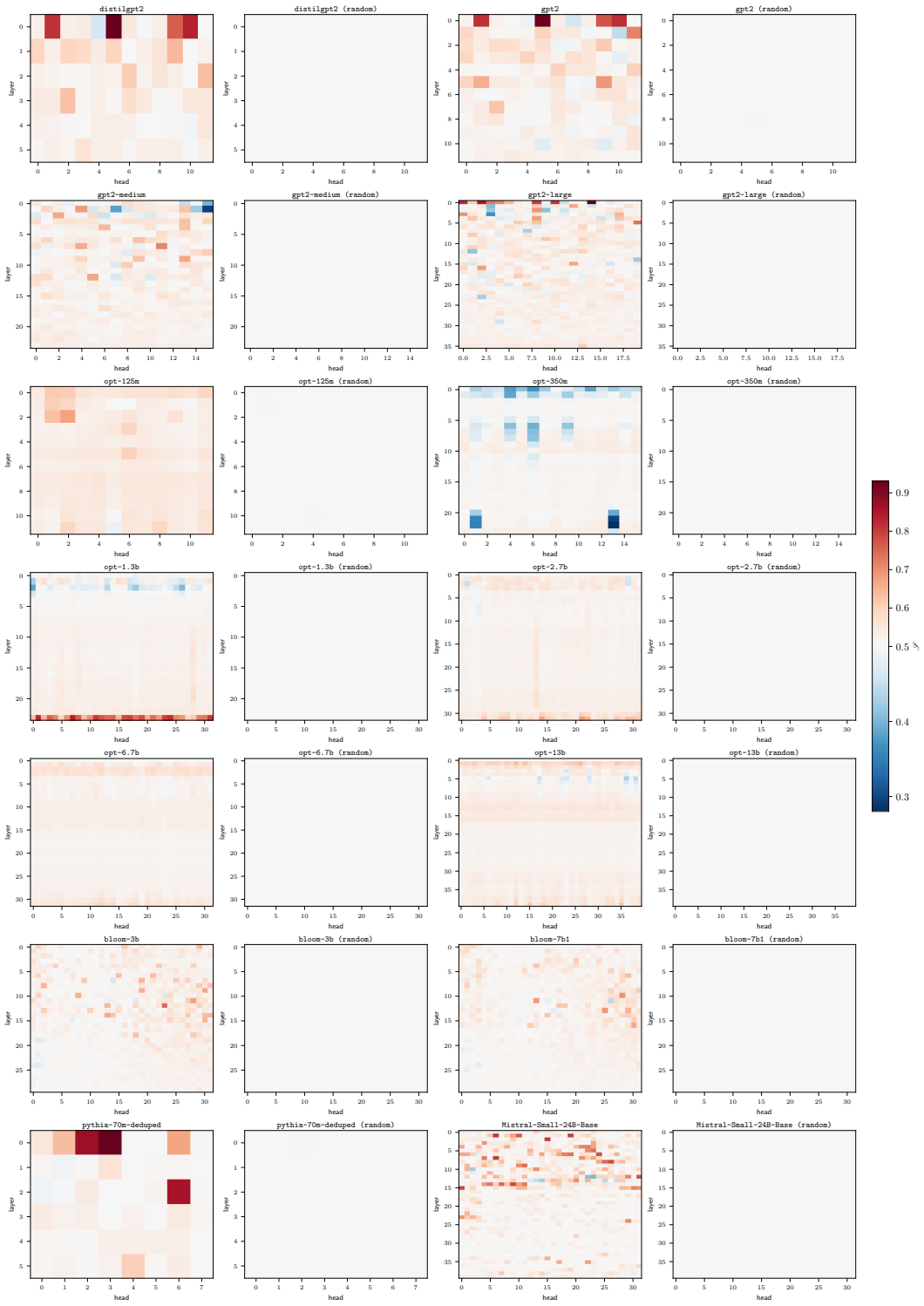

FIGURE 13. Symmetric energy $\mathscr{S} = \|S\|^2/\|L\|^2$ at each attention head. For each model, the trained heads (left) are paired with independent random products $L = K^T Q$ (right), with iid standard normal $K, Q \in \mathbb{R}^{n\times N}$ of the same dimensions. Rows index layers and columns index heads, both starting at zero. The shared color scale is white at $1/2$, red above and blue below; the random baseline has expectation $(N+1)/(2N)$. RoPE models use relative position $d = 0$.

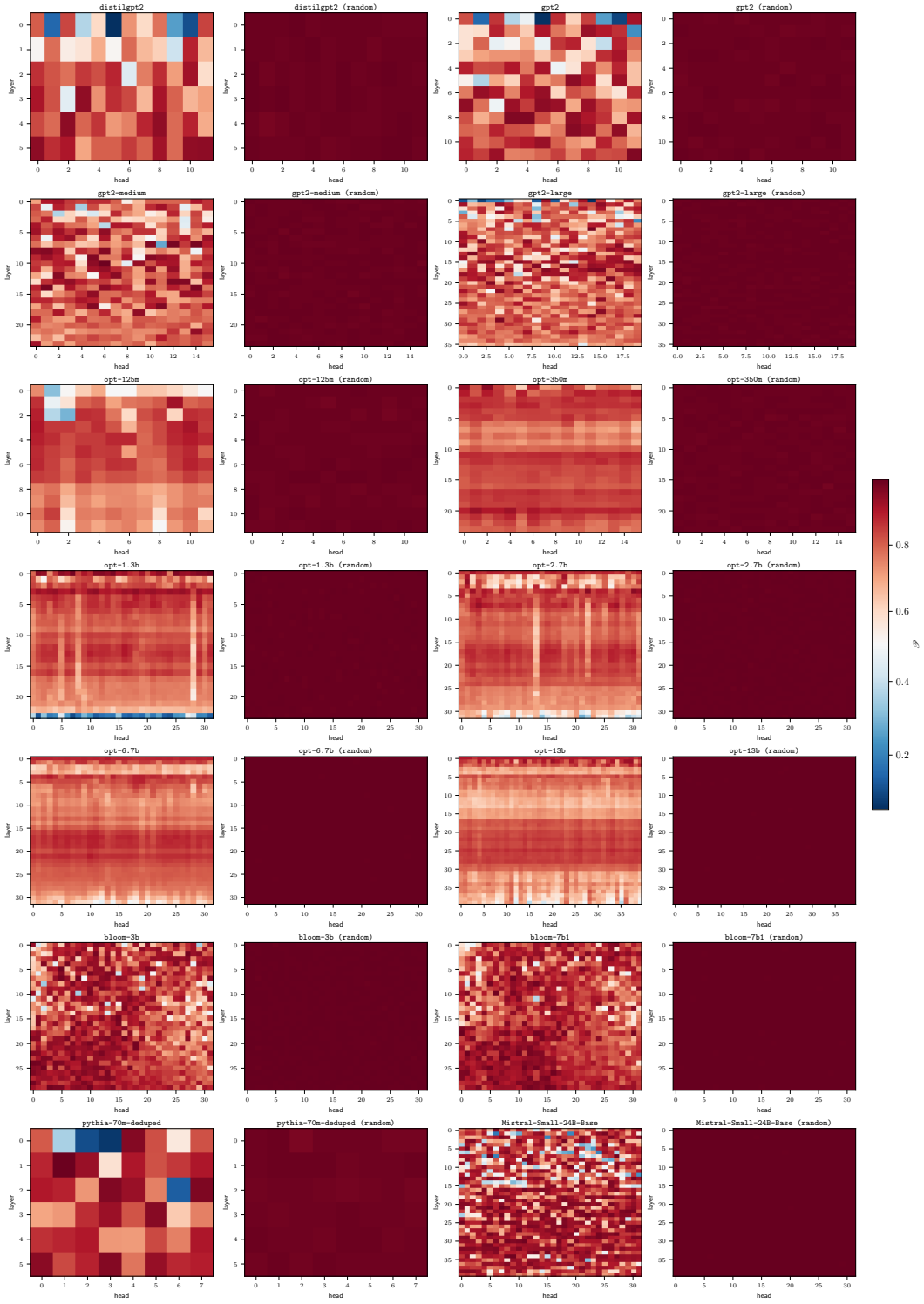

FIGURE 14. Pairing score $\mathscr{P} = 1 - \|\lambda_+ - \lambda_-\|/\|S\|$ at each attention head. Each model's trained heads (left) are paired with independent Gaussian QK products of the same dimensions (right), as in Figure 13. Rows index layers and columns index heads. A score of one means the symmetric spectrum is symmetric about zero. The shared color scale is white at $1/2$, red above and blue below. RoPE models use relative position $d = 0$.

References

[1] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

[2] BigScience Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, J. Tow, A. M. Rush, S. Biderman, et al. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[3] S. Darveshi. Spectral taxonomy of QK circuits in transformer models. LessWrong, 2025. https://www.lesswrong.com/posts/Yig9fc7wAxKqG63Do/spectral-taxonomy-of-qk-circuits-in-transformer-models.

[4] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2013.

[5] M. Migliarini. The self-hating attention head: A deep dive in GPT-2. LessWrong / AI Alignment Forum, 2025. https://www.lesswrong.com/posts/wxPvdBwWeaneAsWRB/the-self-hating-attention-head-a-deep-dive-in-gpt-2-1.

[6] Mistral AI Team. Mistral Small 3. https://mistral.ai/news/mistral-small-3/, Jan. 2025.

[7] O. Press, N. A. Smith, and M. Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Technical Report*, 2019.

[9] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[10] M. Saponati, P. Sager, P. V. Aceituno, T. Stadelmann, and B. Grewe. The underlying structures of self-attention: symmetry, directionality, and emergent dynamics in transformer training. *arXiv preprint arXiv:2502.10927*, 2025.

[11] J. Su, M. Ahmed, Y. Lu, S. Pan, B. Wen, and Y. Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[13] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.